Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Disaster Robotics in the Deep Learning Era] {Disaster Robotics in the Deep Learning Era: A Systematic Review of Platforms, Autonomy, Methods, Datasets, Challenges, and Future Research Directions

Georgios Angelidis
Harokopio University of Athens
Aggelos Psiris
Harokopio University of Athens
Thomas Lagkas
Democritus University of Thrace
Vasileios Argyriou
Kingston University London
Panagiotis Sarigiannidis
University of Western Macedonia
Iraklis Varlamis
Harokopio University of Athens
Georgios Th. Papadopoulos
g.th.papadopoulos@hua.gr

Harokopio University of Athens

Research Article

Keywords: Disaster robotics, search and rescue, field robotics, deep learning, foundation models, human-robot teaming

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11587551/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Disaster Robotics in the Deep Learning Era: A Systematic Review of Platforms, Autonomy, Methods, Datasets, Challenges, and Future Research Directions

Georgios Angelidis[1], Aggelos Psiris[1], Thomas Lagkas[2], Vasileios Argyriou[3], Panagiotis Sarigiannidis[4], Iraklis Varlamis[1], Georgios Th. Papadopoulos[1,5*]

[1]Department of Informatics and Telematics, Harokopio University of Athens, Athens, Greece.
[2]Department of Informatics, Faculty of Science, Democritus University of Thrace, Kavala, Greece.
[3]Department of Networks and Digital Media, Kingston University, Kingston upon Thames, United Kingdom.
[4]Department of Electrical and Computer Engineering, University of Western Macedonia, Kozani, Greece.
[5]Archimedes Research Unit, Athena Research Center, Athens, Greece.

*Corresponding author(s). E-mail(s): g.th.papadopoulos@hua.gr;

Abstract

Disaster robotics enables responders to sense and act at a disaster site from a safe distance, converting responder risk into an equipment one. The field is increasingly shaped by deep learning and, most recently, by foundation models. However, existing surveys adopt a single perspective or platform class, predate the deep learning era or remain largely narrative. Moreover, none grades every reviewed work on a common validation maturity scale. This paper presents a systematic review of disaster robotics in the deep learning era, restricted to disaster-connected works evaluated on disaster data or in real, replicated or simulated emergency settings. The evolution of the field is traced through successive research phases, from teleoperated platforms to foundation-model interfaces and responder-grounded field autonomy. A multi-criteria taxonomy is formed, covering response tasks, robotic platforms, perception, autonomy, learning strategies, foundation model types, human-robot teaming, coordination, communication and deployment, disaster domains, and validation

maturity. Each criterion is analyzed comparatively, with critical insights into the strengths and limitations of its categories under field conditions. Additionally, public datasets and benchmarks are reviewed. Current challenges are also paired with future research directions, from data acquisition and field hardening to verifiable autonomy, responder trust and evaluation. The analysis shows that systems deployed at real incidents still rely predominantly on model-based methods. It also shows that recent research leans increasingly on simulation and benchmark evaluation. Finally, foundation models have entered the field mainly as perception front ends and mission-level planners rather than as embodied policies, without yet reaching a real incident.

Keywords: Disaster robotics, search and rescue, field robotics, deep learning, foundation models, human-robot teaming

1 Introduction

Disaster robotics assists in disaster prevention, preparedness, response and recovery. Within it, rescue robots enable responders and other stakeholders to sense and act at a distance from the site of a disaster or extreme incident (Murphy 2014). These systems serve as extensions of emergency professionals (Murphy et al. 2016). Teleoperated robots were already used for recovery after the Three Mile Island and Chernobyl nuclear accidents (Murphy 2014; Nagatani et al. 2013). Later, the World Trade Center collapse of 2001 marked the first known use of robots for technical search in an urban search-and-rescue response (Casper and Murphy 2003; Murphy 2004b). Since then, first-hand records of subsequent incidents have documented what robots achieved and where they fell short (Murphy and Stover 2008; Nagatani et al. 2013; Whitman et al. 2018; Manzini et al. 2023b). On the basis of these records, deployment reviews derived the technical and sociotechnical challenges of the field (Murphy 2014; Murphy et al. 2016). The societal importance of the field follows from three conditions of response work, namely: a) Victim mortality in a structural collapse increases drastically after the first 48 hours (Murphy 2000, 2004a). In contrast, in the five deployments of 2010 clearly aimed at survivor search, a robot arrived on average 4.2 days after the incident (Murphy 2014). b) Rescue work can be as dangerous to the rescuers as the initial event, since moving through a damaged structure can cause further collapse (Murphy 2000). Moreover, searching confined spaces has cost rescuers their lives (Casper and Murphy 2003). c) A robot can enter a void too small, too hot or too toxic for a rescuer (Casper and Murphy 2003). Likewise, it can operate inside a contaminated site while its operators remain at a safe distance (Chiou et al. 2022; Sakaue et al. 2025). In this way, the robot converts responder risk into an equipment one, since robots are expendable (Casper and Murphy 2003), whereas rescuers are not (Liu and Nejat 2013).

However, the literature that answers to this promise is markedly heterogeneous in what it demonstrates. Specifically, part of the general-purpose robotics and embodied artificial intelligence literature names emergency response as an application without any concrete connection to it. Admitting such work would overstate the maturity of the field. Therefore, the central scoping decision of this review is a two-part restriction, namely: a) A concrete tie to disaster response. This tie is established through a disaster or emergency site, a recognized response task, disaster data, responders as intended users, a competition or test course for

disaster robotics, or a hazardous setting representative of response missions. b) Evaluation in such a setting. Such evaluations range from disaster data and disaster-specific simulation, through laboratory, arena, competition, field and responder evaluations, to deployment at real incidents. Studies of emergency personnel, datasets, test methods and competition records are admitted as well. Additionally, the level attained by each work is recorded explicitly. Within this restriction, the scope remains broad. It spans ground, aerial and marine robots, manipulators and overhead sensing, as well as hazards from structural collapse to wildfire, flood, subterranean and maritime emergencies. Hence, what is excluded is the assertion of applicability, rather than a class of platform or method.

Disaster environments impose six main constraints that jointly separate the field from conventional mobile robotics and laboratory-scale embodied artificial intelligence, namely: a) Degraded sensing. Darkness hinders cameras, while smoke, dust and fog corrupt LiDAR returns (Ebadi et al. 2024; Dhrafani et al. 2025). Moreover, feature-poor tunnels and corridors can render LiDAR-based localization degenerate (Tuna et al. 2024; Petrlík et al. 2025). b) Unstructured environments unknown in advance (Delmerico et al. 2019). In such environments, a collapsed structure must be treated as if on the verge of further collapse (Murphy 2004a). Furthermore, no prior map or reliable satellite positioning is typically available underground (Ebadi et al. 2020). c) Scarce communication and computation. Rescue robots typically operate without satellite positioning and wireless communication (Murphy 2014). In particular, subterranean sites are communication-denied (Tranzatto et al. 2024), while deployed mesh networks typically reach only the vicinity of the entrance (Agha et al. 2022). Insufficient wireless connectivity has also been a pervasive issue even in large-scale aerial responses (Manzini et al. 2023b). In parallel, weight and power limit onboard computation and payload (Murphy et al. 2016), most acutely on small aerial platforms (Cladera et al. 2025; Amessegher et al. 2025). d) Safety criticality. A run-time failure can result in death, injury or loss of property (Yanco and Drury 2007). In addition, robots have been lost during real responses (Murphy et al. 2009; Nagatani et al. 2013). Besides, responders require heavily vetted technology for actual deployment (Delmerico et al. 2019). e) Chronic data scarcity, both for training learned models (Gupta et al. 2019; Scheele et al. 2025) and for characterizing how robots perform in disaster situations (Murphy and Stover 2008). f) Human factors. Operators and information consumers act under extreme time pressure, personal risk and possible fatigue, while the problem holder may be a technical search specialist with little or no robot training (Murphy 2004a).

Over the past decade, deep learning has reshaped a considerable part of the field. This shift from hardware development toward intelligence technologies is also visible in the publication record of disaster robotics (Hu et al. 2026). The reshaping concerns three main respects, namely: a) Perception. Deep learning is used by all recent damage assessment works, as well as increasingly by hazard detection and victim search ones. These works are supported by annotated aerial and satellite datasets for person detection, emergency scene classification and building damage assessment (Božić-Štulić et al. 2019; Kyrkou and Theocharides 2019; Gupta et al. 2019). b) State estimation. The most visible hardening of state estimation against the above degradations comes from the resilient multi-robot exploration and mapping of the DARPA Subterranean (SubT) Challenge teams. In these systems, deep networks enter mainly as onboard object detectors (Tranzatto et al. 2022; Agha et al. 2022; Scherer et al. 2022; Hudson et al. 2022). c) Control. Policies trained in simulation by reinforcement learning were

carried onto legged and aerial platforms in competitions and field trials (Tranzatto et al. 2022; Hickling et al. 2026; Pal et al. 2025).

Most recently, pre-trained foundation models (FMs) have entered the field as: a) Promptable segmentation models adapted to overhead change and damage mapping (Zheng et al. 2024b; Ahn et al. 2025). b) Language-conditioned mission planners (Ravichandran et al. 2025; Cladera et al. 2025). c) Command layers that translate operator instructions into robot tasks (Bark et al. 2025; Jarabo-Peñas et al. 2025). d) Vision-language assistants for disaster imagery (Irvin et al. 2025; Wang et al. 2025). The present review accordingly emphasizes the literature within the last 3 years, while retaining the milestone works that led there. However, one of its own findings is that foundation-model adoption is real, but shallow. This adoption is also concentrated at the two ends of the autonomy stack. In particular, the models that plan, navigate or interact with operators are prompted rather than fine-tuned on disaster data (Cladera et al. 2025; Zhou et al. 2024; Shi et al. 2025). In contrast, the assistants for disaster imagery are fine-tuned on it (Irvin et al. 2025; Wang et al. 2025).

Literature surveys have addressed disaster robotics and its neighboring fields. However, none of them analyzes the contemporary, deep learning-driven landscape of disaster robotics in a systematic, multi-criteria manner. As detailed in Table 1, existing surveys: a) Adopt a single perspective, such as control (Liu and Nejat 2013), localization and mapping (Chen et al. 2026) or human-robot teaming (Adams et al. 2024). b) Are restricted to a single platform class, most often the uncrewed aerial vehicle (UAV) (Recchiuto and Sgorbissa 2018; Khan et al. 2022). c) Predate the deep learning era (Murphy 2014; Murphy et al. 2016) or focus on learning only in passing (Delmerico et al. 2019; Khan et al. 2022). In other cases, they are bibliometric rather than technical (Hu et al. 2026), or they address the modern methodological landscape outside the disaster setting (Firoozi et al. 2025). d) Adopt a largely narrative discussion or tabulate works along a few descriptive attributes (Liu and Nejat 2013; Khan et al. 2022), rather than coding the whole corpus along a comprehensive set of criteria. e) Do not grade every reviewed work on a common validation maturity scale. This holds even where they contrast real-disaster achievements with research readiness (Murphy et al. 2016; Delmerico et al. 2019), or record the experimental environment (Liu and Nejat 2013) or benchmark type (Chen et al. 2026) of the works they summarize. As a result, the reader is unable to distinguish a capability that has met an emergency from one that has met a benchmark.

In order to close these gaps, this review provides a holistic, systematic and in-depth review of **disaster robotics in the deep learning era**, under the scope restriction stated above. Its main contributions are the following:

- Analysis of the **evolution of disaster robotics research** in five phases, from teleoperated platforms and field-deployment lessons to FM interfaces and responder-grounded field autonomy. Each phase is accompanied by its key systems and milestones.
- A **multi-criteria taxonomy** along ten criteria (response task, platform, perception, autonomy, learning strategy, FM type, teaming, deployment, domain and validation maturity) and 67 categories. Since it classifies only disaster-connected works, the taxonomy also exposes what the corpus lacks. For example, the corpus contains no vision-language-action model or generalist robot policy applied to a response task, whereas the capabilities of such models are often demonstrated in controlled tabletop manipulation (Firoozi et al. 2025).

Table 1 Comparative analysis of representative surveys in the field of disaster robotics and its neighboring domains.

Survey	Scope	Field-deployment focus	Primary contributions	Limitations
Liu and Nejat (2013), J. Intell. Robot. Syst.	Urban search-and-rescue robotics, control perspective	Partial	• Low-level, semi-autonomous, learning-based and multi-robot control, with the test environment of each scheme tabulated	• Control only, mainly ground robots, pre-deep learning • No dataset review; environments tabulated, not analyzed
Murphy (2014), MIT Press	Foundational monograph on deployments and fieldwork	Yes	• 34 known disaster deployments to April 2013 (ground, aerial, marine), failures and fieldwork guidance	• Pre-deep learning; organized by modality and deployment record • No taxonomy or comparison of methods
Murphy et al. (2016), Springer Handbook of Robotics	Handbook treatment of disaster robotics	Yes	• Tasks, robot types, deployment outcomes, lessons per subsystem, and simulators, test beds and standards	• Predates learning-based perception and autonomy (deployments up to 2013) • No per-method comparison or dataset review
Recchiuto and Sgorbissa (2018), J. Field Robot.	Post-disaster assessment with uncrewed aerial vehicles	Partial	• Platforms, multi-robot software, sensing, mapping, coverage, navigation and human-swarm interfaces, with real deployments	• Single platform class and response phase • Learning as one approach class; no dataset review • Per-aspect classifications, no joint taxonomy
Delmerico et al. (2019), J. Field Robot.	Rescue robotics from 2014 to early 2018 versus first-responder needs	Yes	• Ground, aerial, marine and amphibious robots and interfaces versus requirements gathered from response experts	• Qualitative, modality-organized, no methodological taxonomy • No deep learning treatment or dataset review
Khan et al. (2022), J. Field Robot.	Uncrewed aerial vehicle networks across disaster management stages	Partial	• Sensor networks, Internet of Things and edge, fog and cloud computing; nine data-collection scenarios; relays	• Aerial-only, networking-centered • Deep learning only in isolated studies; no dataset review • Validation conditions not tabulated

- A **per-criterion comparative analysis** with critical insights on the relative strengths and limitations of the categories under field conditions.
- An explicit **validation maturity axis**, from simulation to real-incident deployment. This axis separates demonstrated from asserted capability, while also making the scarcity of incident-level evidence measurable.
- A **source-organized review of public datasets and benchmarks**, with observations on their coverage, modalities and provenance.
- A structured discussion of **current challenges** and the most promising **future research directions**.

Table 1 (continued)

Survey	Scope	Field-deployment focus	Primary contributions	Limitations
Adams et al. (2024), Annu. Rev. Control Robot. Auton. Syst.	Military and first-response human-robot teaming	Partial	• Teammate roles mapped onto a military mission; near- and mid-term teaming challenges and recommendations	• Single (teaming) perspective, military-centered • No platform, perception, autonomy or dataset analysis
Firoozi et al. (2025), Int. J. Robot. Res.	Foundation models in robotics	No	• Foundation models in robot perception, decision making and control, and the challenges hindering their adoption	• Not disaster-specific, no field-validation axis • Search and rescue only as a connectivity-limited setting
Chen et al. (2026), J. Field Robot.	Localization and mapping for search-and-rescue ground robots	Partial	• Localization, mapping and active exploration in complex environments; normalized comparison of LiDAR-based algorithms	• Single perspective; no response tasks or teaming • Learning noted only within techniques; no dataset review
Hu et al. (2026), J. Field Robot.	Disaster robotics publications, 1992-2025	No	• PRISMA-screened mapping of 2,201 Web of Science publications, tracing the shift toward intelligence technologies	• Bibliometric, keyword clusters rather than a method taxonomy • No comparative analysis or dataset review
Current survey	Deep learning-era disaster robotics, limited to disaster-linked works evaluated in emergency conditions (real, replicated or simulated) or on disaster data	Yes	• Ten-criterion, 67-category taxonomy; research phases and milestones • Per-criterion comparative analysis; validation maturity axis • Datasets and benchmarks by source; challenges and future directions	• Reported evidence, no common-protocol re-evaluation or ranking • Partially overlapping criteria

The remainder of the manuscript is organized as follows. Section 2 describes the review methodology, while Section 3 traces the evolution of the field. Section 4 then describes the ten criteria and their categories, which Sections 5-14 analyze in turn. Subsequently, Section 15 reviews public datasets and benchmarks. Sections 16 and 17 discuss current challenges and future research directions, respectively. Finally, Section 18 concludes the manuscript.

2 Literature review methodology

The literature review followed a structured and iterative methodology. Its steps are described below and summarized in Fig. 1.

2.1 Scope and objectives formulation

The review targets deep learning-era research on disaster robotics, i.e., approaches that employ robotic platforms for response tasks. The platforms range from tracked, wheeled, legged and serpentine ground robots, as well as aerial and marine vehicles, to manipulators and airborne or spaceborne sensing systems. The response tasks comprise reconnaissance and mapping, victim search and detection, damage and structural assessment, hazard detection and monitoring, rubble access and void assessment, physical intervention and casualty handling, as well as reporting and decision support. These approaches are evaluated in real, replicated or simulated emergency conditions or on disaster data. Additionally, the studies of emergency responders that inform their design are included. However, deep learning was not a strict selection requirement. The reason is that most milestone deployments, as well as a considerable part of current field systems, do not employ deep neural networks. Emphasis was placed on the works published since 2023, which account for 100 of the 159 primary studies. Emphasis was also placed on milestone works with a substantial contribution to the field. In this context, the review aimed at tracking: a) The main categories of methods along multiple criteria. b) The publicly available datasets. c) The key trends and the evolution of the field. d) The current challenges and future research directions.

2.2 Literature search

Six databases were queried, namely IEEE Xplore, Google Scholar, Scopus, DBLP, arXiv and Web of Science. The queries combined targeted keywords (e.g., 'disaster robotics', 'rescue robot', 'search and rescue', 'urban search and rescue', 'first responder', 'subterranean exploration', 'victim detection', 'damage assessment') through Boolean operators. In parallel, they excluded application domains outside field-deployed disaster response (e.g., 'warehouse', 'agriculture', 'surgical', 'autonomous driving', 'manufacturing'). The search focused on 2023-2026. However, seminal and milestone works introduced earlier were also included, given the long field-deployment history of the domain. The Scopus query is given below as an example.

```
TITLE-ABS-KEY ( ( "disaster robot*" OR "rescue robot*" OR "search and rescue" OR "urban
search and rescue" OR "first responder*" OR "emergency response" ) AND ( "robot*" OR
"unmanned aerial vehicle" OR "unmanned ground vehicle" OR "autonomous system*" ) ) AND
NOT TITLE-ABS-KEY ( "warehouse" OR "agricultur*" OR "surgical" OR "autonomous driving"
OR "manufacturing" OR "prosthe*" ) AND PUBYEAR > 2022 AND PUBYEAR < 2027 AND ( SRCTYPE
(j) OR SRCTYPE (p) )
```

The keywords were refined over multiple iterations. Moreover, the reference list of each retained work was analyzed in order to identify additional studies.

2.3 Eligibility criteria

Table 2 summarizes the inclusion (IC) and exclusion (EC) criteria, as well as the definition of a contextual reference (CR). The criteria were applied uniformly, irrespective of the publication year. In contrast, the contextual references were not subject to them, and they are excluded from every statistic of this section.

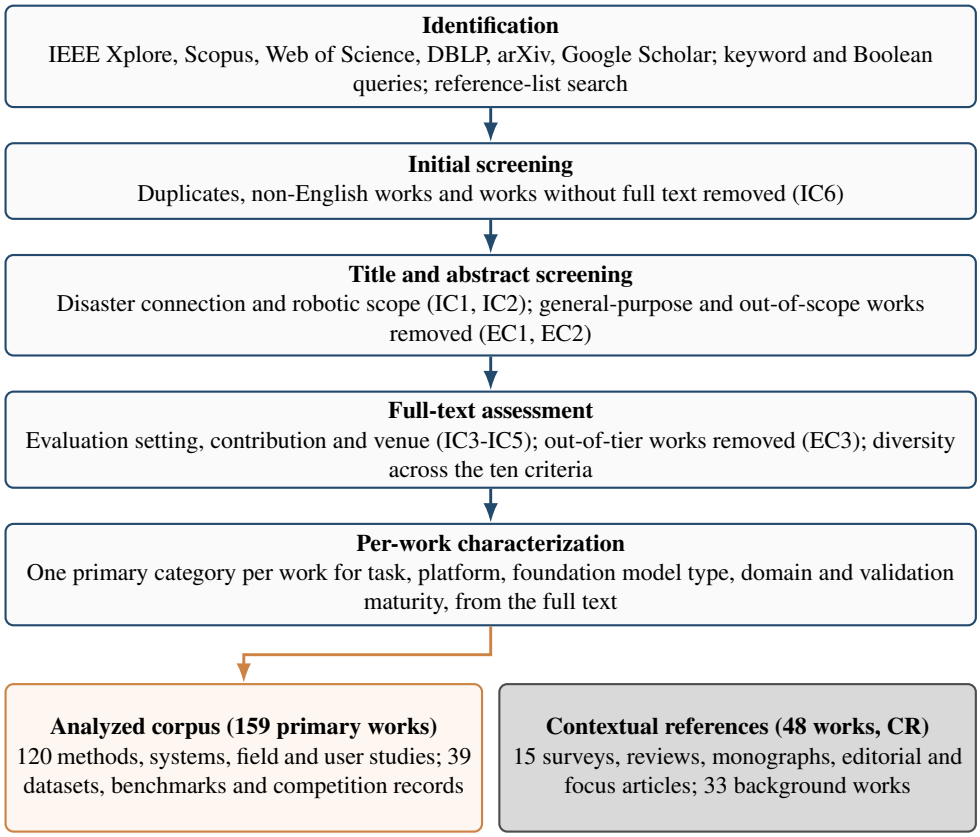


Fig. 1 Corpus construction and study selection procedure. The decision rules and eligibility criteria (Table 2) applied at each stage are shown, as well as the resulting analyzed corpus and contextual references

2.4 Screening

The selection followed the stages of Fig. 1, namely: a) Initial screening (IC6). b) Title and abstract screening (IC1, IC2, EC1, EC2). c) Full-text assessment (IC3-IC5, EC3). This stage also ensured diversity across the ten criteria of Section 4. d) Per-work characterization. This stage assigned each retained work, from its full text, a single primary category for response task, robotic platform, foundation model type, disaster domain and validation maturity. Additionally, it recorded whether the work employs deep neural networks. Of the retained works, 145 involve a disaster or emergency explicitly (IC1). Moreover, ten were admitted for operating in hazardous environments representative of response missions, and four for their link to the DARPA Robotics or Subterranean Challenges.

The analyzed corpus comprises 159 primary works. These include 120 methods, robotic systems, and field and user studies, as well as 39 datasets, benchmarks and competition

Table 2 Inclusion (IC) and exclusion (EC) criteria for the selection of the primary works, and definition of the contextual references (CR).

Code	Criterion
IC1	Concrete disaster connection, i.e., a disaster or emergency site, a response task, disaster data, responders as users, a disaster robotics competition or test course, or a hazardous environment representative of response missions (e.g., mines, tunnels, flooded cavities, nuclear facilities)
IC2	A robotic platform, a perception, autonomy, learning or teaming method for response robots, overhead sensing that supports the response, or a dataset, benchmark, test method or competition for such systems
IC3	Evaluation at a real incident, with responders, in field trials, test arenas, competitions or the laboratory, or on disaster data or disaster-specific simulation; requirement and user studies with emergency personnel, disaster datasets, test methods and competition records are also admitted
IC4	Substantial theoretical and/or experimental contribution, or a first-hand record of a real-incident deployment or a disaster robotics competition
IC5	Prominent peer-reviewed robotics, field robotics, artificial intelligence, computer vision or human-robot interaction venue (including workshop, companion and demonstration proceedings), or multidisciplinary journal
IC6	Written in English, with accessible full text
EC1	General-purpose methods whose relevance to disaster response is only asserted, i.e., without any of the connections of IC1
EC2	Out-of-scope application domains (e.g., warehouse, agriculture, surgical, autonomous driving, manufacturing)
EC3	Preprints without a peer-reviewed version, and venues outside the tier of IC5
CR	Surveys, reviews, monographs, and editorial or focus articles, as well as general-purpose robotics, computer vision and foundation model works cited as technological background, and out-of-tier original papers of datasets used by primary works

records. In addition, 48 contextual references are cited, which are excluded from all statistics. They comprise 15 surveys, reviews, monographs and editorial or focus articles, 10 of which are the prior surveys of Table 1. The remaining 33 are background works. They comprise 32 general-purpose works without a concrete disaster connection, as well as the out-of-tier original paper of a dataset used by a primary work. In total, 207 works are cited.

2.5 Bibliometric analysis

Fig. 2 shows six main trends over the 159 primary works, namely: a) Journal and conference papers are balanced, led by the Journal of Field Robotics (27) and IEEE SSRR (22). b) 100 works date from 2023-2026. c) Reconnaissance and mapping leads the response tasks (52 works). d) Tracked and wheeled ground vehicles fall from 48.1% of the works up to 2015 to at most 30.0% since 2023. In contrast, aerial vehicles, and satellite and airborne sensing reach 45.0% in 2023-2026. e) Real-incident deployments fall from 9 of 27 works up to 2015 to 7 of 100 in 2023-2026. f) Subterranean and mine incidents (35), and earthquake and structural collapse (27) lead the specific disaster domains. Additionally, deep neural networks are employed by 71 works (44.7%). Their share rises from zero before 2016 to 69.6% in 2025-2026. In contrast, the remaining 88 works, including 17 of the 19 real-incident deployments, employ no deep neural networks and rely mostly on model-based perception, estimation,

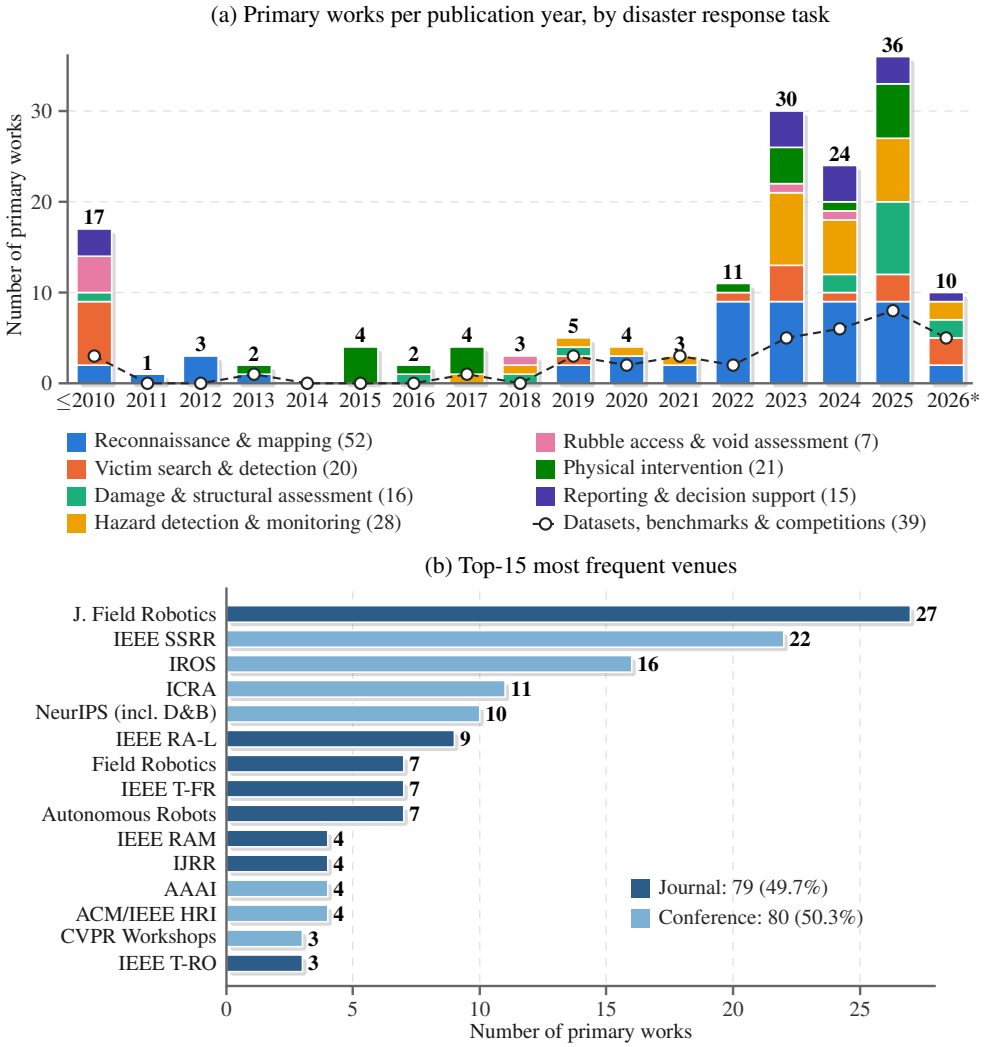


Fig. 2 Bibliometric statistics of the 159 considered primary works by: a) Year and response task. b) Venue (top 15). c) Platform per period. d) Validation maturity per period. e) Disaster domain. "Not applicable" marks works without an evaluated system, "Platform-agnostic" and "Multi-hazard or unspecified" mark works not tied to one platform or disaster domain, n denotes the works per period, and parenthesized numbers denote the category totals. Finally, 2026* denotes a partially covered year

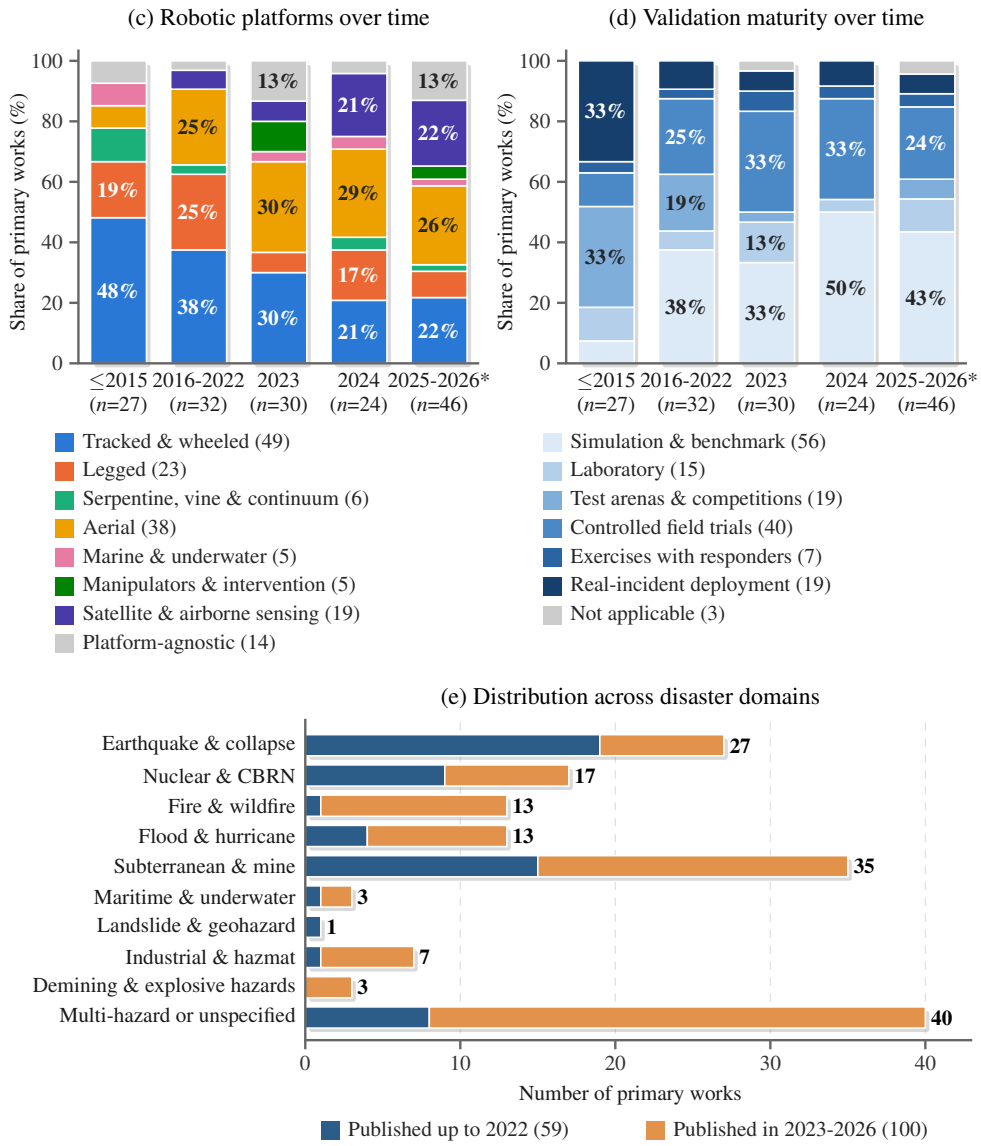


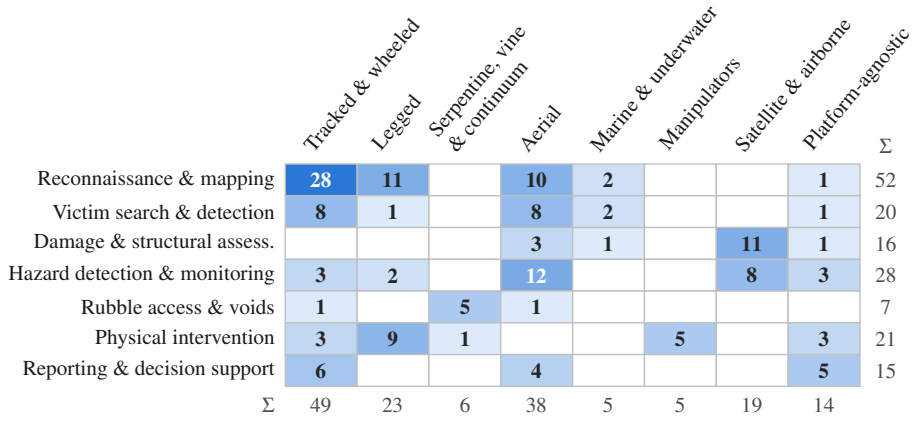
Fig. 2 (continued)

planning and control. These statistics describe research activity rather than the relative merit of the approaches.

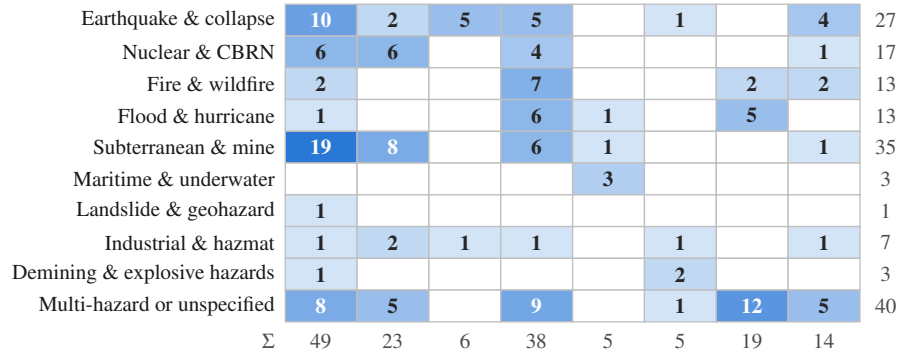
2.6 Cross-criteria analysis

Fig. 3 crosses the five criteria that assign every work to exactly one category. The other criteria are analyzed qualitatively in Sections 5-14. Four main observations emerge, namely: a)

(a) Response task × robotic platform



(b) Disaster domain × robotic platform



(c) Foundation model type × validation maturity

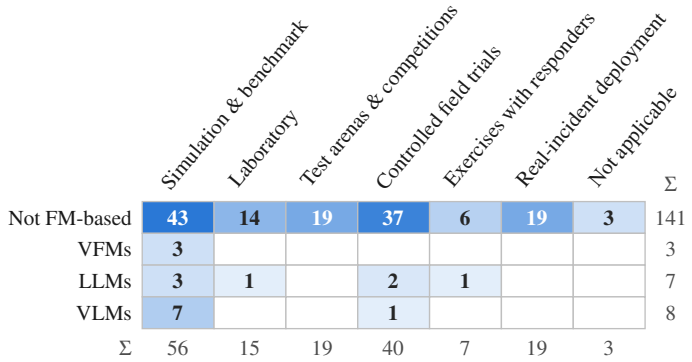


Fig. 3 Cross-criteria distribution of the 159 selected primary works, namely: a) Disaster response task versus robotic platform. b) Disaster domain versus robotic platform. c) Foundation model type versus validation maturity. Each work is counted once per panel, under its primary category along each criterion. As a result, the totals (Σ) sum to 159 and match Fig. 2. Abbreviations used: VFMs, vision foundation models; LLMs, large language models; VLMs, vision-language models

The corpus is concentrated, with 26 of the 56 task-platform and 41 of the 80 domain-platform combinations empty. b) Response tasks and disaster domains are bound to specific embodiments. For example, damage assessment is bound to satellite and airborne sensing (11 of 16 works), while maritime emergencies are bound to marine vehicles. c) The 18 foundation-model-based works are evaluated mostly in simulation or on benchmarks. None of them is evaluated in a standardized arena or at a real incident. d) Between the works up to 2022 and those of 2023-2026, simulation and benchmark evaluation rises from 23.7% to 42.0%. In parallel, real-incident deployment falls from 20.3% to 7.0%.

2.7 Limitations of the current review

Like any systematic review, the present one is shaped by the design choices behind its corpus, its taxonomy and the characterization of the works. These choices bound the scope of its findings. Three main limitations arise from them, concerning the nature of the analyzed evidence, the structure of the taxonomy and the composition of the corpus. Each limitation is stated below, together with its implication for the interpretation of the findings.

- Reliance on reported evidence. The analysis rests on heterogeneous reported evidence rather than on a re-evaluation under a common protocol. Hence, no ranking of methods or pooled statistical synthesis is attempted.
- Partially overlapping taxonomy criteria. The ten criteria are complementary, but not strictly orthogonal. Hence, each work receives one primary category per coded criterion. Secondary aspects, in turn, are discussed in the criterion sections, but they are not counted in Figs. 2 and 3.
- Curated and time-bounded corpus. The venue (IC5, EC3), the language and access (IC6) criteria, as well as the partial coverage of 2026, make the corpus criteria-driven rather than exhaustive. Therefore, the statistics indicate research activity trends rather than a census.

3 Disaster robotics research evolution

Academic research into rescue robotics started in 1995, after the Great Hanshin-Awaji earthquake and the Oklahoma City bombing (Murphy 2014). Since then, it has moved from tethered, teleoperated single platforms toward heterogeneous, largely autonomous teams. This evolution has been driven by real disasters, competitions, advances in computer vision and machine learning, as well as evidence gathered with responders. Fig. 4 and Table 3 organize this trajectory into five phases, by the year in which a system was introduced or deployed. However, an event series, such as the SubT circuits, may begin in the preceding phase. Additionally, the general-purpose works that enabled them are cited, namely for Phase 2 (Shen et al. 2011; Bachrach et al. 2012; Endres et al. 2014; Zhang and Singh 2014; Whelan et al. 2015), Phase 3 (Hutter et al. 2016; Gehring et al. 2021; Hwangbo et al. 2019; Mohta et al. 2018; Zhou et al. 2019; Papachristos et al. 2019; Mintchev et al. 2017; Cadena et al. 2016; McCormac et al. 2017; Bircher et al. 2016), Phase 4 (Rosinol et al. 2020; Tian et al. 2022; Huang et al. 2022; Zhao et al. 2021; Cao et al. 2021; Lee et al. 2020; Miki et al. 2022; Kumar et al. 2021; Loquercio et al. 2021; Song et al. 2021; Zhou et al. 2022; Schwarz et al. 2021) and Phase 5 (Kirillov et al. 2023; Firoozi et al. 2025; Brohan et al. 2023; Zitkovich et al. 2023; Kim et al. 2024; Black et al. 2025).



Fig. 4 Main phases in disaster robotics research and key/milestone works (UGV/UAV = uncrewed ground/aerial vehicle, sUAS = small uncrewed aerial system)

Table 3 Key systems, benchmarks, and methodological milestones in the evolution of disaster robotics research, limited to works directly connected to a disaster domain. UGV and UAV denote uncrewed ground and aerial vehicles, while SAR denotes synthetic aperture radar.

System / Work	Year	Phase	Focus; platform / method	Key contribution
RoboCup Rescue (Kitano et al. 1999; Kitano and Tadokoro 2001)	1999	1	Benchmark domain; simulation & ground	Framed large-scale disaster response as a grand challenge for multi-agent and intelligent-systems research.
NIST reference test arenas (Jacoff et al. 2003)	2000	1	Evaluation; standardized courses	Reproducible arenas and a metric scoring victim finding, mapping accuracy and operator count in rescue-robot competitions.
World Trade Center deployment analysis (Casper and Murphy 2003)	2001	1	Field deployment, human-robot interaction; teleoperated ground	First human-robot interaction study during a real urban search-and-rescue response, yielding 17 findings and 11 recommendations.
Souryu IV/V (Arai et al. 2008)	2006	1	In-rubble search; articulated crawler	Serially connected crawlers with metal-reinforced tracks, tested on rubble and step fields.
Quince at Fukushima (Nagatani et al. 2011, 2013)	2011	2	Nuclear reconnaissance; tracked ground	Redesigned and deployed inside damaged reactor buildings; radiation, tether and communication resilience as design requirements.
Marine ROV teams at Tōhoku (Murphy et al. 2012)	2011	2	Marine survey; underwater (ROV)	Port clearing and victim recovery; reopened a port and found wreckage in areas marked clear by divers.
Collaborative UGV-UAV mapping (Michael et al. 2012)	2011	2	Interior 3D mapping; ground + aerial	Joint 3D map of a real earthquake-damaged building, obtained via marsupial micro aerial vehicle (MAV) deployment.
DRC affordance-based autonomy (Fallon et al. 2015)	2013	2	Shared autonomy; humanoid	Affordance-templated perception and whole-body planning for intent-level command over low-bandwidth links.
DRC-HUBO+ / Team KAIST (Lim et al. 2017)	2015	2	Physical intervention; transformable humanoid	Winning DRC Finals entry, switching between walking and a knee-wheel mode with a larger support polygon.
Amatrice earthquake deployment (Kruijff-Korbayová et al. 2016)	2016	3	Field deployment; ground + aerial team	Ground-aerial survey of two damaged historical churches, delivering textured 3D models to the fire brigade.
Mexico City snake-robot search (Whitman et al. 2018)	2017	3	In-rubble search; articulated snake robot	Ruled out an open void as a survivor location, while exposing the limits of a researcher-operated system.
Hurricane Harvey sUAS operations (Fernandes et al. 2018)	2017	3	Field deployment; small uncrewed aerial systems	112 flights by nine models; first-person-view control essential early, data management unable to keep pace.
xBD (Gupta et al. 2019)	2019	3	Damage assessment; satellite imagery dataset	Multi-disaster pre/post building-damage dataset with a joint damage scale; basis of the xView 2.0 challenge.

Table 3 (continued)

System / Work	Year	Phase	Focus; platform / method	Key contribution
GBPlanner (Dang et al. 2019, 2020)	2019	3	Autonomous exploration; graph-based planning	Bifurcated local/global graph exploration planner tailored to confined subterranean geometry.
NeBula (Agha et al. 2022; Morrell et al. 2024)	2020	4	Risk-aware autonomy; heterogeneous team	Belief-space, risk-aware stack with multi-modal positioning; won the SubT Urban Circuit, second in the Tunnel Circuit.
CERBERUS (Tranzatto et al. 2022, 2024)	2021	4	Subterranean autonomy; legged + aerial team	Winning SubT Final Event system; complementary multi-modal SLAM and exploration under perceptual degradation.
Team CSIRO Data61 at SubT (Hudson et al. 2022; Kottege et al. 2025)	2021	4	Subterranean autonomy; ground + aerial team	Homogeneous sensing and a common navigation stack across a heterogeneous fleet; tied the top SubT Final Event score (Kottege et al. 2025).
German Rescue Robotics Center (DRZ) task-force lessons (Surmann et al. 2024)	2021	4	Field deployment; ground + aerial team	Lessons from three task-force missions exploring buildings damaged by fire and flood.
Nuclear response field exercise (Chiou et al. 2022)	2022	4	Responder exercise; teleoperated ground	Five-day exercise with a teleoperated mobile manipulator; operator workload the main limitation, variable autonomy recommended.
X-ICP (Tuna et al. 2024)	2023	5	Degradation-robust localization; LiDAR registration	Localizability-aware registration that detects ill-conditioned geometry in tunnels and featureless planar spaces.
Kuro Siwo (Bountos et al. 2024)	2024	5	Flood mapping; multi-temporal SAR benchmark	SAR dataset of 43 flood events (2015-2022) across six continents, released with a baseline benchmark.
DAVI (Ahn et al. 2025)	2025	5	Damage assessment; vision FM + change detection	Building damage in unlabeled target regions, combining a source-trained model with a segmentation foundation model.
SPINE (Ravichandran et al. 2025)	2025	5	Language-conditioned planning; LLM over semantic map	Online planning from incomplete natural-language missions over an open-vocabulary semantic map.
Fukushima radiation-tolerant remote system (Sakaue et al. 2025)	2025	5	Nuclear intervention; radiation-tolerant teleoperation	Controller kept in a low-dose area, remote I/O on the robot; used for zeolite accumulation in decommissioning.
DisasterM3 (Wang et al. 2025)	2025	5	Disaster vision-language; multi-modal benchmark	123k instruction pairs; the 14 evaluated VLMs struggle with disaster reasoning, while four VLMs fine-tuned on the dataset improve.

Phase 1: Teleoperated platforms and field-deployment lessons (pre-2011): Robots were mobile sensor carriers under continuous low-level control. They ranged from serpentine crawlers and active borescopes for rubble voids (Hirose and Fukushima 2004; Hatazaki et al. 2007; Arai et al. 2008) to marsupial teams (Murphy 2000). These systems were benchmarked in RoboCup Rescue and the arenas of the National Institute of Standards and Technology (NIST) (Kitano et al. 1999; Kitano and Tadokoro 2001; Jacoff et al. 2003). The World Trade Center response of 2001 and later deployments exposed operator errors, traced mainly to poor interfaces and the limits of human perception (Casper and Murphy 2003; Murphy 2004b,a; Murphy et al. 2009), as did arena competition studies (Yanco and Drury 2007). They also exposed mobility limits on unfamiliar terrain (Murphy and Stover 2008; Micire 2008). Moreover, deployments extended to surface-aerial infrastructure inspection (Murphy et al. 2008). In contrast, autonomy remained confined to mapping and victim detection in test arenas and outdoor trials (Kleiner and Dornhege 2007; Kleiner and Kümmerle 2007).

Phase 2: Field-proven mobility, mapping and humanoid disaster response (2011-2015): At Fukushima Daiichi, redesigned tracked robots performed interior reconnaissance (Nagatani et al. 2011, 2013). These deployments documented operational failures, as well as an operator who trusted neither full nor shared autonomy (Nagatani et al. 2013). In parallel, the tsunami prompted underwater surveys (Murphy et al. 2012), while a ground-aerial team mapped an earthquake-damaged building (Michael et al. 2012). The DARPA Robotics Challenge (DRC) yielded whole-body control, affordance-based shared autonomy and a winning walking-wheeled design (Pratt and Manzo 2013; Krotkov et al. 2017; Feng et al. 2015; Fallon et al. 2015; Kuindersma et al. 2016; Lim et al. 2017). However, it was also marked by frequent falls, slow task execution and operator-side bottlenecks (Johnson et al. 2015; Yanco et al. 2015; Norton et al. 2017). Furthermore, aerial autonomy for urban search and rescue without global navigation satellite system (GNSS) positioning matured (Tomić et al. 2012).

Phase 3: Deep-learning perception and field autonomy (2016-2019): Research shifted toward onboard perception and autonomy. Legged robots reached field applications (Bellicoso et al. 2018), while deep reinforcement learning was applied to rescue-robot exploration (Niroui et al. 2019). Moreover, distributed multi-UAV control addressed propagating hazards in simulation (Pham et al. 2017), and graph-based planners were field-tested in mines (Dang et al. 2019, 2020). In parallel, dedicated disaster datasets appeared (Božić-Štulić et al. 2019; Gupta et al. 2019; Kyrkou and Theocharides 2019), and deep networks outperformed a handcrafted detector on one of them (Božić-Štulić et al. 2019). At real disasters, however, robots remained largely teleoperated, as documented for the responses at Amatrice, Mexico City and Hurricane Harvey (Kruijff-Korbayová et al. 2016; Whitman et al. 2018; Fernandes et al. 2018). Moreover, teleoperation was the dominant approach for semi-autonomous field-ready robots (Delmerico et al. 2019).

Phase 4: Resilient multi-robot autonomy at field scale (2020-2022): The DARPA SubT Challenge integrated separate capabilities into field-scale heterogeneous teams operating under degraded perception and communication (Tranzatto et al. 2022; Agha et al. 2022; Scherer et al. 2022; Hudson et al. 2022; Rouček et al. 2022). Each team was permitted a single human supervisor (Orekhov and Chung 2022), whose load limited team performance in some cases (Chung et al. 2023). Along with these teams, outlier-resilient multi-robot simultaneous localization and mapping (SLAM), failure-tolerant odometry and large-volume exploration planning emerged (Ebadi et al. 2020; Chang et al. 2022; Reinke et al. 2022; Kulkarni et al.

2022). Meanwhile, the notion of a disaster broadened to public-health emergencies (Yang et al. 2020; Gao et al. 2021). Task forces also deployed robots at fires and floods (Surmann et al. 2024). Moreover, the Hurricane Ian response recorded human errors in most missions, as well as wireless demands that forced physical data transfer (Manzini et al. 2023b; Murphy et al. 2024; Manzini et al. 2023c). Finally, emergency marine robotics re-emerged in field trials (Rymansaib et al. 2023), together with surface vehicles for mapping flooded underground sites (Villarroel et al. 2024).

Phase 5: Foundation-model interfaces and responder-grounded field autonomy (2023-present): Foundation models entered as perception front ends and planners over existing stacks, rather than as embodied policies. They took the form of promptable segmentation for change and damage analysis (Zheng et al. 2024b; Ahn et al. 2025), overhead vision-language models (Irvin et al. 2025; Soni et al. 2025; Wang et al. 2025; Danish et al. 2025), as well as language-model planners, copilots and voice interfaces (Bark et al. 2025; Ravichandran et al. 2025; Shi et al. 2025; Cladera et al. 2025; Zhu et al. 2025; Jarabo-Peñas et al. 2025). In contrast, no vision-language-action policy has yet been applied to a disaster response task in the surveyed venues. In parallel, the SubT systems reached archival form (Tranzatto et al. 2024; Kottege et al. 2025; Petrlík et al. 2025; Dixit et al. 2025; Morrell et al. 2024; Ebadi et al. 2024), while disaster benchmarks multiplied (Bountos et al. 2024; Gupta et al. 2025; Tiwari et al. 2025; Huang et al. 2026; Manzini et al. 2026; Malizia et al. 2026). Furthermore, responder requirements, trust and fatigue (Daun and von Stryk 2023; Ruan et al. 2025; Mehta et al. 2024), field wireless links (Bravo-Arrabal et al. 2023), as well as the energy-aware planning of cellular-connected robot fleets (Romero et al. 2024), were studied. However, the operational record remains thin (Sakaue et al. 2025; Lichtenfeld et al. 2025; Manzini et al. 2026).

4 Key criteria and main categories of disaster robotics methods

This section provides a compact overview of contemporary disaster robotics, i.e., of the works of the current research phase (Phase 5 in Section 3, 2023-present). The overview is organized along ten complementary criteria, each addressing a key aspect of a method and yielding a set of main categories (Fig. 5). Criteria and categories are only defined in this section. In contrast, the in-depth analysis of the corresponding works is provided in Sections 5-14. Moreover, only works tied concretely to disaster response, in one of the senses listed in Table 2, are classified. The reason is that a taxonomy populated by general-purpose methods of merely asserted relevance would misrepresent the maturity of the field.

- **Disaster response task**: The operational objective that a method serves within the response effort. This objective motivates specific robot designs and shapes the sensor payload (Murphy et al. 2016). The main tasks addressed are: a) Reconnaissance and mapping: Building a geometric and semantic model of an unknown or altered site (Petrlík et al. 2025; Surmann et al. 2024; Ravichandran et al. 2025). b) Victim search and detection: Locating people from visual, thermal or acoustic evidence (Cao et al. 2024; Horyna et al. 2023; Gupta et al. 2025). c) Damage and structural assessment: Estimating the extent and severity of damage to buildings and roads (Ahn et al. 2025; Manzini et al. 2026; Huang et al. 2026;

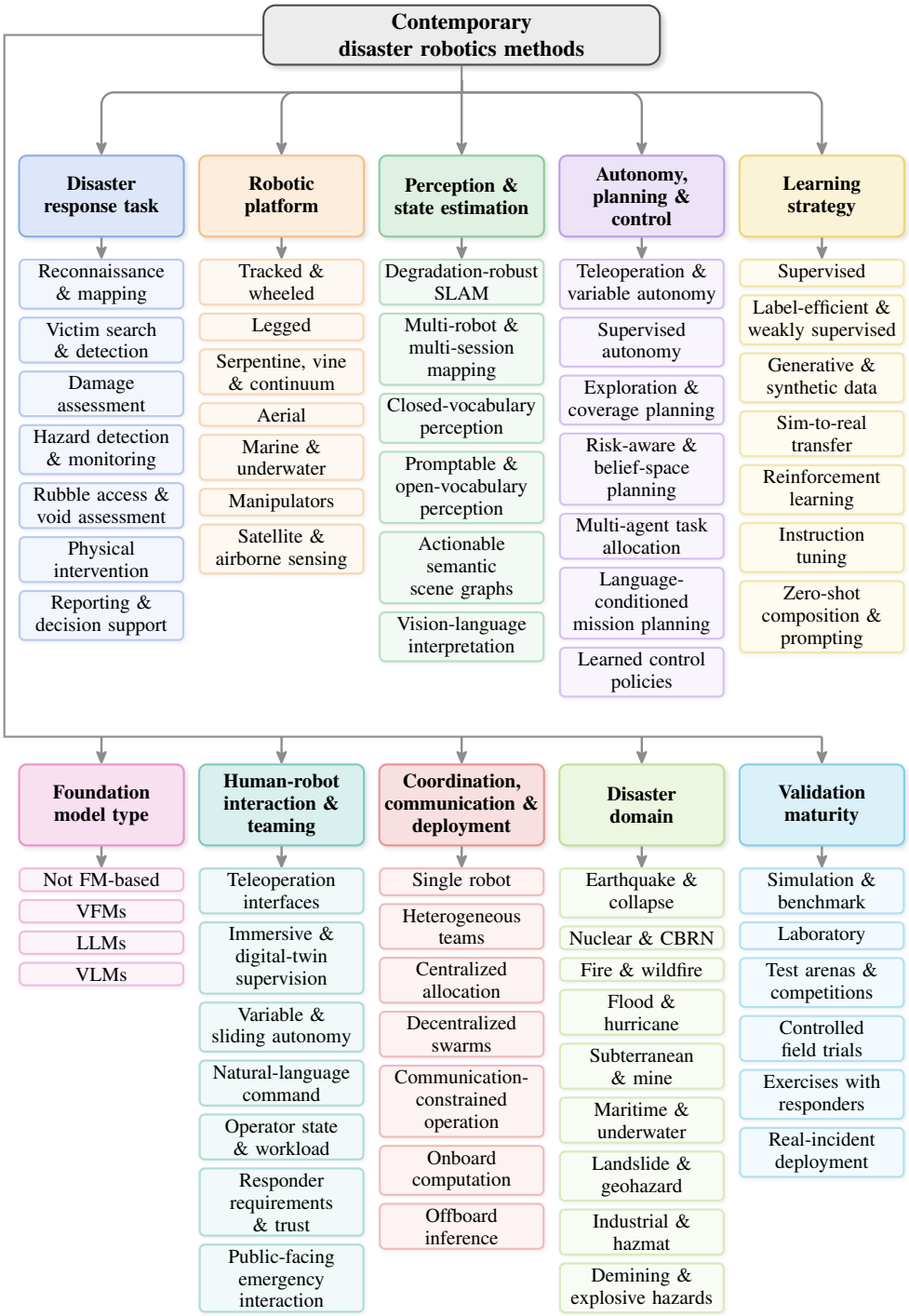


Fig. 5 Key criteria considered for analyzing the contemporary disaster robotics literature, together with the main method categories that result from them

Wang et al. 2025). d) Hazard detection and monitoring: Tracking fire, smoke, gas, radiation, flooding or concealed explosives (Fernández Talavera et al. 2023; Pal et al. 2025; Nordström et al. 2025; Mascarich et al. 2023; Rhodes et al. 2023; Bountos et al. 2024). e) Rubble access and void assessment: Reaching and characterizing the confined voids of a collapse (Bal et al. 2023; McFarland et al. 2024). f) Physical intervention and casualty handling: Acting upon the environment or upon the victim (Lichtenfeld et al. 2025; Hong et al. 2023; Chaoxia et al. 2025; Peiros et al. 2023). g) Reporting and decision support: Converting acquired data into the situational picture of incident command (Manzini et al. 2023c; Murphy et al. 2024; Ruan et al. 2025).

- **Robotic platform**: The embodiment, which determines mobility, carrying capacity and endurance. The platform is selected, together with the payload, according to the type of disaster (Wang et al. 2023; Murphy et al. 2016). The main platform categories are: a) Tracked and wheeled ground vehicles: Uncrewed ground vehicles (UGVs) favored for their robustness and payload capacity (Wang et al. 2023; Fernández Talavera et al. 2023; Sakaue et al. 2025). b) Legged platforms: Machines offering greater mobility over debris-strewn terrain and stairs (Kottege et al. 2025; Cruz Ulloa et al. 2025). c) Serpentine, vine and continuum platforms: Elongated or growing mechanisms that penetrate voids and confined conduits (McFarland et al. 2024; Lee et al. 2025). d) Aerial vehicles: Rotorcraft and reconfigurable airframes for wide-area search and confined entry (Manzini et al. 2023b; Petrlík et al. 2025; Brown et al. 2026; Schomer and Adams 2026). e) Marine and underwater vehicles: Surface vessels and underwater robots for water search and the mapping of flooded galleries (Rymansaib et al. 2023; Zhou et al. 2026; Villarroel et al. 2024). f) Manipulators and intervention systems: Arms and end effectors for contact with the environment or a casualty (Hong et al. 2023; Xiao et al. 2023; Peiros et al. 2023). g) Satellite and airborne sensing: Overhead platforms whose imagery is analyzed offboard (Bountos et al. 2024; Ahn et al. 2025; Wang et al. 2025).
- **Perception and state estimation**: How a system senses, localizes itself within and represents a damaged environment exposed to darkness, smoke and dust (Ebadi et al. 2024; Chen et al. 2026). The main methodological categories are: a) Degradation-robust localization and mapping: Estimation hardened against obscurants and self-similar geometry, including radar and thermal sensing (Ebadi et al. 2024; Tuna et al. 2024; Zeng et al. 2024; Dhrafani et al. 2025; Chen et al. 2026). b) Multi-robot and multi-session mapping: Merging, sharing or relocalizing maps across robots and mission phases (Kottege et al. 2025; Petrlík et al. 2025; Dahlquist et al. 2025; Bal et al. 2023; Morrell et al. 2024). c) Closed-vocabulary disaster perception: Supervised recognition over a label set fixed at training time (Salas-Espinales et al. 2023; Zhao et al. 2024a; Amessegher et al. 2025). d) Promptable and open-vocabulary perception: Recognition driven by point or text prompts and open-set detectors (Zheng et al. 2024b; Ahn et al. 2025; Cladera et al. 2025). e) Actionable semantic scene graphs: Relational maps annotated with affordances, in which planned actions are grounded (Bark et al. 2023, 2025). f) Vision-language disaster interpretation: Joint image-text description, questioning and

reasoning about disaster scenes (Wang et al. 2025; Irvin et al. 2025; Soni et al. 2025; Danish et al. 2025; Yamanaka et al. 2025; Cladera et al. 2025; Shi et al. 2025).

- **Autonomy, planning and control**: How decisions are produced and at which level of abstraction the human intervenes. The main paradigms are: a) Teleoperation and variable autonomy: Continuous or intermittent human control (Daun and von Stryk 2023; Mehta et al. 2024; Sakaue et al. 2025; Lichtenfeld et al. 2025). b) Supervised autonomy: A single operator directing a team at the task level over intermittent links (Biggie et al. 2023; Kottege et al. 2025). c) Exploration and coverage planning: Deciding where to go next in an unknown or partially known volume (Petrlík et al. 2025; Blake and Schubert 2025; Schomer and Adams 2026). d) Risk-aware traversability and belief-space planning: Reasoning over terrain risk, hazards and localization uncertainty (Dixit et al. 2025; Ginting et al. 2023; Morrell et al. 2024). e) Multi-agent task allocation: Distributing work across a team as tasks emerge and budgets bind (O'Brien et al. 2023; Maresca et al. 2025; Dahlquist et al. 2025; Hickling et al. 2026). f) Language-conditioned mission planning: Large models decomposing natural-language missions into grounded, checked actions (Ravichandran et al. 2025; Shi et al. 2025; Bark et al. 2025; Cladera et al. 2025; Zhu et al. 2025). g) Learned control and decision policies: Policies trained by reinforcement learning for dynamics that resist hand design (Chaoxia et al. 2025; Blake and Schubert 2025; Hickling et al. 2026).
- **Learning strategy**: How competence is acquired. In this respect, real disaster data are scarce and costly, while deployment differs from training (Huang et al. 2026; Ahn et al. 2025). The strategies observed are: a) Supervised learning: Training on labeled disaster imagery or sensor data (Manzini et al. 2026; Huang et al. 2026; Salas-Espinales et al. 2023). b) Label-efficient and weakly supervised learning: Reducing annotation through physical priors, unpaired labels or pseudo labels (Sami et al. 2024; Zheng et al. 2024a; Ahn et al. 2025). c) Generative and synthetic data production: Synthesis of labeled images, change pairs or hazard-changing simulated environments (Zheng et al. 2025; Chaoxia et al. 2025; Zhou et al. 2024). d) Sim-to-real transfer: Training in simulation and closing the reality gap (Chaoxia et al. 2025; Hickling et al. 2026; Pal et al. 2025; Tranzatto et al. 2024). e) Reinforcement learning: Learning through interaction against a reward signal (Hickling et al. 2026; Cao et al. 2024; Blake and Schubert 2025; Chaoxia et al. 2025; Zhou et al. 2024). f) Instruction tuning of vision-language models: Specialization on Earth-observation or disaster instruction corpora (Soni et al. 2025; Irvin et al. 2025; Wang et al. 2025; Revankar et al. 2025; Chen et al. 2025). g) Zero-shot composition and prompting: Assembling a capability from frozen pre-trained models (Zheng et al. 2024b; Yamanaka et al. 2025).
- **Foundation model type**: The large pre-trained model around which a method is organized. Its modalities shape which part of the autonomy stack it can address (Firoozi et al. 2025). The types encountered are: a) Not foundation-model-based: Classical estimation and planning, task-specific models or teleoperation (Surmann et al. 2024; Lichtenfeld et al. 2025; Sakaue et al. 2025; Manzini et al. 2026; Murphy et al. 2024; Manzini et al. 2023b,c). b) Vision foundation models (VFMs): General-purpose or promptable visual models, such as the Segment Anything Model (SAM) (Kirillov et al. 2023), used as label-free front ends or as generators of synthetic training data (Zheng et al. 2024b;

Ahn et al. 2025; Zheng et al. 2025). c) Large language models (LLMs): Models translating instructions into plans, task decompositions or behavior-tree calls (Ravichandran et al. 2025; Bark et al. 2025; Jarabo-Peñas et al. 2025; Zhu et al. 2025; Shi et al. 2025). d) Vision-language models (VLMs): Models jointly exploiting images and text for grounding, question answering and scene interpretation (Irvin et al. 2025; Danish et al. 2025; Yamanaka et al. 2025; Wang et al. 2025; Tiwari et al. 2025).

- **Human-robot interaction and teaming**: The modality through which humans and robots exchange intent and information. It forms a first-class criterion, since human error, fatigue and workload are a major source of mission failure (Murphy et al. 2024; Mehta et al. 2024; Biggie et al. 2023). The main modalities are: a) Teleoperation interfaces: Immersive or conventional displays and controls for remote operation (Stedman et al. 2023; Illing et al. 2025). b) Immersive and digital-twin supervision: Robot fleets coupled to remote supervisors through synchronized virtual replicas (Baniqued et al. 2024). c) Variable and sliding autonomy: Authority shifted between operator and system at run time (Daun and von Stryk 2023; Biggie et al. 2023). d) Natural-language command: Spoken or written instruction of a robot or a team (Jarabo-Peñas et al. 2025; Bark et al. 2025; Cladera et al. 2025). e) Operator state, workload and error: Measurement of workload, trust, awareness and fatigue, as well as of the errors that occur in operation (Orlov Savko et al. 2024; Senaratne et al. 2026; Mehta et al. 2024; Murphy et al. 2024). f) Responder requirements and trust: Elicitation of what emergency personnel need from robots before trusting them (Ruan et al. 2025; Daun and von Stryk 2023; Kleiser et al. 2026). g) Public-facing emergency interaction: Interaction with people other than the operators, such as evacuees or responders at road incidents (Hu et al. 2023; Lee et al. 2023).
- **Coordination, communication and computational deployment**: How many robots act together, how they communicate and where computation runs. Disaster sites, in this respect, can lose power and communication infrastructure (Surmann et al. 2024; Manzini et al. 2023c; Steinbauer-Wagner et al. 2025). The main categories are: a) Single-robot operation: One platform acting alone (Fernández Talavera et al. 2023; Sakaue et al. 2025; Brown et al. 2026). b) Heterogeneous teams: Ground and aerial platforms exploiting complementary mobility and viewpoints (Kottege et al. 2025; Tranzatto et al. 2024; Cladera et al. 2025; Surmann et al. 2024). c) Centralized coordination and allocation: Team-level planning at one decision point, such as an edge server or a lead robot (Maresca et al. 2025; Romero et al. 2024; Werner et al. 2024). d) Decentralized swarms: Robots planning locally in peer-to-peer schemes, even with sparse or no radio links (O'Brien et al. 2023; Dah-Achinanon et al. 2023; Horyna et al. 2023). e) Communication-constrained operation: Systems and field measurements for intermittent or absent links (O'Brien et al. 2023; Dah-Achinanon et al. 2023; Bravo-Arrabal et al. 2023; Manzini et al. 2023c). f) Onboard computation: Perception and autonomy under the platform's size, weight, power and latency budgets (Amessegher et al. 2025; Petrlík et al. 2025; Steinbauer-Wagner et al. 2025). g) Offboard and cloud-dependent inference: Computation offline or on remote servers. It assumes a link that the site may not provide (Jarabo-Peñas et al. 2025; Yamanaka et al. 2025; Cladera et al. 2025).

- **Disaster domain**: The hazard environment, which shapes sensing, mobility and time constraints. Moreover, it conditions the transferability of results between studies (Murphy et al. 2016; Wang et al. 2025). The main domains are: a) Earthquake and structural collapse: Unstable rubble, confined voids and unknown post-event geometry (Daun and von Stryk 2023; Bal et al. 2023; McFarland et al. 2024; Huang et al. 2026). b) Nuclear, radiological and CBRN incidents: Hazards imperceptible to humans that degrade onboard electronics. CBRN denotes chemical, biological, radiological and nuclear hazards (Sakaue et al. 2025; Werner et al. 2024; Mascarich et al. 2023). c) Fire and wildfire: Propagating hazards that transform the scene (Zhou et al. 2024; Pal et al. 2025; Tiwari et al. 2025; Fernández Talavera et al. 2023; Frering et al. 2023; Chaoxia et al. 2025). d) Flood and hurricane response: Large-area events relying on aerial survey and rapid flood mapping (Manzini et al. 2023b; Bountos et al. 2024; Sami et al. 2024). e) Subterranean and mine incidents: Tunnels, caves, mines and underground structures (Chung et al. 2023; Petrlík et al. 2025; Tranzatto et al. 2024; Nordström et al. 2025; Brown et al. 2026; Villarroel et al. 2024). f) Maritime and underwater emergencies: Open-water search, submerged survey and recovery (Zhou et al. 2026; Rymansaib et al. 2023). g) Landslide and geohazard monitoring: Unstable ground and volcanic eruptions (Murphy and Stover 2008; Wang et al. 2025). h) Industrial and hazmat emergencies: Plant fires, chemical releases and gas leaks (Steinbauer-Wagner et al. 2025; Rhodes et al. 2023; Lichtenfeld et al. 2025). i) Humanitarian demining and explosive hazards: Detection and disposal of landmines and explosive ordnance (Malizia et al. 2026; Xiao et al. 2023).
- **Validation maturity**: The strength of the evidence supporting a method. It forms an explicit criterion, since the distance between demonstrated capability and operational practice recurs in the field (Daun and von Stryk 2023; Lichtenfeld et al. 2025; Steinbauer-Wagner et al. 2025). The identified levels are: a) Simulation and benchmark evaluation: Assessment confined to simulated environments or offline datasets (Zhou et al. 2024; Wang et al. 2025; Blake and Schubert 2025; Maresca et al. 2025). b) Laboratory conditions: Physical experiments in controlled, mostly indoor settings (Peiros et al. 2023; Mascarich et al. 2023; Chaoxia et al. 2025). c) Standardized test arenas and competitions: Reproducible courses supplying comparable metrics (Kottege et al. 2025; Hickling et al. 2026; Cruz Ulloa et al. 2025; Chung et al. 2023). d) Controlled field trials: Real unstructured environments without an emergency in progress (Dahlquist et al. 2025; Brown et al. 2026; Rymansaib et al. 2023). e) Exercises with responders: Trials conducted jointly with operational response organizations (Mehta et al. 2024; Frering et al. 2023; Jarabo-Peñas et al. 2025; Steinbauer-Wagner et al. 2025). f) Deployment at a real incident: Use at an actual emergency, during the response or its recovery (Surmann et al. 2024; Manzini et al. 2023b,c; Murphy et al. 2024; Lichtenfeld et al. 2025; Sakaue et al. 2025; Manzini et al. 2026).

5 Disaster response tasks

Disaster robotics methods are grouped in this section by their purpose in the emergency operation. This purpose determines the sensing modality carrying the decisive evidence and the

time scale of a useful answer. Moreover, the size of a robot affects both the tasks it suits and how soon after a disaster it can be used (Murphy et al. 2016). The main tasks, as defined in Section 4, are: a) Reconnaissance and mapping. b) Victim search and detection. c) Damage and structural assessment. d) Hazard detection and monitoring. e) Rubble access and void assessment. f) Physical intervention and casualty handling. g) Reporting and decision support.

5.1 Reconnaissance and mapping

Reconnaissance and mapping, a prerequisite of search and rescue, provides responders with situational awareness of places that may be life-threatening to enter (Ebadi et al. 2024; Petrlík et al. 2025). This awareness is obtained through site models acquired by exploration. These models are robust to degraded sensing or restricted access, merged across robots and sorties, and semantically enriched.

According to the representation produced and its sensing regime, the main categories are:

- Degradation-robust metric localization and mapping: Mapping that remains valid without satellite positioning, under the darkness, dust and smoke of subterranean sites (Ebadi et al. 2024), as well as under their geometric self-similarity, without markers or site-specific tuning (Tuna et al. 2024). The achieved accuracies reach decimeters for autonomous tunnel navigation (Zeng et al. 2024), while the SubT Challenge required errors within 5 m per reported survivor or object across kilometers of underground networks (Ebadi et al. 2024).
- Multi-robot and multi-session map construction: Merging the partial maps of several platforms or sorties under bandwidth limits and drift. In this way, detections reach the operator within a one-hour SubT run (Tranzatto et al. 2024; Kottege et al. 2025; Morrell et al. 2024). In contrast, inconsistent shared maps can make a robot reject an unvisited goal, while drift can void a correct detection (Petrlík et al. 2025).
- Exploration and coverage planning as the acquisition strategy: The map as a by-product of choosing where to go next. Approaches include supervised exploration (Biggie et al. 2023), frontier exploration merged with an operator's other tasks (O'Brien et al. 2023), two-drone (Hickling et al. 2026) and emergency-specific coverage (Blake and Schubert 2025), as well as borehole-inserted drone exploration of mine voids (Brown et al. 2026). In contrast, auction-based inspection in a pre-scanned mine executes rather than acquires the map (Dahlquist et al. 2025).
- Semantically annotated and queryable reconnaissance: Object-centric, relational or language-addressable site models. These include affordance-annotated maps for changing mission goals (Bark et al. 2023), semantic belief graphs of unknown sites with rubble, stairs and narrow passages (Ginting et al. 2023), as well as intent-level language missions of kilometer-scale air-ground teams (Cladera et al. 2025). Besides, a prompted multi-modal model inspects mine tunnels without a map (Yamanaka et al. 2025).
- Reconnaissance in access-limited media: Sites that exclude the conventional platform or its radio link. Examples are partially flooded underground galleries surveyed from the water surface (Villarroel et al. 2024), as well as shallow urban waters mapped with side-scan sonar (Rymansaib et al. 2023). Further examples are underground buildings mapped by two robots relaying through multi-node networking (Wang et al. 2023), and

smoke-filled buildings closed to rescuers by potentially toxic gases (Cruz Ulloa et al. 2025).

5.2 Victim search and detection

Victim search and detection works under occlusion and low visibility, where vision-based methods are prone to fail, as well as under time pressure, where any delay can have dramatic consequences (Gupta et al. 2025; Recchiuto and Sgorbissa 2018). Its methods combine visual, thermal and acoustic evidence with probabilistic, multi-agent and canine-informed search.

According to the evidence discriminating a human from clutter and the level at which the search is decided, the main categories are:

- Visual and thermal detection: Learned detection and segmentation on platform imagery, with first-responder and civilian classes (Salas-Espinales et al. 2023). Onboard detectors relieve operators from scanning video, yet they miss a person lying on the ground absent from the training data (Amessegher et al. 2025). Long-wave infrared detection of crews near prescribed burns also belongs to this category, while detectors trained without fire data often classify fire as people (Jong et al. 2023).
- Acoustic and non-visual evidence: Cues surviving the occlusion that defeats vision, such as speech, screams, cries, knocking or clapping. The drone's own rotor noise masks these cues, which a purpose-built search-and-rescue audio corpus addresses (Gupta et al. 2025).
- Probabilistic search planning: Sensing effort allocated over a belief about victim location. It serves volunteer mountain teams whose consumer-grade drones lack wide-area search autonomy (Schomer and Adams 2026), as well as maritime accident sites ill-served by predefined search paths (Zhou et al. 2026). In contrast to belief-based allocation, language-model reasoning targets people possibly behind closed doors by adding door-opening and obstacle-clearing subtasks, which enlarges the indoor search scope over traditional methods (Shi et al. 2025).
- Multi-agent and swarm search: Search distributed across a team, for targets whose unpredictable movement defeats repeated coverage sweeps (Cao et al. 2024). Search is also distributed across swarms that verify a candidate object (Horyna et al. 2023) or relay a found target to the base station (Dah-Achinanon et al. 2023). Moreover, it is distributed through energy-aware exploration of fire- or earthquake-damaged factories (Maresca et al. 2025).
- Human- and canine-derived search behavior: Robot roles derived from the practice they should complement. In this respect, four interviewed search-and-rescue dog handlers judged robots unable to replace dogs for lack of olfaction, while they favored low-attention signaling that does not distract the dogs (Kleiser et al. 2026).

5.3 Damage and structural assessment

Damage and structural assessment establishes the state of accessibility of the stricken area, with a value that depends on timely delivery (Recchiuto and Sgorbissa 2018; Manzini et al. 2026). Its methods range from wide-area change detection to per-structure grading. In parallel, promptable, generative and vision-language models relax fixed-label supervision.

According to the spatial scale and the supervision regime of the underlying model, the main categories are:

- Wide-area overhead change assessment: Pre- and post-event overhead imagery compared over an entire affected area. This includes building damage detection in unseen regions without their labels, as for the 2023 Türkiye earthquake (Ahn et al. 2025). It also includes fast change detectors trained with single-temporal supervision for building damage or road flooding (Zheng et al. 2024a).
- Object-level assessment of buildings and infrastructure: Damage resolved per structure or road segment. Pixel-level grading is applied to ground-level photographs posted within minutes of an earthquake, which show street-scale damage that aerial images miss (Huang et al. 2026). Moreover, road flooding, obstruction and destruction labels are obtained from small uncrewed aerial systems (sUAS) for routing aid and evacuation (Manzini et al. 2026).
- Promptable and zero-shot change assessment: Prompt-driven rather than fixed-label assessment, suited to disaster scenes containing categories that no training set anticipated. An example is a training-free adaptation of the Segment Anything Model that detects unseen change types through point queries (Zheng et al. 2024b).
- Generative supervision for change assessment: Synthesis of the multi-temporal change pairs that are costly to collect and annotate. A generative model produces such pairs, while its pre-trained detectors transfer to building damage assessment (Zheng et al. 2025). This model extends a stochastic change simulation first evaluated on generic building change detection (Zheng et al. 2023).
- Vision-language damage reasoning: Assessment posed as question answering or description. The tasks include rescue-place localization, restoration advice and long-form reports (Wang et al. 2025). In addition, Earth-observation assistants locate buildings by damage type or grade a user-marked region (Irvin et al. 2025; Soni et al. 2025). Furthermore, benchmarks ask which disaster caused the damage or how many buildings are damaged (Danish et al. 2025).

5.4 Hazard detection and monitoring

Hazard detection and monitoring sets exclusion zones and safe entry windows for hazards that may escape the senses of responders, as ionizing radiation does (Nordström et al. 2025; Werner et al. 2024). Its methods cover radiological, chemical, fire, flood and concealed explosive hazards. The signatures of these hazards range from static to spreading during observation.

According to the physical nature and spatio-temporal behavior of the hazard signature, the main categories are:

- Radiological sources and fields: Dose-field mapping and source localization from sparse, stochastic counts. The approaches comprise aerial spectroscopic mapping (Mascarich et al. 2023) and receding-horizon interception of moving targets to detect transient radioactivity (Yadav et al. 2023). Moreover, several sources are localized in real time without long gradient- or contour-tracing measurements (Werner et al. 2024). Finally,

alpha sampling is simulated during a mock initial entry into an undocumented nuclear facility (Stedman et al. 2023).

- Chemical, gaseous and airborne releases: Detection, measurement and source localization of released substances. Examples include post-blast gas monitoring in underground mines (Nordström et al. 2025) and a sniffer robot sparing firefighters manual measurements inside the danger zone (Steinbauer-Wagner et al. 2025). Additionally, the origin of an urban airborne release is estimated, whereas manual sensor probes obtain it only at high risk (Rhodes et al. 2023).
- Fire, smoke and thermal hazards: Hazards that may move during observation. They comprise smoke plumes under shifting winds (Pal et al. 2025) and wildfires under aerial surveillance (Zhao et al. 2024a). Moreover, they include hotspots that persist after surface fires are extinguished, which can spread over several hectares (Frering et al. 2023). Indoor fire sources and cold smoke (Fernández Talavera et al. 2023), as well as degraded 360-degree video of firefighting training (Tiwari et al. 2025), are also addressed.
- Areal inundation and flood extent: The spatial footprint of water delineated from overhead observation. This is achieved through synthetic aperture radar (SAR) time series of 43 flood events worldwide for rapid flood mapping (Bountos et al. 2024), as well as through elevation-guided flood-extent estimation (Sami et al. 2024).
- Concealed explosive hazards: Static but hidden hazards. They are addressed by multisensor data for landmine detection among vegetation and leaf litter (Malizia et al. 2026), as well as by explosive-residue detection on a teleoperated disposal robot (Xiao et al. 2023).

5.5 Rubble access and void assessment

The confined voids of a collapse, which have the highest chance of containing trapped survivors, are entered by robots when too small or too dangerous for rescuers (Bal et al. 2023; Murphy et al. 2016). The corresponding methods detect voids from outside, or enter them with growing devices or mobile platforms. In addition, they derive access requirements from rescue crews.

According to how the confined space is reached and characterized, the main categories are:

- Standoff void detection: Voids inferred from three-dimensional reconstructions of the rubble surface without entering them. An example is the ex post facto photogrammetric analysis of the Surfside collapse, which detected nine of ten expert-observed voids (Bal et al. 2023).
- Growing and continuum entry mechanisms: Elongating devices that advance without requiring purchase against the rubble. These comprise everted vine robots evaluated against the portability and deployment constraints of field teams (McFarland et al. 2024). They also comprise helical soft growing structures reaching spaces where long ventilation hoses are hard to use (Lee et al. 2025).
- Mobile access platforms: Ground platforms driven into the accessible parts of a damaged structure. An example is a quadruped crossing floor debris and detouring through pipes where collapses block the passages (Cruz Ulloa et al. 2025).

- Field-derived requirements for access: A taxonomy of void-entry difficulty derived from the challenges of rescue crews, with design changes prompted by the first trial and by responder feedback (McFarland et al. 2024).

5.6 Physical intervention and casualty handling

Physical intervention demands contact under uncertainty, while manipulation at deployments, always teleoperated, failed in at least one task per mission (Murphy et al. 2016). The corresponding methods span heavy manipulation of the environment, telepresence handling of hazardous objects and learned fire suppression. They also span contact with casualties, as well as long-horizon work in contaminated facilities.

According to the target of the physical action and the human authority retained over it, the main categories are:

- Environment manipulation and heavy intervention: Extraction of patients or soldiers with impaired mobility, lifting more than 120 kg, as well as explosive ordnance disposal (EOD) in place of human personnel (Hong et al. 2023). Moreover, a teleoperated vehicle opened a critical valve to avert a chemical-plant explosion (Lichtenfeld et al. 2025). In addition, hazardous gas is suctioned from otherwise inaccessible volumes (Lee et al. 2025).
- Hazardous-object handling under telepresence: The operator retained in the loop with perception extended to the contact itself. An example is approaching concealed ordnance that only touch can reach, with unintrusive contact to avoid detonation (Xiao et al. 2023). Another one is retrieving a biohazard-labeled bottle from clutter under immersive and screen-based control (Illing et al. 2025).
- Suppression as a learned control problem: An intervention policy learned rather than left to remote control. An example is an extinguishing strategy transferred from simulation to a firefighting robot, which guides the water jet toward the flames (Chaoxia et al. 2025).
- Contact with the human body: Casualty manipulation bound by biomechanical safety constraints, as in repositioning an unconscious casualty's limbs for extraction (Peiros et al. 2023).
- Sustained remote intervention in contaminated facilities: Unavoidably long-horizon remote work. An example is the suction of 26 tons of submerged zeolite sandbags, adsorbents for radioactive materials, on the underground floors of the Fukushima Daiichi plant, estimated to take about 200 days (Sakaue et al. 2025).

5.7 Reporting and decision support

Reporting and decision support is the interface between robots and the response organization. Its value depends on delivering information, rather than raw sensor data, to decision makers in time (Manzini et al. 2023c; Adams et al. 2024). The corresponding methods cover data products, responder acceptance, operator workload and error, language-mediated command, as well as shared situational pictures.

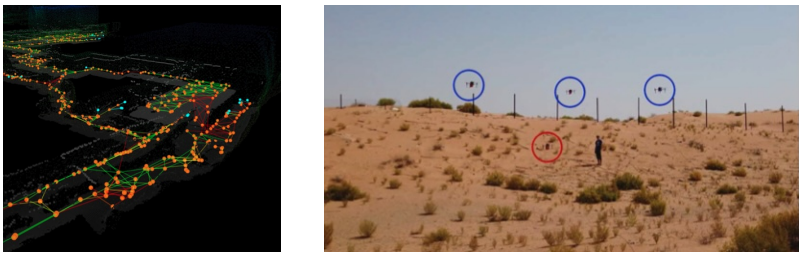
According to the addressed stage of the data-to-decision pipeline, the main categories are:

Table 4 Disaster response tasks: Comparative analysis and key insights.

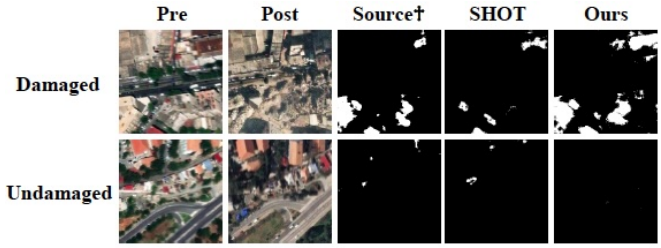
Response task	Reconnaissance & mapping	Victim search & detection	Damage & structural assessment	Hazard detection & monitoring	Rubble access & void assessment	Physical intervention	Reporting & decision support
Primary function	• Build a valid geometric and semantic model of a changed site • Inform entry into hazardous areas	• Localize trapped or moving people under occlusion • Allocate sensing effort over a victim-location belief	• Gauge how extensive and severe built-environment damage is • Prioritize response and access routes	• Detect, localize and track the agent that makes a site dangerous • Define exclusion zones and entry windows	• Reach and characterize confined voids in rubble • Point to where trapped survivors are likely	• Physically alter the environment or the victim's condition • Suppression, manipulation, disposal, extraction	• Convert acquired data into the incident-command picture • Mediate responder intent and machine output
Operational time scale	• Minutes to hours per sortie • Hour-long scored runs in the SubT Final Event	• Seconds onboard; hours of search • Bounded by victim survivability	• Hours to days • Repeat passes over the event	• Seconds to hours • Continuous while the hazard propagates	• Minutes per void entry • Serialized by physical entry	• Minutes to months • Decommissioning horizons in contaminated plants	• Operational tempo of the incident • Shift-length supervision windows
Dominant evidence source	• LiDAR, radar, visual-inertial range data • Cross-platform map fusion	• Visual, thermal and acoustic cues • Prior search-area models	• Pre/post-event overhead imagery • Text-paired disaster corpora	• Radiation counts, gas, thermal, water extent • Sparse, stochastic, non-visual signatures	• Tip-camera video from inside the void • Field-elicited access requirements	• Contact forces, tactile and chemical sensing • Operator haptic and immersive feedback	• Outputs of all other tasks • Responder studies and operator-state measurement
Input	• Onboard range and inertial streams, partial maps from teammates	• Onboard RGB, infrared and audio, terrain and drift priors	• Satellite, airborne, sUAS and ground imagery, instructions and questions in text	• Specialized hazard sensors, overhead imagery for areal hazards	• Aerial images of the rubble surface, body pressure and tip-camera feedback	• Object and body pose, operator commands, simulated interaction	• Maps, detections, damage and hazard products, spoken or written responder instruction
Output	• Metric, semantic or queryable site model	• Victim detections and localized search plan	• Per-pixel or per-object damage grades and reports	• Hazard maps, source estimates, flood extents	• Candidate voids and in-void views for responders	• Executed physical action on environment or casualty	• Situational picture, alerts and grounded mission plans

Table 4 (continued)

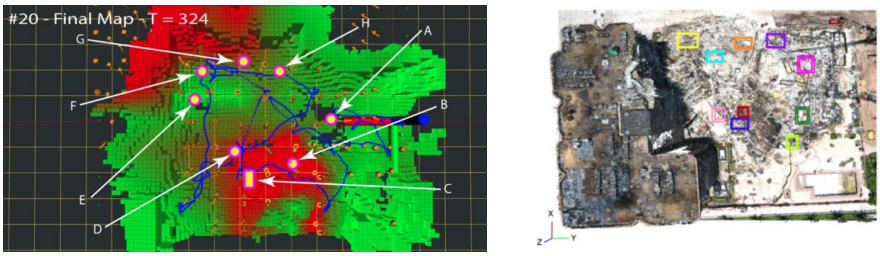
Response task	Reconnaissance & mapping	Victim search & detection	Damage & structural assessment	Hazard detection & monitoring	Rubble access & void assessment	Physical intervention	Reporting & decision support
Key strengths	• Most mature; validated at competition and field scale • Spatial frame for localizing victims and artifacts	• Directly aligned with the operational objective • Scales through multi-agent and swarm search	• Relatively latency-tolerant, so offboard computation is legitimate • Highest share of benchmark-based evaluation	• Addresses hazards inaccessible to human senses • Couples estimation with informative-path planning	• Access to spaces few other platforms can enter • Novel growing and continuum morphologies	• High potential operational impact per action • Removes responders from direct hazard exposure	• Governs how promptly robot data reach incident command • Grounded in real responder evidence
Critical limitations	• Degrades where geometry is ill-conditioned • Semantic layers not yet tested at real incidents	• Victim-focused real-incident reports limited to 2001-2005 deployments • Acoustic and occluded-victim evidence still data-poor	• Presupposes offboard links often unavailable • Current vision-language models struggle with disaster tasks	• Sparse, noisy signatures; slow information gain • Hazard-specific sensors and estimation models	• Very small corpus; few field-validated systems • Portability and deployment logistics limit field use	• Among the least mature; mostly arena- and lab-bound • No dataset surrogate for contact evaluation	• Few dedicated methods despite high leverage • Trust conditional on semantic accuracy
Typical validation maturity	• Controlled field trials; benchmarks, simulation and competitions • Occasional real incidents	• Simulation and controlled field trials; early real incidents	• Benchmark evaluation, offline; occasional real-incident use	• Benchmarks and simulation; some field trials and exercises	• Laboratory and controlled field trials; an isolated real deployment	• Test arenas and laboratory; isolated real incidents	• Simulation; real incidents and responder exercises
Indicative works	• X-ICP (Tuna et al. 2024), CERBERUS (Tranzatto et al. 2024), team CSIRO Data61 (Kottege et al. 2025)	• HMA-SAR (Cao et al. 2024), swarm search and rescue (Horyna et al. 2023), DroneAudioSet (Gupta et al. 2025)	• DAVI (Ahn et al. 2025), EIDSeg (Huang et al. 2026), DisasterM3 (Wang et al. 2025)	• Aerial radiation mapping (Mascarich et al. 2023), smoke-plume tracking (Pal et al. 2025), Kuro Siwo (Bountos et al. 2024)	• Void detection (Bal et al. 2023), portable vine robots (McFarland et al. 2024)	• HURCULES (Hong et al. 2023), EOD telepresence (Xiao et al. 2023), sim-to-real extinguishing (Chaoxia et al. 2025)	• Hurricane Ian data delivery (Manzini et al. 2023c), responder thresholds (Ruan et al. 2025), LLM copilot (Bark et al. 2025)



(a) (b)



(c)



(d) (e)



(f) (g)

Fig. 6 Representative literature methods per disaster response task: a) Reconnaissance and mapping (Kottege et al. 2025). b) Victim search and detection (Horyna et al. 2023). c) Damage and structural assessment (Ahn et al. 2025). d) Hazard detection and monitoring (Mascarich et al. 2023). e) Rubble access and void assessment (Bal et al. 2023). f) Physical intervention and casualty handling (Hong et al. 2023). g) Reporting and decision support (Jarabo-Peñas et al. 2025)

- Field data-product generation and delivery: What robots actually deliver at an incident. At Hurricane Ian, sorties of the Center for Robot-Assisted Search and Rescue (CRASAR) delivered a few representative images or short videos, while coverage surveys requested by incident command were stitched in the field into low-resolution orthomosaics (Manzini et al. 2023b). Moreover, data were carried physically to the nearest intact emergency operations center, despite six wireless networks (Manzini et al. 2023c). Live air observation with georeferenced orthophotos and point clouds was also provided (Surmann et al. 2024).
- Semantic situational awareness and responder acceptance: Which machine-produced semantics responders find usable and at what reliability. For instance, responders required a mean accuracy of 74.6% to trust semantic outputs, and of 67.8% to consider them useful (Ruan et al. 2025). Additionally, assistance functions for reconnaissance missions were derived from deployment reports and responder workshops (Daun and von Stryk 2023).
- Operator state, workload and human error: The cost of supervision and the errors that occur in practice. This category spans human-error analysis of disaster flight operations (Murphy et al. 2024) and team vulnerabilities under fatigue (Mehta et al. 2024), as well as datasets of human behavior and cognitive state in simulated response tasks (Orlov Savko et al. 2024).
- Language-mediated command and reporting: A large model placed between the responder and the system. Such a model turns natural-language task descriptions into grounded plans for remote urban search-and-rescue operation (Bark et al. 2025). It also decides, upon a voice command, which responder team to supply with material during six parallel rescue missions (Jarabo-Peñas et al. 2025). Moreover, it infers the details of high-level missions while mapping (Ravichandran et al. 2025), or creates team tasks from features detected online (Zhu et al. 2025).
- Shared situational pictures and team-level supervision: A heterogeneous fleet presented to a single supervisor as one coherent picture, through an immersive digital twin of a nuclear robot fleet (Baniqued et al. 2024).

5.8 Comparative analysis and key insights

Table 4 compares the seven tasks. These separate into information-producing ones (reconnaissance, victim search, damage assessment, hazard monitoring) and physically committing ones (rubble access, intervention). Reporting and decision support, in turn, forms the interface to the response organization. Representative methods per task are illustrated in Fig. 6. The main insights are that: a) Benchmark availability follows the evidence modality rather than operational priority. Overhead imagery is richly benchmarked. In contrast, acoustic data remain scarce, while radiological and contact data are absent. b) Foundation models concentrate in damage assessment, with few in the remaining tasks and none in rubble access. c) Reporting carries high operational leverage. However, its value is bounded by how promptly data reach decision makers.

6 Robotic platforms and morphologies

Disaster robotics methods are also grouped by the embodiment through which they sense and act at the disaster site. This is because the modality and size of a platform shape its design and capabilities, while the type of disaster influences the platform used (Murphy et al. 2016). The main platform categories, as defined in Section 4, are: a) Tracked and wheeled ground vehicles. b) Legged platforms. c) Serpentine, vine and continuum platforms. d) Aerial vehicles. e) Marine and underwater vehicles. f) Manipulators and intervention systems. g) Satellite and airborne sensing.

6.1 Tracked and wheeled ground vehicles

Tracked and wheeled ground vehicles, the most frequently used robots at disasters, remain relevant where a hardened chassis is needed, as an explosion-proof tracked design for underground rescue (Murphy et al. 2016; Wang et al. 2023). These vehicles are engineered for confined interiors, fire, radiation or detection payloads. Moreover, they serve as autonomy and interface testbeds.

Regarding the environmental constraint around which the chassis and its payload are engineered, the main categories are:

- Confined-interior reconnaissance platforms: Vehicles sized and articulated for damaged buildings and underground spaces. Examples are two tracked robots of 140 and 150 kg with swing arms for stairs (Wang et al. 2023). Another one is a small four-wheeled robot whose fanless computer copes with dripping water in tunnels about 2 to 5 m wide (Yamanaka et al. 2025). Moreover, onboard computing directly affects the size and power of ground rescue robots (Daun and von Stryk 2023).
- Ground platforms for fire environments: Vehicles carrying sensing or suppression equipment into fire scenes. An example is a low-cost tracked robot reporting temperature, humidity and gas readings from indoor fires (Fernández Talavera et al. 2023). At the other end, the heavy water cannon of a hydrant-fed firefighting robot, with a maximum jet range of about 10 m, resists precise control (Chaoxia et al. 2025).
- Radiation-tolerant platforms: Vehicles whose onboard electronics have a lifespan limited by cumulative dose. An example is a 600 kg Fukushima Daiichi robot, wheeled and thruster-driven in water about 1.5 m deep. Keeping its controller in a low-dose area extends its lifespan 1.8 times (Sakaue et al. 2025).
- Chemically and explosively instrumented platforms: Vehicles built around a detection payload rather than around mobility. An example is a ground vehicle and its robotic arm carrying LiDAR, visible and infrared cameras for landmine detection (Malizia et al. 2026). Another one is a modeled vehicle with a single fast-response chemical sensor at a fixed sampling height, whose time budget better suits a ground than an aerial vehicle (Rhodes et al. 2023).
- General-purpose ground vehicles as autonomy and interface testbeds: Vehicles whose contribution is the software they carry rather than their morphology. Examples are Husky robots chosen for rugged off-road capability and battery capacity (Stedman et al. 2023), or whose 10 cm ground clearance sets the step height that risk mapping must detect (Dixit et al. 2025). Moreover, a four-flipper Telemax carries a manipulator, 3D LiDAR

and 360-degree camera (Bark et al. 2023), while a robot is presented to a language-model copilot as actions with preset costs (Bark et al. 2025). Further examples are a GNSS-reporting vehicle with an onboard pure-pursuit controller (Jarabo-Peñas et al. 2025), as well as TurtleBot (Cao et al. 2024) and Jackal robots (Maresca et al. 2025).

6.2 Legged platforms

Legged platforms entered disaster settings only recently, as quadrupeds reaching the stairs and narrow passageways unsuitable for tracked platforms (Murphy et al. 2016; Kottege et al. 2025). They serve as members of heterogeneous field teams and as standalone walkers on collapse terrain. Moreover, they act as substrates for risk-aware traversability and as carriers of hazard-sensing payloads.

Regarding the role that the legged platform plays within the wider system, the main categories are:

- Team members within heterogeneous field stacks: One morphology among several under a shared perception and autonomy stack. DARPA SubT Challenge finalists fielded IP67-rated 55 kg quadrupeds walking for 80 min (Tranzatto et al. 2024), a commercial quadruped with a 14 kg payload limit (Kottege et al. 2025), or quadrupeds with high-traction rubber foot covers (Morrell et al. 2024).
- Standalone navigation on reconstructed collapse terrain: Legged machines pursuing efficient and reliable locomotion over the terrain of a damaged site on their own rather than within a team (Cruz Ulloa et al. 2025).
- Substrates for risk-aware traversability: Legged platforms for terrain-risk and belief-space formulations, with the gait itself as a planning variable. For instance, a Spot quadruped switches to a stair gait, a three-footed crawl on predicted slip or a 16 cm crouch under low ceilings (Dixit et al. 2025). In another case, the quadruped pitches up to map a staircase before climbing it (Ginting et al. 2023).
- Carriers for responder-operated hazard sensing: Commercial quadrupeds fitted with a detection payload. An example is a sniffer robot whose initially planned tracked base failed the mobility requirements and was replaced by a legged one (Steinbauer-Wagner et al. 2025).

6.3 Serpentine, vine and continuum platforms

Serpentine, vine and continuum platforms reach voids that humans cannot enter and conventional search cameras cannot reach, continuing the snake robots used at earlier collapses (McFarland et al. 2024; Whitman et al. 2018). These platforms either advance by eversion and growth, or act inside the void with a soft helical body.

Regarding the mechanism of advance and the function performed once the platform is inside the void, the main categories are:

- Everting and growing bodies: Mechanisms extending from the tip by eversion, so that the body grows without sliding relative to the ground. An example is a vine robot of about 13 cm lay-flat diameter, which grew to 3 m and was later fed with air from a breathing-apparatus tank (McFarland et al. 2024).

- Soft helical structures for in-void action: Mechanisms whose body is itself the working element. An example is a soft growing robot whose inflatable helical structure is unpressurized during eversion. The structure is then pressurized to keep its inner channel open for hazardous-gas suction (Lee et al. 2025).

6.4 Aerial vehicles

Aerial vehicles, the most used autonomous systems in post-disaster assessment, monitor wide areas at the price of limited flight time and payload (Recchiuto and Sgorbissa 2018). They range from micro aerial vehicles in tunnels and morphing airframes passing boreholes to open-area sUAS. Moreover, they comprise multi-vehicle teams and carriers of non-visual sensors.

Regarding the flight regime and the airframe adaptation that it demands, the main categories are:

- Confined-space and subterranean micro aerial vehicles: Rotorcraft flying in the dust, darkness and narrow tunnels of underground sites. These vehicles are constantly close to walls, where any contact can cause a crash. Moreover, they carry 3D LiDAR, onboard computers and their own lighting (Petrlík et al. 2025; Dahlquist et al. 2025; Nordström et al. 2025).
- Reconfigurable and morphing airframes: Vehicles that fold to pass a restricted access point, such as a 150 mm borehole. They then unfold and undock from their deployment system to fly in the void beyond (Brown et al. 2026).
- Open-area survey sUAS: Multirotor and fixed-wing small aircraft flown over an affected area. In one hurricane response, rotorcraft flew nearly all sorties, alongside a fixed-wing vertical-takeoff aircraft imaging long ranges from limited launch space (Manzini et al. 2023b). Other aircraft yielded orthomosaics at 1.93 to 12.7 cm per pixel (Manzini et al. 2026). Moreover, a battery-limited aircraft was launched from a pad dug into mountain snow (Schomer and Adams 2026), while a quadrotor imaged a collapse in RGB and thermal from about 84 m (Bal et al. 2023). However, such aircraft have limited computing for real-time analysis of wildfire video (Zhao et al. 2024a).
- Multi-vehicle aerial teams and swarms: Groups of aerial platforms sharing a search or monitoring task. In one swarm, vehicles localize one another through onboard ultraviolet markers (Horyna et al. 2023). In other cases, three M300 quadcopters are meshed through onboard Manifold 2 computers (Dah-Achinanon et al. 2023), while three F450 vehicles carry 44 g Compton cameras suited to sub-1 kg vehicles (Werner et al. 2024). Moreover, two 1 m-class quadrotors carry LiDAR, RGB-D camera and companion computer for GNSS-denied indoor flight (Hickling et al. 2026). Finally, fleets pair manipulation and fire-suppression drones with casualty-transport ground vehicles (Zhu et al. 2025).
- Aerial carriers for non-visual sensors: Platforms whose design is driven by the payload rather than the flight envelope. They carry a lightweight scintillator on a collision-tolerant micro flying robot (Mascarich et al. 2023), Geiger-Müller counters (Yadav et al. 2023) or microphone arrays recording human-presence cues under the rotor ego-noise of a frame-mounted drone (Gupta et al. 2025). In contrast, a camera-based counterpart

pairs a 12-megapixel camera with an onboard Jetson Orin Nano to descend into a smoke plume (Pal et al. 2025).

6.5 Marine and underwater vehicles

Marine and underwater vehicles address flooded galleries, harbors and open water, searching areas that can be unsafe or prohibited for divers (Rymansaib et al. 2023; Murphy et al. 2012). They operate as surface vessels in flooded structures or as sonar-equipped surface vehicles over submerged scenes. They also operate as autonomous vehicles searching open water.

Regarding the operating medium and the survey geometry, the main categories are:

- Surface vessels for partially flooded structures: Small boats on the water surface of narrow flooded galleries in caves and mines. An example is a plastic bait-boat catamaran with two differentially driven propellers, which moves only in two dimensions (Villarroel et al. 2024). Meanwhile, its 3D LiDAR and vertical laser rangefinder scan the gallery above the water.
- Surface vehicles for submerged scene investigation: Uncrewed surface vessels imaging a submerged search area with side-scan sonar. An example is a catamaran of about 1.2 m and 20 kg driven by two rear thrusters (Rymansaib et al. 2023). Its latest version raises its submerged sensors and thrusters on hinges for transport.
- Autonomous vehicles for open-water search: Platforms whose problem is the drifting search area rather than the site itself. An example is an underactuated rudder-and-propeller torpedo vehicle of 2.34 m and 87.6 kg (Zhou et al. 2026). Other conditions being fixed, its side-scan sonar sweep width is set only by its height above the bottom.

6.6 Manipulators and intervention systems

Manipulators and intervention systems let a robot act physically on a disaster site rather than only observe it, sampling the environment or interacting with survivors (Murphy et al. 2016). They comprise heavy task-priority arms, effectors coupled to the operator and soft task-specific bodies. Moreover, they include casualty-contact systems and manipulators within supervised teams.

Regarding what the end effector must contact and under whose authority the contact is made, the main categories are:

- Heavy-duty task-priority manipulators: Dual-arm upper bodies on hybrid tracked-legged bases. An example is a 680 kg robot whose replaceable lower arms, grippers and transfer bed switch it between manipulation and casualty rescue under task-priority whole-body control (Hong et al. 2023). In contrast, a teleoperated vehicle compact enough to be carried up a stairwell opened a valve with a custom-fabricated extension of its last arm segment (Lichtenfeld et al. 2025).
- Specialized end effectors with operator feedback: Effectors coupled to the operator through haptic, tactile or immersive channels. An example is a telepresence dual-arm manipulator for EOD, mounted on an industrial arm, whose gripper carries a dovetail adapter for changing tools (Xiao et al. 2023). Another one is a Telemax arm with a

parallel gripper and gripper camera on a ruggedized tEODor tracked base (Illing et al. 2025).

- Soft and morphology-specific effectors: Compliant bodies engineered for one task in one geometry rather than conventional arms, such as the helical suction structure in Lee et al. (2025).
- Casualty-contact manipulation: Systems whose contact target is the human body. An example is a Franka Emika Panda arm on a mobile platform, whose end-effector grasp pose must coincide with the grasp location on the limb (Peiros et al. 2023).
- Manipulation within supervised team architectures: Manipulation as one resource among many under a team-level controller. An example is a Husky ground vehicle carrying a UR5 arm for remote handling. It is grouped in a virtual digital-twin implementation with two other ground vehicles and a Tello aerial vehicle (Baniqued et al. 2024).

6.7 Satellite and airborne sensing

Satellite and airborne sensing covers broad regions at lower cost than ground surveys and provides a first general picture of the situation (Ahn et al. 2025; Recchiuto and Sgorbissa 2018). However, it is a platform category only in a loose sense. The reason is that its works contribute analysis pipelines constrained by revisit time, sensor modality, imaging altitude and annotation needs.

Regarding the sensing geometry and the revisit regime of the imagery, the main categories are:

- Spaceborne multi-temporal constellations: Repeat satellite acquisition that lets the analysis be posed as a change problem. Examples are the Sentinel-1 SAR series of 43 flood events, whose post-event acquisition delay depends on the revisit time. The radar is unaffected by the cloud cover and precipitation that blind multi-spectral sensors during floods (Bountos et al. 2024). Another example is pre- and post-event pairs of the 2023 Türkiye earthquake, freely provided by Maxar (Ahn et al. 2025).
- Multi-sensor and multi-modal Earth observation: Optical, radar, temporal and increasingly textual streams within one interface. These streams are combined on purpose-built multi-hazard resources (Wang et al. 2025), as well as in general Earth-observation assistants and benchmarks (Soni et al. 2025; Irvin et al. 2025; Danish et al. 2025). In addition, 0.3 m aerial imagery acquired during hurricane flooding is fused with terrain elevation (Sami et al. 2024).
- Low-altitude aerial imagery as an overhead product: Orthomosaics flown by sUAS, treated as an overhead data product rather than as robot perception. Their analysis must handle a non-uniform misalignment of prior road lines that satellite imagery does not exhibit (Manzini et al. 2026).
- Change models that relax paired-annotation requirements: Approaches reducing the dependence on co-registered, annotated pre- and post-event pairs. They rely on training-free adaptation of the Segment Anything Model (Kirillov et al. 2023) to bi-temporal change detection (Zheng et al. 2024b), and on generative synthesis of change data from single-temporal imagery (Zheng et al. 2025). Moreover, detectors are trained from unpaired single-date labels (Zheng et al. 2024a).

Table 5 Robotic platforms and morphologies: Comparative analysis and key insights.

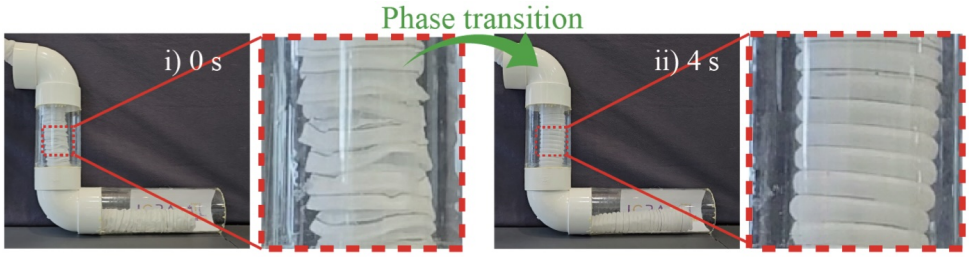
Platform	Tracked & wheeled	Legged	Serpentine, vine & continuum	Aerial	Marine & underwater	Manipulators & intervention	Satellite & airborne sensing
Primary role	• Carry heavy, hardened payloads into a damaged structure • Serve as autonomy and interface testbeds	• Negotiate discontinuous, unstable footing • Extend team reach beyond wheeled limits	• Enter voids too small for people • Search inside or remove hazardous gas	• Rapid obstacle-independent coverage • Reach shafts, tunnels and interiors	• Operate in flooded and open water • Shorten divers' time underwater	• Change the state of environment or casualty • Deliver the response action itself	• Cover broad geographic regions • Feed offboard analysis pipelines
Mobility envelope and access limit	• Passage width and step height • Defeated by rubble discontinuity	• Steps, gaps and loose footing • Bounded by slip and fall risk	• Interstitial voids and conduits • Reach set by stored body length	• Free space above and within structures • Bounded by endurance and airflow	• Water column and surface • Bounded by currents and debris	• Workspace of the arm on its base • Bounded by reach and base stability	• Global in area; misses street-scale damage • Bounded by revisit interval and cloud
Typical sensing payload	• LiDAR, cameras, thermal • Gas and multi-spectral instruments	• LiDAR-inertial, cameras • Gas and radiation detectors	• Tip camera and light • Minimal by mass budget	• Cameras, LiDAR, thermal • Radiation detectors, gas sensors, microphones	• Sonar and underwater cameras • Surface LiDAR on vessels	• Force, tactile and chemical sensing • Haptic channels to the operator	• Optical, SAR and multi-spectral • Text paired with imagery
Dominant autonomy mode	• Teleoperation and variable autonomy	• Supervised autonomy within a team	• Direct joystick teleoperation	• Autonomous in trials, piloted at incidents	• Autonomous in trials, teleoperated at incidents	• Mainly teleoperated	• Offline, offboard inference
Domains where it dominates	• Subterranean and mine • Earthquake, nuclear and CBRN	• Subterranean and mine • Nuclear and CBRN	• Earthquake and structural collapse	• Flood, hurricane and wildfire • Subterranean and mine	• Maritime and underwater • Flooded subterranean galleries	• Explosive hazards • Industrial and hazmat	• Multi-hazard benchmarks • Flood and hurricane
Key strengths	• Payload, endurance and hardening • Longest real-incident record	• Stairs and passages unsuited to tracks • Proven at field scale in teams	• Grows through apertures and around turns • Advances without purchase on rubble	• Rapid needs assessment in the first 72 hours • Hundreds of sorties at a single disaster	• Sonar search in turbid water • Maps galleries beyond conventional survey	• Built to act on the world • Valve opening averted an explosion threat	• Coverage no vehicle can match • Richest data and model support

Table 5 (continued)

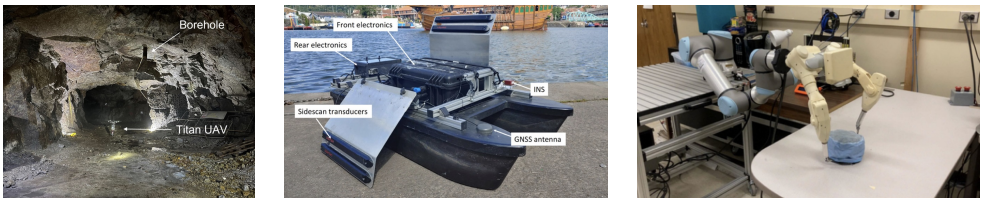
Platform	Tracked & wheeled	Legged	Serpentine, vine & continuum	Aerial	Marine & underwater	Manipulators & intervention	Satellite & airborne sensing
Critical limitations	• Radiation shortens electronics lifespan • Added devices slow the robot	• No disaster-specific survey • Payload and endurance penalty	• Thin corpus; isolated real-incident evidence • Portability limits field use	• Endurance, weather and regulation • Payload-limited; propellers disturb gas sensing	• Among the smallest corpora of the mobile classes • No competition-driven development	• Weak links and limited operator assistance • No recorded contact datasets	• Presupposes off-board compute and links • Not an embodiment; no site access
Typical validation maturity	• Simulation, field trials and real incidents	• Competitions and controlled field trials	• Laboratory and controlled field trials	• Field trials, simulation and real incidents	• Controlled field trials; early real incidents	• Laboratory; isolated real-incident records	• Benchmark evaluation, offline
Indicative platforms and works	• Fukushima remote system (Sakaue et al. 2025), indoor firefighting UGV (Fernández Talavera et al. 2023), team CSIRO Data61 (Kottege et al. 2025)	• Sniffer quadruped (Steinbauer-Wagner et al. 2025), CERBERUS (Tranzatto et al. 2024), arena-tested quadruped (Cruz Ulloa et al. 2025)	• Portable vine robots (McFarland et al. 2024), helical gas-suction robot (Lee et al. 2025)	• SubT aerial autonomy (Petrlík et al. 2025), Prometheus (Brown et al. 2026), CRASAR sUAS (Manzini et al. 2023b)	• RoboBoat (Villarroel et al. 2024), underwater investigation surface vessel (Rymansaib et al. 2023), maritime search underwater vehicle (Zhou et al. 2026)	• HURCULES (Hong et al. 2023), EOD telepresence (Xiao et al. 2023), casualty manipulation (Peiros et al. 2023)	• Kuro Siwo (Bountos et al. 2024), DAVI (Ahn et al. 2025), DisasterM3 (Wang et al. 2025)



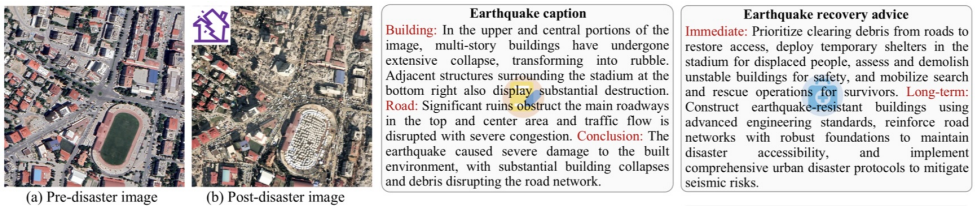
(a) (b)



(c)



(d) (e) (f)



(a) Pre-disaster image (b) Post-disaster image

(g)

Fig. 7 Representative literature platforms per morphology: a) Tracked and wheeled ground vehicles (Sakaue et al. 2025). b) Legged platforms (Cruz Ulloa et al. 2025). c) Serpentine, vine and continuum platforms (Lee et al. 2025). d) Aerial vehicles (Brown et al. 2026). e) Marine and underwater vehicles (Rymansaib et al. 2023). f) Manipulators and intervention systems (Xiao et al. 2023). g) Satellite and airborne sensing (Wang et al. 2025)

6.8 Comparative analysis and key insights

Table 5 compares the seven platform classes. The main insights are that: a) Apart from tracked and wheeled vehicles as general-purpose carriers, the classes differ mainly in the physical barrier each one addresses (terrain discontinuity, confinement, ground obstacles, water, contact and distance). Hence, they mostly complement one another. This is consistent with a substantial minority of the primary works being coded with more than one platform. b) Access and payload tend to trade off, as for a vine robot whose steering is restricted by its 145 g camera mount (McFarland et al. 2024). c) The corpus is markedly asymmetric. Tracked, wheeled and aerial platforms are primary in most works, whereas the marine and continuum classes rest on few ones. d) Foundation models are most common in overhead sensing, the class that presupposes offboard computation and connectivity. In contrast, damage segmentation from ground-level social-media photographs (Huang et al. 2026) is tied to no platform at all. Representative platforms per morphology are illustrated in Fig. 7.

7 Perception and state estimation

Perception and state estimation concerns how a method senses a damaged environment, localizes the platform within it and represents what has been observed. This conditions any downstream autonomy, while it is directly exposed to the perceptual degradation of disaster sites (Ebadi et al. 2024; Chen et al. 2026). The main categories, as defined in Section 4, are: a) Degradation-robust localization and mapping. b) Multi-robot and multi-session mapping. c) Closed-vocabulary disaster perception. d) Promptable and open-vocabulary perception. e) Actionable semantic scene graphs. f) Vision-language disaster interpretation.

7.1 Degradation-robust localization and mapping

Rescue robots rely on localization and mapping to move autonomously and safely through a site. However, darkness, dust, smoke and self-similar surfaces break the assumptions of conventional SLAM (Ebadi et al. 2024; Chen et al. 2026). Methods counter this failure by detecting ill-conditioned geometry, substituting failing sensors, hardening whole stacks and exposing their uncertainty.

According to the failure mode that the method is designed to survive, the main categories are:

- Localizability-aware registration: Methods detecting ill-conditioned geometry before trusting a registration result. They apply controlled updates along partially localizable directions, while keeping the odometry prior along nonlocalizable ones, where registration would otherwise diverge (Tuna et al. 2024).
- Modality substitution under sensor denial: Methods replacing the failing modality. An example is millimeter-wave radar in tunnels where poor illumination, dust and water vapor strain LiDAR- and camera-based SLAM (Zeng et al. 2024). Another one is stereo long-wave infrared depth estimation for aerial platforms, trained on smokeless frames and generalizing qualitatively to smoke (Dhrafani et al. 2025).
- Field-hardened complete stacks: Robustness as a property of the whole estimation pipeline. It is achieved through loosely coupled fusion with health checks on each

data source (Ebadi et al. 2024), as well as laser odometry falling back from a visual-thermal-inertial prior to kinematic or inertial ones (Tranzatto et al. 2024). Moreover, intensity-based filtering removes propeller-raised dust from aerial LiDAR scans (Petrlík et al. 2025).

- Estimation uncertainty exposed to downstream decisions: Confidence carried forward instead of a point solution. Examples are a categorical terrain belief absorbing semantic misclassification (Ginting et al. 2023), or odometry uncertainty propagated to each point of an aggregated point cloud (Dixit et al. 2025).

7.2 Multi-robot and multi-session mapping

Map sharing is especially demanding at disaster sites, because self-similar locations raise the risk of false loop closures. Moreover, communication is limited by bandwidth, latency and packet loss underground (Ebadi et al. 2024; Chen et al. 2026). The corresponding methods fuse maps at a base station, across morphologies, over constrained links or across sorties.

According to the constraint that governs how map information is shared, the main categories are:

- Centralized fusion at a base station: Submaps or keyed scans merged at a base station into one multi-robot map. The merged map provides a globally consistent estimate (Tranzatto et al. 2024), or delivers search-and-rescue data to the operator with an outlier-robust back end against false loop closures (Morrell et al. 2024).
- Cross-morphology map merging: Methods reconciling maps built from different viewpoints and sensing geometries. In particular, aerial and ground semantic maps of the same site are registered through GNSS into the map over which a ground robot plans (Cladera et al. 2025).
- Bandwidth- and link-constrained sharing: What can actually be transmitted from a disaster site. At one end, map frames are synchronized peer-to-peer whenever a link exists, with each agent optimizing its own copy of the team map (Kottege et al. 2025). At the other end, high-resolution aerial maps require upload for cloud processing, whereas low-resolution ones are generated on a laptop in under an hour (Manzini et al. 2023c).
- Multi-session and cross-sortie consistency: Map information carried across sorties into the same site. For instance, compact topological-volumetric maps are handed from each returning aerial robot to the next, in order to target unexplored parts (Petrlík et al. 2025). Moreover, a point-cloud map from an earlier mapping phase is used for relocalization by every inspection agent (Dahlquist et al. 2025). In addition, multi-day collapse imagery is reconstructed in one coordinate frame, with far lower translation error than separately aligned reconstructions (Bal et al. 2023).

7.3 Closed-vocabulary disaster perception

Closed-vocabulary perception needs labeled examples of every category. However, damage varies by disaster type, while a decision boundary learned for one type must shift for another (Ahn et al. 2025). The fixed label sets are applied onboard in real time, offline at pixel level, to non-visual signals or to degraded imagery.

According to the sensing modality and the deployment budget under which the fixed label set is applied, the main categories are:

- Onboard real-time visual recognition: Models compressed toward the budget of an aerial platform, so that results arrive while the robot can still act on them. An example is a low-latency YOLO variant implemented on a field-programmable gate array (FPGA) for onboard use, which detects humans and cars in RGB or event-camera data. Its indoor flight tests ran on an embedded GPU (Amesseghер et al. 2025). Another one is a wildfire video classifier whose policy networks drop redundant portions of the video (Zhao et al. 2024a).
- Offline semantic segmentation of disaster imagery: Pixel-level delineation without a real-time constraint. It relies on attention-augmented encoder-decoders for ground-vehicle search-and-rescue imagery (Salas-Espinales et al. 2023) and mask transformers on ground-level earthquake photographs (Huang et al. 2026). Moreover, it relies on road-condition models for sUAS orthomosaics (Manzini et al. 2026) and elevation-gated flood segmentation (Sami et al. 2024). Labels are sometimes synthesized as bi-temporal building change pairs derived from pre-disaster images and building footprints (Zheng et al. 2025).
- Non-visual and multi-sensor recognition: Signals other than ordinary color imagery. These include human sounds buried in propeller ego-noise (Gupta et al. 2025). They also include LiDAR scans and visible to long-wave infrared images of landmines, taken from a ground robot and its arm in afternoon light and, on one track, at twilight (Malizia et al. 2026).
- Recognition under sensor degradation: A fixed label set applied to imagery corrupted by the hazard itself. An example is smoke-degraded firefighting video annotated with eight action and six object classes. This video is also tagged with smoke grade, heat distortion and lighting, in order to score recognition against degradation severity (Tiwari et al. 2025).

7.4 Promptable and open-vocabulary perception

Promptable and open-vocabulary perception is appealing, because deep change detection models do not generalize to change types and geographic areas beyond those seen during training (Zheng et al. 2024b). The corresponding models are applied frozen, used as pseudo-label sources or queried during the mission.

According to how the model is specialized to the disaster domain in the absence of domain annotation, the main categories are:

- Frozen promptable backbones applied directly: A general-purpose promptable segmentation model (Kirillov et al. 2023) used with no parameter update. An example is a change detector matching the object-mask embeddings of two acquisition dates in that model's latent space. In this way, unseen change types are segmented zero-shot (Zheng et al. 2024b).
- Promptable models as pseudo-label sources: Damage read from a text-prompted segmentation model. The confidence drop of the model for the prompt "Building" between pre- and post-event images is thresholded. It is then merged with a source-trained

model's change map into pseudo labels. Hence, the frozen component supplies supervision rather than the final prediction (Ahn et al. 2025).
- Open-vocabulary interpretation in the control loop: A multi-modal model queried during the mission over camera frames and LiDAR scans rendered as bird's-eye-view traversability images. The same prompted model that selects the waypoints also recognizes abnormal objects, such as a safety helmet (Yamanaka et al. 2025).

7.5 Actionable semantic scene graphs

Actionable semantic scene graphs sit between a geometric map, which lacks object semantics, and a caption, which supports no action. They interface perception with language-conditioned planning, while sparing operators raw sensor data (Ravichandran et al. 2025; Adams et al. 2024). The graphs are annotated with affordances, queried by a language front end or carry their own uncertainty.

According to what the graph is constructed to be consumed by, the main categories are:

- Affordance-annotated maps for robot action: Each object and each room carries the operations a platform can perform on them, as well as their costs. Objects are labeled by projecting 360-degree instance masks onto LiDAR points. Hence, planning queries the map for available actions, instead of inferring them from geometry (Bark et al. 2023).
- Graphs as the query surface for a language front end: Representations read by a large model or constraining its outputs. An example is a graph of rooms, entities, connections and affordances that grows with new detections, and against which extracted tasks are matched (Bark et al. 2025). In another case, region nodes from free space and object nodes from open-vocabulary detections are joined by traversability and visibility edges, as well as enriched with semantic descriptions (Ravichandran et al. 2025).
- Graphs carrying their own uncertainty: Representations in which each roadmap node carries a Gaussian pose belief and a categorical belief over terrain classes, such as stairs, rubble or unknown. These beliefs are sharpened by information-gathering self-loop edges (Ginting et al. 2023).

7.6 Vision-language disaster interpretation

Current vision-language models transfer poorly to disaster tasks, even when trained on aerial and satellite imagery (Wang et al. 2025). Moreover, they fall far below human experts on degraded firefighting video (Tiwari et al. 2025). These shortcomings motivate disaster-specific corpora, temporal and multi-sensor assistants, wider geospatial benchmarks and interpretation embedded in robot actions.

According to what is interpreted and how the interpretation is evaluated, the main categories are:

- Disaster-specific instruction corpora and their diagnostics: Resources specializing and testing general models on co-registered pre-event optical and post-event optical or radar imagery. Their tasks range from disaster recognition and damaged-object counting to

Table 6 Perception and state estimation: Comparative analysis and key insights.

Perception category	Degradation-robust SLAM	Multi-robot & multi-session mapping	Closed-vocabulary perception	Promptable & open-vocabulary perception	Actionable semantic scene graphs	Vision-language interpretation
Primary function	• Estimate pose and geometry where sensing degrades • Anchor maps and detections in one frame	• Merge partial maps across robots and sorties • Maintain one coherent picture for the operator	• Recognize a fixed set of disaster-relevant classes • Detect people, damage, fire or landmines	• Recognize categories no training set anticipated • Locate change and damage across dates	• Represent objects, relations and affordances • Bridge metric maps and language planning	• Describe, question and reason about a scene • Produce human-readable interpretation
Sensing input	• LiDAR, radar, visual-thermal-inertial • Proprioception	• Per-robot or per-sortie maps and imagery • Link-quality measurements	• RGB, event, thermal, audio, 3D	• Image pairs or robot views, plus prompts	• Metric map with object detections	• Optical, radar or video imagery with text
Output representation	• Pose estimate and metric map • Localizability or belief annotations	• Merged team map, central or per agent	• Class labels, boxes or per-pixel masks	• Masks, change maps or waypoint choices	• Queryable, executable relational graph	• Descriptions, answers, reports, referring masks
Supervision requirement	• Mostly model-based; some learned components	• Geometric matching; learned loop-closure ranking	• Labor-intensive annotation, or synthetic change pairs	• None, or target-region pseudo labels	• Pre-trained or open-vocabulary detectors; affordance definitions	• Remote-sensing instruction tuning, or zero-shot use
Where computation runs	• Onboard, in real time	• Onboard, base station or cloud	• Onboard if compressed, else offline	• Offline on GPUs; cloud model in the field	• Onboard graph, offboard language model	• Mostly offboard; some onboard captioning
Key strengths	• Mature given high-quality multi-modal sensing • Prerequisite for successful operation	• Turns a team into one sensing system • Accurate for moderate team sizes	• Higher precision than zero-shot detectors • Deployable within onboard budgets	• No annotation campaign required • Categories defined at the incident	• Grounded and queryable at once • Directly consumable by planners	• Output legible to responders • One model spans many Earth-observation tasks

Table 6 (continued)

Perception category	Degradation-robust SLAM	Multi-robot & multi-session mapping	Closed-vocabulary perception	Promptable & open-vocabulary perception	Actionable semantic scene graphs	Vision-language interpretation
Critical limitations	• Degenerate geometry can still break registration • Largely geometric; manual parameter tuning	• Bandwidth limits what maps can be shared • False loop closures where places look alike	• Bound to the annotated hazards and geographies • Degrades under ego-noise or misaligned priors	• Lower precision than supervised oracles • Prompt sensitivity; heavy compute	• Evaluations often assume ideal detections • Affordance schemas not standardized	• Struggle with cross-sensor input and counting • Embodied use often relies on cloud uplinks
Typical validation maturity	• Competitions and controlled field trials	• Competitions, field trials, incident data	• Benchmarks; lab and operational checks	• Benchmarks; isolated field test	• Simulation; some field trials	• Benchmarks; few field trials
Indicative works	• X-ICP (Tuna et al. 2024), Millimeter-wave radar SLAM (Zeng et al. 2024), SubT SLAM review (Ebadi et al. 2024)	• NeBula addendum (Morrell et al. 2024), team CSIRO Data61 (Kottege et al. 2025), air-ground collaboration (Cladera et al. 2025)	• SAR Nets (Salas-Espinales et al. 2023), low-latency onboard detection (Amesegher et al. 2025), DroneAudioSet (Gupta et al. 2025)	• DAVI (Ahn et al. 2025), Segment Any Change (Zheng et al. 2024b), vision-LLM navigation (Yamanaka et al. 2025)	• Affordance mapping (Bark et al. 2023), SPINE (Ravichandran et al. 2025), semantic belief graphs (Ginting et al. 2023)	• DisasterM3 (Wang et al. 2025), TEOChat (Irvin et al. 2025), GEOBench-VLM (Danish et al. 2025)

relational reasoning and long-form reports (Wang et al. 2025). In addition, vision-language and open-vocabulary models are probed zero-shot on degraded firefighting video (Tiwari et al. 2025).

- Temporal and multi-sensor Earth-observation assistants: Models extending interpretation across acquisition dates and sensor types. Some localize buildings and grade their damage within conversations about image sequences (Irvin et al. 2025). Others fuse multi-spectral, radar and time-series channels into visual tokens, which extends damage assessment to earthquake detection from bi-temporal radar pairs (Soni et al. 2025).
- Geospatial benchmarks placing disaster tasks in a wider suite: Disaster-type classification and damaged-building counting over pre- and post-event pairs, posed as items of a wider suite. This suite also spans object detection, referring segmentation and non-optical imagery (Danish et al. 2025).
- Interpretation embedded in an embodied loop: Models informing the actions of a platform rather than the analysis of an archive. For instance, they select waypoints executed without position feedback in tunnels where SLAM accuracy degrades (Yamanaka et al. 2025). Moreover, they caption onboard (LLaVA) the objects that ground robots detect and localize from RGB-D data, for an air-ground semantic map (Cladera et al. 2025). Furthermore, they list the objects in view with confidence scores and spatial relations, such as a box blocking a door, while segmented object point clouds parameterize behavior-tree actions (Shi et al. 2025).

7.7 Comparative analysis and key insights

Table 6 compares the six perception categories. The main insights are that: a) The categories form a ladder of abstraction, from pose and geometry through fixed and open label sets to relations and language. Robot-borne systems mostly climb this ladder in order, as a scene graph built over LiDAR odometry (Ravichandran et al. 2025). However, a vision-language navigator bypasses an unreliable SLAM estimate (Yamanaka et al. 2025), while overhead change and interpretation works need no robot pose (Zheng et al. 2024b; Wang et al. 2025). b) Computation runs onboard for registration and compressed detection, as well as at a base station or peer-to-peer for team maps. In contrast, it runs remotely for language reasoning or cloud-hosted waypoint selection. c) Supervision does not grow steadily up the ladder. Pose and map estimation largely need no annotation, whereas closed-vocabulary models are bound to the regions and disaster types of their training data (Ahn et al. 2025). Promptable ones need no training (Zheng et al. 2024b) or rely on pseudo labels (Ahn et al. 2025). Moreover, vision-language models improve after tuning on a disaster corpus (Wang et al. 2025). d) The semantic layers remain thinly evidenced. Scene graphs come from few works and groups, while vision-language models transfer poorly to disaster tasks.

8 Autonomy, planning and control

Autonomy, planning and control concerns how decisions are made and the abstraction level at which the human enters, from continuous command to open-ended instruction. This level determines how many platforms one responder can direct (Murphy et al. 2016; Adams et al. 2024). The main paradigms, as defined in Section 4, are: a) Teleoperation and variable autonomy. b) Supervised autonomy. c) Exploration and coverage planning.

d) Risk-aware traversability and belief-space planning. e) Multi-agent task allocation. f) Language-conditioned mission planning. g) Learned control and decision policies.

8.1 Teleoperation and variable autonomy

Teleoperation remains the typical mode of rescue robots in actual disasters, although the documented problems of manual teleoperation argue for increasing autonomy (Daun and von Stryk 2023; Murphy et al. 2016). The corresponding methods keep direct control, switch authority during the mission, extend the operator's perception or derive the division of authority from requirements and errors.

Regarding how control authority is split between operator and system, and when the split changes, the main categories are:

- Continuous direct control as fielded practice: Remote control through an offboard controller, which leaves the robot unable to act once the link is cut (Sakaue et al. 2025). In another case, a second pilot steers an autonomously filming drone back out of a hall (Surmann et al. 2024). Moreover, manual teleoperation is used to reach a reactor of known position, while the proprietary radio runs no assistance function (Lichtenfeld et al. 2025).
- Run-time authority switching: Control reassigned during the mission. For instance, a supervisor teleoperates a robot into a tunnel before leaving its exploration to the planner (Biggie et al. 2023). Similarly, a requirements analysis for rescue robots calls for a lower autonomy level or operator takeover once the conditions of an assistance function no longer hold (Daun and von Stryk 2023).
- Perceptual extension of the operator: Authority left with the human while perception improves. This is achieved through immersive displays, with commands still issued from handheld controllers (Stedman et al. 2023; Illing et al. 2025). Moreover, haptic and chemical feedback is provided, with a learned classifier only assisting object identification (Xiao et al. 2023). Finally, in a digital twin, evaluators struggling in narrow passages suggested waypoint-based shared autonomy (Baniqued et al. 2024).
- Requirement- and error-derived constraints: Empirical guides to the division of authority. Disaster sites resist high autonomy. Hence, assistance functions must complement operator skills (Daun and von Stryk 2023). Moreover, human errors are evenly split between manual and autonomous flights (Murphy et al. 2024), while teleoperating pilots err more at mission termination as fatigue grows (Mehta et al. 2024).

8.2 Supervised autonomy

Autonomy can raise the neglect tolerance of each robot and let one operator supervise many, whereas a larger team otherwise requires more trained operators (Adams et al. 2024; Murphy et al. 2016). In this paradigm, the supervisor directs a heterogeneous team at task level, relies on autonomy that persists through communication loss, or specifies team-level missions.

Regarding the level of abstraction at which the operator issues intent, the main categories are:

- Task-level direction of a heterogeneous team: Independent single-robot autonomy under automated mission management, freeing a single supervisor from agent-level micromanagement in subterranean exploration (Biggie et al. 2023).
- Autonomy that persists through communication loss: A robot beyond the link that keeps executing. For instance, it performs return-to-synchronize tasks that a teammate holding its data may bid on (Kottege et al. 2025). Moreover, it homes after a supervisor-set out-of-range budget, or backtracks until reconnected (Tranzatto et al. 2024). In another case, it detours to a high-signal checkpoint or deploys relay radios (Morrell et al. 2024).
- Mission-level specification: An objective stated for the team as a whole, which the operator then retasks on the findings it reports. An example is an air-ground collaboration whose ground robot spent nearly 90% of a mission in autonomous mode, while takeovers arose mainly from dynamic or undetected obstacles (Cladera et al. 2025).

8.3 Exploration and coverage planning

Exploration was rewarded only indirectly in subterranean competitions, through accurately reported artifacts, while unreliable early information makes preplanned drone paths infeasible in mountain rescue (Petrlík et al. 2025; Schomer and Adams 2026). The corresponding planners pursue volumetric coverage, operational objectives, target beliefs or sparse physical signals. In addition, they allocate frontiers across teams.

Regarding what the planner is constructed to maximize and under which budget, the main categories are:

- Volumetric coverage of unknown space: Frontiers or viewpoints selected by the unexplored space they would reveal against path cost. An example is aerial subterranean exploration weighing ray-scored frontier viewpoints against surface-coverage viewpoints (Petrlík et al. 2025). Another one is a borehole-deployed drone discounting frontier entropy by path length (Brown et al. 2026).
- Coverage against an operational objective: Objectives defined by the response rather than by geometry. An example is a reinforcement-learned nuclear-emergency coverage policy, whose reward weighs radiation limits, turns, flight time and return to base (Blake and Schubert 2025). In contrast, without coverage planning, a route passes through operator-selected gas-measurement waypoints in a blast area before re-entry (Nordström et al. 2025).
- Probabilistic search over a target belief: Sensing effort allocated to where a missing person or wreckage is likely to be. Prioritized viewpoints are ordered as a clustered traveling-salesperson problem, trading early coverage of likely regions against energy (Schomer and Adams 2026). Alternatively, a coverage path is fitted to the region predicted from a wreck's drift (Zhou et al. 2026).
- Informative-path planning for sparse physical signals: Trajectories accumulating evidence from individually uninformative measurements. Uncertainty-aware planning steers toward locations allowing fast, high-fidelity radiation-field estimation (Mascarich et al. 2023). Moreover, receding-horizon selection weighs obstacle proximity against closeness to a moving target (Yadav et al. 2023). In addition, tree search toward the estimated release source keeps exploring better sampling locations (Rhodes et al. 2023).

- Frontier allocation across a team: Deciding not only where to go but which robot goes there, and re-deciding as the map grows. In this case, a frontier cluster is valued mainly by its location relative to all agents, rather than by the volume behind it (O'Brien et al. 2023).

8.4 Risk-aware traversability and belief-space planning

Risk-aware planning reasons explicitly over what the robot does not know. Navigating on an incorrect terrain classification can be catastrophic, while always assuming the most likely terrain type loses time (Ginting et al. 2023). The uncertainty arises from terrain and traction, semantic labels or localization.

Regarding the source of the uncertainty over which the planner reasons, the main categories are:

- Terrain and traction risk: The risk that a given path is not physically negotiable. It is assessed on the tail of a traversability distribution rather than on a single score (Dixit et al. 2025). Alternatively, candidate routes are sampled in a digital twin of the laser-scanned site and ranked before execution (Cruz Ulloa et al. 2025).
- Semantic uncertainty attached to the map: The confidence of a semantic label carried on the representation itself. In this way, a policy minimizing traversal time to the target decides when an information-gathering action should precede commitment to a terrain-specific locomotion controller (Ginting et al. 2023).
- Localization uncertainty admitted into the plan: Localization error entering the plan indirectly, by inflating the variance of the traversability estimates over which a risk-aware planner optimizes (Dixit et al. 2025). In contrast, a planner responding to such uncertainty is only identified as a possible generalization of a subterranean planning framework (Morrell et al. 2024).

8.5 Multi-agent task allocation

Disaster task allocation faces tasks unknown in advance, as well as links that may be lost often and for long (O'Brien et al. 2023). The allocation is computed centrally or locally under degraded links, bound by energy budgets, extended to tasks discovered mid-mission or mediated by language.

Regarding where the allocation decision is computed and which budget binds it, the main categories are:

- Centralized allocation with a global view: A central coordinator computing the assignment. In one case, an auctioneer assigns operator-selected inspection tasks on the agents' bids, while each agent finishes its task after a link loss before returning to range (Dahlquist et al. 2025). In another case, one vehicle plans information-gaining waypoints for the team and assigns them by spatial clustering (Werner et al. 2024).
- Decentralized allocation under degraded links: Local decisions by each robot. For instance, swarms turn toward a candidate object by propagating detections between neighbors (Horyna et al. 2023). Moreover, sparse teams search through a distributed belief map before relaying toward a detected target (Dah-Achinanon et al. 2023).

In addition, auctions without a designated auctioneer resolve the conflicts left by communication failures through prioritized rules (O'Brien et al. 2023).

- Budget-bound allocation: Energy rather than task structure as the binding constraint. The allocation is made either online, with a contested target going to the robot with the better position and battery level (Maresca et al. 2025), or offline, by sizing the smallest sufficient fleet and its initial paths before deployment (Romero et al. 2024).
- Allocation over tasks discovered mid-mission: A task set growing during the search. For instance, frontiers are bid on while the map expands, alongside operator-created tasks that take priority over autonomous bids (O'Brien et al. 2023). Alternatively, moving targets are handled implicitly by a learned multi-agent search policy, with a rule-based A* fallback for a stalled agent (Cao et al. 2024).
- Language-mediated allocation: A large model proposing strategies for decomposing each task, while branch-and-bound search with integer programming assigns and schedules the resulting subtasks to minimize the overall makespan (Zhu et al. 2025).

8.6 Language-conditioned mission planning

Language-conditioned planning rests on the contextual reasoning of large language models (Ravichandran et al. 2025). However, it lacks formal verification and plan-quality guarantees, unless assignment is delegated to model-based optimization (Zhu et al. 2025). The resulting plans are grounded in online maps, verified before execution, decomposed for teams with explanation or issued as spoken commands.

Regarding what the language model is permitted to produce and what verifies its output before execution, the main categories are:

- Semantic plans grounded in an online map: Plans made against a scene representation built during the mission. In one case, validated subtasks are realized over a receding horizon as the semantic map grows (Ravichandran et al. 2025). In another case, grounded subtasks are handed to a graph search, which selects autonomous, assisted or teleoperated execution by cost unless the operator constrains it (Bark et al. 2025).
- Executable structures verified before running: Symbolic actions emitted by a decision model and mapped by an execution one onto predefined function schemas. A behavior tree runs these actions once their preconditions hold. In contrast, unmet postconditions trigger a re-evaluation of the scene and replanning (Shi et al. 2025).
- Team-level decomposition with explanation: A team mission translated into verifiable temporal-logic formulas and decomposed with explicit reasoning. In this case, an operator checks ordering, strategies, assigned plans and execution status (Zhu et al. 2025). Alternatively, a model sets the classes an aerial mapper searches for and justifies each ground-robot behavior sequence, replanning once untraversable edges are removed (Cladera et al. 2025).
- Spoken command at the edge of the mission stack: Language at the command layer. Each goal derived from a voice command is previewed and confirmed by the operator. Subsequently, an edge-tier planner generates a safe global path, which the onboard controller tracks within operational windows authorized by the command post (Jarabo-Peñas et al. 2025).

Table 7 Autonomy, planning and control: Comparative analysis and key insights.

Autonomy paradigm	Teleoperation & variable autonomy	Supervised autonomy	Exploration & coverage planning	Risk-aware & belief-space planning	Multi-agent task allocation	Language-conditioned planning	Learned control policies
Primary function	• Place a human in direct control of the platform • Shift authority as difficulty and links change	• Let one operator direct a team at task level • Continue acting beyond the link	• Decide where to go next in unknown space • Allocate sensing effort over an objective	• Plan over what the robot does not know • Trade information gain against risk	• Distribute work across a team • Re-decide as tasks and budgets change	• Turn an open instruction into grounded actions • Accept under-specified missions	• Close the loop where dynamics resist design • Learn reward-shaped guidance and allocation
Locus of decision	• Mainly the operator; some onboard functions	• Operator sets tasks; robots decide how	• Per robot or team, onboard or pre-planned	• Onboard planner over a belief	• Team-level, centralized or peer-to-peer	• Large model, off-board, over a scene representation	• Trained policies, onboard or at a ground station
Human involvement	• Near-constant; often a crew per robot	• Intermittent; one operator per team	• Sets the area and the objective	• Minimal; risk level tuned by heuristics	• Sets the team goal and budgets	• States the mission; may confirm the plan	• Designs the reward; rarely acts at run time
Required inputs	• Live video and telemetry • Operator commands	• Shared map and task list • Robot reports and data streams	• Occupancy map and frontier estimates • Search-area or hazard priors	• Traversability, semantic and pose distributions	• Team state, task set, energy and link budgets	• Scene description plus a text instruction	• Simulator or logged interaction; a reward
Assumed communication	• Continuous; cable, Wi-Fi or proprietary radio	• Intermittent; designed for loss	• Not needed en route; shared when in range	• None required during execution	• From full connectivity to none, by scheme	• Internet link to a hosted model	• None onboard; Wi-Fi to a central allocator
Key strengths	• A substantial incident record • Human judgment retained throughout	• Breaks the one-operator-one-robot ceiling • Few supervisor interventions per run	• Objectives can encode dose or likelihood • Weighs gain against path or energy cost	• Relaxes risk to escape when trapped • Makes risk an explicit, settable quantity	• Exploits complementary platforms • Adapts to tasks found mid-mission	• Accepts missions as responders state them • Replans online on new observations	• Handles dynamics that resist modeling • Trains in simulation when real trials are costly

Table 7 (continued)

Autonomy paradigm	Teleoperation & variable autonomy	Supervised autonomy	Exploration & coverage planning	Risk-aware & belief-space planning	Multi-agent task allocation	Language-conditioned planning	Learned control policies
Critical limitations	• Does not scale beyond operator count • Fatigue raises errors at mission termination	• Supervisor workload caps fleet size • Reports beyond the link arrive late or never	• Volumetric planners use no target prior • Map drift hampers pruning shared frontiers	• Localization error only inflates variance • Risk tolerance rarely elicited from responders	• Several assume 5G or full connectivity • Not yet validated with responders	• Few verification guarantees • Query latency of seconds or more	• Task-specific; no generalist policies • Sim-to-real gap unquantified on real sites
Typical validation maturity	• Incidents and exercises; interfaces in lab	• Competitions and controlled field trials	• Field trials; radiation cases partly simulated	• Competitions and controlled field trials	• Field trials and simulation	• Simulation to field trials; isolated exercise	• Simulation, then laboratory, arena or field tests
Indicative works	• Rescue task-force deployments (Surmann et al. 2024), Fukushima remote system (Sakaue et al. 2025), immersive teleoperation (Stedman et al. 2023)	• Flexible supervised autonomy (Biggie et al. 2023)	• SubT aerial autonomy (Petrlík et al. 2025), terrain-aware search (Schomer and Adams 2026), nuclear coverage planning (Blake and Schubert 2025)	• STEP (Dixit et al. 2025), semantic belief graphs (Ginting et al. 2023)	• Dynamic task allocation (O'Brien et al. 2023), REACT (Maresca et al. 2025), swarm search and rescue (Horyna et al. 2023)	• SPINE (Ravichandran et al. 2025), LLM behavior trees (Shi et al. 2025), DEXTER-LLM (Zhu et al. 2025)	• Sim-to-real extinguishing (Chaoxia et al. 2025), cooperative deep reinforcement learning search (Hickling et al. 2026), HAZARD (Zhou et al. 2024)

8.7 Learned control and decision policies

Learned control remains confined to narrow, task-specific policies. It continues an earlier behavioral-cloning method, whose control rules cloned the skills of an expert operator (Murphy et al. 2016). The corresponding methods transfer intervention policies from simulation, learn cooperative decisions by reinforcement and train in simulated dynamic environments.

Regarding what makes the task resist hand design and where the training experience is obtained, the main categories are:

- Sim-to-real transfer of an intervention policy: Policies trained against a simulated physical process, since learning on the real process is costly, and then transferred to hardware. An example is an agent rotating a heavy water cannon in discrete steps, from jet images translated into the real domain (Chaoxia et al. 2025). Another one is a smoke-tracking policy mapping segmentation masks to discrete maneuvers, compared with a centroid-based proportional-integral-derivative controller (Pal et al. 2025).
- Reinforcement learning for cooperative decision making: Policies split across decision levels. A graph-attention actor-critic allocates tasks on a ground station, which each drone queries over Wi-Fi for its next target, while the guidance policy runs onboard. For this arrangement, intermittent communication is expected to call for fully onboard autonomy (Hickling et al. 2026).
- Simulated dynamic environments as training substrate: Scenes altered by fire, flood or wind, for comparing learned, rule-based, tree-search and language-model decision pipelines repeatably on one rescue task (Zhou et al. 2024). In another case, radiation exposure varies with time and altitude, while a twin delayed deep deterministic policy gradient agent controls heading and altitude through continuous angular adjustments. However, this agent stagnates in mountainous terrain, where rewards turn sparse (Blake and Schubert 2025).

8.8 Comparative analysis and key insights

Table 7 compares the seven paradigms. The main insights are that: a) The assumed communication regime discriminates the paradigms more sharply than their planning machinery. Teleoperation requires a continuous link, while supervised autonomy is designed for its loss. In contrast, field-tested language-conditioned planners reintroduce a dependence on an internet- or cloud-hosted model (Ravichandran et al. 2025; Cladera et al. 2025; Jarabo-Peñas et al. 2025). b) Uncertainty is treated inconsistently. It is explicit and settable in risk-aware planning, while some exploration and allocation planners carry it as a belief or uncertainty-aware objective. Moreover, it is handled by verifying model output in language-conditioned planning, and it is met in learned control only by randomizing training conditions. c) Learned control remains task-specific. Hence, vision-language-action models outputting low-level closed-loop control after learning from web and robot data (Firoozi et al. 2025) have no counterpart.

9 Learning strategies

Learning strategy concerns how competence is acquired. This weighs heavily in disaster settings, since damage labels are costly to acquire and datasets span a limited range of regions and disaster types. Moreover, models built on them struggle in unseen regions (Ahn et al. 2025). The main strategies, as defined in Section 4, are: a) Supervised learning. b) Label-efficient and weakly supervised learning. c) Generative and synthetic data production. d) Sim-to-real transfer. e) Reinforcement learning. f) Instruction tuning of vision-language models. g) Zero-shot composition and prompting.

9.1 Supervised learning

Supervised learning remains the reference that synthetically pre-trained change detectors still trail (Zheng et al. 2025). Yet, a model so trained reflects the regions and disaster types of its annotated data (Ahn et al. 2025). Its labels come from purpose-built benchmarks, opportunistic imagery, small hand-labeled sets or non-visual and degraded-sensing corpora. In other cases, they are shaped by onboard budgets.

According to the provenance of the labeled data and the constraint that this provenance imposes, the main categories are:

- Purpose-built overhead disaster benchmarks: Corpora assembled deliberately for a disaster task. They offer scale and comparability at the cost of fixing the hazard and the sensing geometry in advance. Examples are global flood maps photointerpreted by SAR experts (Bountos et al. 2024), as well as drone road-damage imagery labeled by non-experts and reviewed in two stages (Manzini et al. 2026).
- Opportunistic and crowd-sourced imagery: Imagery never collected for the purpose. It trades controlled acquisition for street-scale views available within minutes of an event. Examples are earthquake photographs from social media, labeled at pixel level by non-experts supervised by domain experts (Huang et al. 2026).
- Small hand-labeled corpora: A small, unbalanced image set labeled for the study. Models trained on it build on a backbone pre-trained on ImageNet, with augmentation and imbalance-aware losses. An example is the eleven-class segmentation of ground images from search-and-rescue exercises (Salas-Espinales et al. 2023).
- Supervision compressed to an onboard budget: Training shaped by the computational budget of the aerial platform, rather than by benchmark accuracy alone. This is achieved through quantization-aware training (Amesegher et al. 2025). Alternatively, a frame-compression policy is trained once networks fitted to frame-level and then clip-level labels are frozen (Zhao et al. 2024a).
- Non-visual and multi-sensor labeled corpora: Corpora of evidence other than ordinary color imagery, whose construction matters for training. An example is drone audio with separately recorded source and rotor sounds that support augmentation (Gupta et al. 2025). Another one is demining imagery with automatically generated, human-refined bounding boxes (Malizia et al. 2026).
- Corpora of degraded sensing: Corpora recorded under hazard-induced corruption rather than clean imagery. An example is smoke-obscured firefighting video with training, validation and test splits, although it is evaluated zero-shot (Tiwari et al. 2025).

9.2 Label-efficient and weakly supervised learning

Annotating change or damage data is labor-intensive (Huang et al. 2026; Zheng et al. 2024a). Label-efficient and weakly supervised learning addresses this cost. It substitutes a cheaper signal for exhaustive labels, namely physical priors, unpaired single-date images or pseudo labels drawn from foundation models.

According to the source of the supervisory signal that replaces exhaustive annotation, the main categories are:

- Physical priors added to supervision: A known property of the phenomenon regularizing a supervised prediction. In this case, flood maps that violate the ordering imposed by terrain elevation are penalized. Moreover, the same prior propagates annotator labels to neighboring pixels, in order to speed up annotation (Sami et al. 2024).
- Relaxation of the paired-observation requirement: Change learned from unpaired single-date labeled images, so that the cost of labeling change in bi-temporal pairs is avoided (Zheng et al. 2024a).
- Foundation-model pseudo labels in place of target annotation: A source-trained damage model updated at test time on unlabeled target imagery. The update relies on entropy minimization and pseudo labels that draw on an external segmentation model, and that are cleaned at pixel and image level. This counters the error propagation of methods updating themselves on their own predictions (Ahn et al. 2025).

9.3 Generative and synthetic data production

Synthetic data answers the expense of collecting, preprocessing and annotating multi-temporal change imagery at scale (Zheng et al. 2025). The corresponding methods generate plausible change pairs with generative change models, or translate simulated images into the real domain. Moreover, they manufacture dynamic hazard environments in which an agent acts.

According to what is synthesized and what the synthesis must preserve, the main categories are:

- Generative change foundation models: Multi-temporal remote sensing imagery generated by modeling how a scene evolves stochastically. Hence, the synthetic pairs carry a plausible change structure, rather than merely a plausible appearance. The generator is trained by self-supervision on unlabeled single-temporal images (Zheng et al. 2025).
- Translation of simulated images into the real domain: Simulated observations made to resemble real ones before a policy is trained on them. An example is a generative adversarial network trained on simulated and real water-jet images paired by a genetic algorithm, with a jet-shape consistency loss (Chaoxia et al. 2025).
- Synthetic environments rather than synthetic images: Dynamic, hazard-driven environments in which an agent acts, so that what is manufactured is interaction and consequence rather than pixels alone (Zhou et al. 2024).

9.4 Sim-to-real transfer

Sim-to-real transfer acquires disaster-relevant behaviors without disaster data, although simulated environments often lack appropriate robot and environmental dynamics (Murphy et al. 2016; Adams et al. 2024). The reality gap to be closed lies in a physical process, terrain and contact, a changing environment, or platform dynamics and sensing.

According to the nature of the reality gap that must be closed, the main categories are:

- Physical-process gap: A simulator that must reproduce a physical process beyond the robot's own dynamics. An example is an extinguishing strategy whose fluid simulator reproduces real water jets only approximately. In this case, the appearance gap is closed by image translation (Chaoxia et al. 2025).
- Terrain and contact gap: Robot mass, ground friction and external forces randomized while learning over generated terrains. An example is a perceptive legged locomotion controller trained in simulation, and deployed without fine-tuning on real data (Tranzatto et al. 2024).
- Environment-dynamics gap: A simulated environment that changes during the episode. An example is a smoke-tracking policy trained under randomly alternating steady and fluctuating winds. This policy is carried to the real domain through binary smoke masks, which one segmentation model produces in both domains (Pal et al. 2025).
- Platform-dynamics and sensing gap: Guidance and task-allocation policies transferred to several platforms. Gaps are reported in platform momentum, which called for an emergency avoidance layer. Moreover, gaps are reported in depth artifacts from reflective surfaces and visually irregular materials absent from the training domain (Hickling et al. 2026).

9.5 Reinforcement learning

Reinforcement learning is adopted for tasks such as aiming a water jet, whose trajectory is hard to predict. Training takes place in simulation, owing to the cost and risk of real training (Chaoxia et al. 2025; Pal et al. 2025). The rewards are defined over team search, physical intervention, coverage under exposure limits or embodied decisions in changing environments.

According to what the reward is defined over, the main categories are:

- Reward over a team-level search outcome: Team-level guidance and task completion for aerial vehicles, rewarded by dense agreement with a reference action derived from an artificial potential field. This reward made training stable, where a standard distance-based reward was insufficient (Hickling et al. 2026). Moreover, the category covers moving-target search by ground robots, with reward shaping against sparse rewards (Cao et al. 2024).
- Reward over a physical intervention outcome: Water-jet particles hitting a flame target, with a time penalty and a potential-based term on the jet-to-target distance that makes the otherwise sparse reward dense (Chaoxia et al. 2025).
- Reward over coverage under exposure limits: A weighted sum of objectives setting coverage against radiation exposure limits. In this sum, a total-variance coverage term

reduces the fragmentation of unmeasured areas. Moreover, a turn penalty reduces, without removing, excessive maneuvering (Blake and Schubert 2025).

- Reward over embodied decisions in a changing environment: Correct pick-up and drop actions rewarded while fire, flood or wind changes the scene. Hence, the evaluated outcome, namely the rescued value of objects that lose value once damaged, depends on timing (Zhou et al. 2024).

9.6 Instruction tuning of vision-language models

Instruction tuning specializes general vision-language models on disaster or Earth-observation corpora. A single disaster dataset both exposes the gaps of current models and improves them after fine-tuning (Wang et al. 2025). The corpora range from disaster-specific to temporal and multi-sensor ones, while evaluation suites bound what tuning achieves.

According to the breadth of the instruction corpus and what the tuning is intended to implement, the main categories are:

- Disaster-specific instruction corpora: Resources built around damage assessment and response reasoning. In one case, recognition answers are annotated by disaster experts on question variants generated by GPT-4o (Wang et al. 2025). In other cases, corpora are used to fine-tune an existing temporal assistant by low-rank adaptation (Revankar et al. 2025), or a general vision-language model in full (Chen et al. 2025).
- Temporal Earth-observation instruction tuning: Instruction following extended across acquisition dates, by a corpus mixing single-image with temporal tasks. In this way, the temporal skill is added without eroding single-image capabilities (Irvin et al. 2025).
- Multi-sensor and multi-task instruction tuning: Corpora spanning optical, radar and multi-spectral inputs, and a broad task set, with better generalization than generic and domain-specific models (Soni et al. 2025).
- Evaluation suites that bound what tuning achieves: Benchmarks setting models trained on satellite and aerial data against generic ones. The best average accuracy goes to a generic model, while a remote-sensing assistant performs better in single-image, though not bi-temporal, disaster classification (Danish et al. 2025).

9.7 Zero-shot composition and prompting

Zero-shot composition meets annotation scarcity by avoiding training rather than by weakening supervision. An example is a training-free adaptation of a promptable segmentation model, motivated by the extreme cost of change detection labels (Zheng et al. 2024b). In this strategy, frozen models serve as perception front ends, pipeline components, in-mission interpreters or planners.

According to which part of the pipeline the frozen model occupies, the main categories are:

- Frozen perception front ends: A promptable segmentation model (Kirillov et al. 2023) supplying the recognition layer, and acquiring a new competence without any training. An example is a zero-shot change detector that improves on the previous unsupervised

Table 8 Learning strategies: Comparative analysis and key insights.

Learning strategy	Supervised	Label-efficient & weakly supervised	Generative & synthetic data	Sim-to-real transfer	Reinforcement learning	Instruction tuning of VLMs	Zero-shot composition & prompting
Primary function	• Fit a model to labeled disaster data • Serve as the in-domain accuracy reference	• Cut the annotation burden • Substitute a cheaper signal for labels	• Manufacture the missing training material • Spare costly collection and labeling	• Acquire behavior without disaster data • Close the gap to hardware	• Learn behavior from interaction • Optimize what can be scored, not specified	• Specialize a general model to the domain • Produce responder-legible outputs	• Build a capability from frozen models • Operate with no domain training data
Form of the supervisory signal	• Human or recording-derived labels	• Physical priors; single-date or pseudo labels	• Labels or pairings generated with the data	• A reward computed in simulation	• A scalar reward from interaction	• Instruction-response pairs	• None; a prompt at inference time
Disaster data requirement	• High; labeled, ideally in-domain	• Moderate; partial, unpaired or unlabeled	• Low; a seed corpus or a scene model	• None; at most real samples to align domains	• None obtainable at a real site	• High; paired imagery and text	• None
Prerequisite artifact	• An annotation campaign	• A domain prior, source model or single-date labels	• A generative or scene model	• A simulator of the task setting	• A simulator and a reward design	• An instruction corpus with disaster tasks	• A suitable pretrained model
Dominant transfer risk	• Bound to annotated hazards and geographies	• Prior or pseudo supervision may not fit the site	• Synthetic structure may not match reality	• Reality gap unmeasured at real incidents	• Policies hinge on reward and simulator	• Tuning may not carry beyond its corpus	• General-domain prior may hallucinate
Key strengths	• Usually exceeds zero-shot accuracy in domain • Compressible to onboard budgets	• Targets the annotation cost directly • Keeps a domain-specific model	• Scales without new events • Reusable across tasks	• Trains where real trials are impractical • Repeatable and safe	• Rewards effects whose dynamics are hard to model • Learns team task allocation	• One model across tasks and sensors • Benchmarks expose task-level gaps	• No annotation campaign at all • Tasks or targets set by prompt at run time

Table 8 (continued)

Learning strategy	Supervised	Label-efficient & weakly supervised	Generative & synthetic data	Sim-to-real transfer	Reinforcement learning	Instruction tuning of VLMs	Zero-shot composition & prompting
Critical limitations	• Accuracy declines on unseen events • Labeling is labor-intensive	• Labels reduced, not removed • Tested on overhead imagery only	• Fidelity untested at real incidents • Idealized physics; gap to full supervision	• No policy exercised at a real incident • Simulator fidelity limits transfer	• Trained in simulation throughout • Sparse rewards hinder training	• Corpora need manual curation • Tuned models uneven across tasks and sensors	• Below supervised accuracy in perception • Prompt-dependent; no tuning on disaster data
Typical validation maturity	• Benchmarks; some field and operational use	• Benchmark evaluation, offline	• Benchmarks, simulation, laboratory hardware	• Laboratory to controlled field trials	• Simulation; limited hardware	• Benchmark evaluation, offline	• Benchmarks to controlled field trials
Indicative works	• EIDSeg (Huang et al. 2026), road damage benchmark (Manzini et al. 2026), DroneAudioSet (Gupta et al. 2025)	• EvaNet (Sami et al. 2024), STAR (Zheng et al. 2024a), DAVI (Ahn et al. 2025)	• Changen2 (Zheng et al. 2025), HAZARD (Zhou et al. 2024), JetGAN (Chaoxia et al. 2025)	• Sim-to-real extinguishing (Chaoxia et al. 2025), cooperative deep reinforcement learning search (Hickling et al. 2026), smoke-plume tracking (Pal et al. 2025)	• Cooperative deep reinforcement learning search (Hickling et al. 2026), nuclear coverage planning (Blake and Schubert 2025)	• DisasterM3 (Wang et al. 2025), TEOChat (Irvin et al. 2025), EarthDial (Soni et al. 2025)	• Segment Any Change (Zheng et al. 2024b), vision-LLM navigation (Yamanaka et al. 2025), SPINE (Ravichandran et al. 2025)

state of the art. In contrast to it, unsupervised methods are re-trained on each new data distribution (Zheng et al. 2024b).

- Frozen components composed into a disaster pipeline: Prompted, unmodified models arranged and configured by role and output prompts, in order to generate subtasks without additional training. Such a pipeline was found less accurate in simulated than in real scenes, which was attributed to texture deficiency and geometric distortion (Shi et al. 2025).
- Frozen models as in-mission interpreters: A frozen model queried during the mission, in place of vision models whose exhaustive training and labeling are judged too costly. Its decisions are shaped by prompt design alone. The prompts include hints computed by deterministic algorithms, as well as chain-of-thought reasoning, which is shown to improve control accuracy (Yamanaka et al. 2025).
- Frozen models as planners over a grounded representation: Unmodified models configured through prompts alone. In one case, a system prompt defines the planner's role, while a graph-editing interface brings map updates into context (Ravichandran et al. 2025). In another case, prompts are used without task-specific retraining, while retraining on real operator language is a possible improvement (Bark et al. 2025). Moreover, multi-stage prompts share core instructions and robot descriptions, with the meta-policy stage proving most critical in ablation (Zhu et al. 2025).

9.8 Comparative analysis and key insights

Table 8 compares the seven learning strategies. The main insights are that: a) The strategies differ chiefly in the artifact that supplies or replaces labeled disaster data. This artifact is an annotation campaign, a domain prior or cheaper label source, a generative model, a simulator, an instruction corpus or a pre-trained model obtained elsewhere. b) Frozen general-purpose models and weakened supervision answer the same annotation cost. The former import a model trained on other data, while the latter retain a task model trained on a cheaper signal. c) A model can still be adapted to the incident after training. This is achieved by a test-time update on unlabeled target imagery (Ahn et al. 2025), or by prompting a frozen model. For instance, single clicks turn class-agnostic change masks into object-centric ones for counting destroyed buildings (Zheng et al. 2024b). d) The transfer risks differ in kind rather than degree. They range from annotation coverage, prior validity and synthetic fidelity to the limited reach of tuning and the hallucinations of prompted general-domain models.

10 Foundation model types

Foundation model type concerns the large pre-trained model, if any, around which a method is built. The modalities of the model bound the part of the problem it can serve. For instance, language models may facilitate task specification from high-level instructions, whereas vision-language models specialize in the analysis of visual data (Firoozi et al. 2025). The main types, as defined in Section 4, are: a) Not foundation-model-based. b) Vision foundation models. c) Large language models. d) Vision-language models.

10.1 Not foundation-model-based

Methods that are not foundation-model-based vary from a manually teleoperated ground robot opening a valve at a chemical plant to an Attention UNet labeling roads during two hurricane responses (Lichtenfeld et al. 2025; Manzini et al. 2026). They draw on model-based estimation, classical planning and control, task-specific networks, purpose-built hardware or integrated field systems.

Regarding what the method is built out of in place of a pre-trained model, the main categories are:

- Model-based estimation and geometric reasoning: Competence drawn from an explicit sensor and motion model rather than from learned priors. An example is localizability analysis of scan-to-map correspondences, free of the extensive ground-truth data on which most learned alternatives rely (Tuna et al. 2024). Moreover, radar SLAM combines Doppler odometry with scan and feature matching (Zeng et al. 2024). Virtually all subterranean competition teams also relied on factor- or pose-graph LiDAR pipelines (Ebadi et al. 2024).
- Classical planning and control stacks: Decision layers designed rather than trained. They are built from conditional value-at-risk traversability with model predictive control (Dixit et al. 2025) and a policy over a semantic-belief roadmap (Ginting et al. 2023). Moreover, they comprise a consensus auction (O'Brien et al. 2023) and risk-aware planning over a sphere-based map (Petrlík et al. 2025).
- Task-specific supervised networks: Learned perception trained for one task and one label set, even when initialized from weights pre-trained on other datasets (Salas-Espinales et al. 2023; Manzini et al. 2026; Huang et al. 2026; Amessegher et al. 2025).
- Purpose-built mechanisms and hardware: Contributions of mechanical design, radiation-tolerant hardware or model-based control. In these systems, the robot is commanded by a remote human operator, aided where needed by optimization-based whole-body control, rather than by any learned model (Hong et al. 2023; McFarland et al. 2024; Sakaue et al. 2025).
- Integrated field systems: Complete stacks in which learned perception enters as artifact detectors. These detectors are trained on images that the team collected in basements, tunnels, mines and caves (Tranzatto et al. 2024), or with phones and robot cameras (Kottege et al. 2025). In another case, convolutional detectors, among them a pruned YOLOv5m6, run at a low rate on limited robot processors (Morrell et al. 2024).

10.2 Vision foundation models

Vision foundation models are adopted mostly zero-shot, where no domain-specific annotation exists. Their promptable formulation lets the categories that matter be specified at run time (Cladera et al. 2025; Zheng et al. 2024b). They serve as frozen segmentation front ends, promptable change detectors, components of disaster pipelines or generators of training data.

Regarding the role the vision foundation model plays and the extent to which it is adapted, the main categories are:

- Frozen promptable segmentation as a general front end: A general-purpose promptable segmentation model (Kirillov et al. 2023) taken unmodified as the perception input of a larger system. For instance, SAM2, chosen for its zero-shot open-set performance on aerial data, segments Grounding DINO detections on an aerial mapper (Cladera et al. 2025). Moreover, OWL-ViT boxes of doors and obstacles feed a behavior tree (Shi et al. 2025). In another case, Grounding DINO detections are used without a segmentation model in a language-driven ground-robot planner (Ravichandran et al. 2025).
- Promptable change detection: The promptable formulation extended from single images to image pairs. Without any training, it carries the promptability and zero-shot generalization of the model over to the pre- and post-event images that damage assessment compares (Zheng et al. 2024b).
- Composition into a disaster-specific pipeline: A frozen segmentation model composed into a disaster pipeline. In this pipeline, SAM, picked among several promptable segmentation models for its simplicity, adds task-agnostic knowledge to a source-trained change model. Only the latter is updated for the target region (Ahn et al. 2025).
- Generative vision foundation models: The pre-trained model producing training material rather than predictions. An example is a diffusion transformer generating image time series with change labels, in order to pre-train change detectors that then operate zero-shot on unseen data (Zheng et al. 2025).

10.3 Large language models

Large language models are chosen for their contextual reasoning. Moreover, they serve as an interface that lets operators without robotics expertise state a mission in natural language (Ravichandran et al. 2025; Jarabo-Peñas et al. 2025). The model emits plans over a semantic map, executable structures, team-level allocations with explanation, or interpretations of operator commands.

Regarding what the model emits and where that output enters the autonomy stack, the main categories are:

- Plans over a semantic representation: A general-purpose model coupled to a semantic map built during the mission. In one case, the base GPT-4 plans over the map serialized as text, while a validator checks each plan against it (Ravichandran et al. 2025). The same planner, run with GPT-4o, also reasons over semantic-metric maps that an aerial and a ground robot build online and share opportunistically (Cladera et al. 2025). In another case, GPT-4o mini, alternatively Mixtral 8x7B Instruct, merely turns the instruction into structured tasks and constraints. These are matched to a scene graph and optimized by graph search (Bark et al. 2025).
- Executable structures: Separately role-prompted models for scene description, action selection from a fixed list and translation into behavior-tree function calls. This arrangement raised instruction execution accuracy more than fourfold over a vision-language model used alone (Shi et al. 2025). In contrast, a model that is not fine-tuned picks one of the available high-level actions at each decision step, from a textual description of the observations in a fire, flood and wind simulation benchmark (Zhou et al. 2024).

- Team-level allocation with explanation: Staged prompts, through which the model decomposes work into subtasks that a search-based optimizer assigns to several platforms. The model's reasoning is exposed in a layered graph for human verification. This design remained effective when DeepSeek-V3 was replaced by GPT-4o, Qwen-2.5-max or Grok-3 (Zhu et al. 2025).
- Command interpretation at the operator interface: Four role-specific agents at the command layer, sharing one cloud-hosted 70-billion-parameter Llama model. Each agent is given system prompts and deterministic tools. The model size would also permit local deployment where connectivity is compromised (Jarabo-Peñas et al. 2025).

10.4 Vision-language models

Vision-language models interpret disaster scenes jointly from images and text, mostly offline on overhead or recorded imagery (Wang et al. 2025; Tiwari et al. 2025). The corresponding works comprise disaster-specific corpora that test and then specialize them, temporal and multi-sensor Earth-observation models, as well as benchmarks bounding their geospatial capability. In contrast, few works let their output steer a platform on site.

Regarding the modality span of the model and whether its interpretation is embodied, the main categories are:

- Disaster-specific multi-modal corpora and diagnostics: Resources built to test and then to specialize models. An example is an evaluation of open-source, commercial and remote sensing models followed by fine-tuning, which improves question answering, reports and referring segmentation (Wang et al. 2025). Moreover, change captions are generated by a proprietary reasoning model as the scoring reference (Chen et al. 2025). In addition, untuned models are evaluated on omnidirectional firefighting video, where their accuracy rises once the panoramic input is rectified (Tiwari et al. 2025).
- Temporal and multi-sensor Earth-observation models: General vision-language architectures instruction-tuned on Earth-observation corpora that include building damage. An example is a Video-LLaVA initialization that outperforms GPT-4o and Gemini 1.5 Pro on several temporal tasks (Irvin et al. 2025), and is later tuned further on a corpus of disaster events (Revankar et al. 2025). Another one is a compact InternViT encoder with a Phi-3-mini language model, tuned in three stages, the last for multi-spectral and radar data (Soni et al. 2025).
- Benchmarks bounding geospatial capability: Generic, commercial and geospatial-specific models scored on one suite, where no single model leads everywhere. LLaVA-OneVision is strongest at localization and counting, GPT-4o at object classification and Qwen2-VL at earthquake estimation from non-optical imagery (Danish et al. 2025). Moreover, on low-altitude disaster imagery, supervised classifiers outperform a zero-shot LLaVA-NeXT on nearly all classes, while remaining competitive with GPT-4o (Scheele et al. 2025).
- Embodied vision-language interpretation: The model's output steering what a platform does on site, rather than supporting offline analysis. In one case, a cloud-hosted GPT-4o is guided by customized prompts, including a deliberation among three personas before each waypoint and velocity command (Yamanaka et al. 2025). In another case, LLaVA

Table 9 Foundation model types: Comparative analysis and key insights.

Foundation model type	Not foundation-model-based	Vision foundation models	Large language models	Vision-language models
Primary function	• Deliver the capability from designed components and task-specific models • Span perception, estimation, planning, control and mechanisms	• Supply recognition without domain annotation • Segment or detect change on prompt	• Turn responder instruction into executable goals • Decompose and allocate missions	• Interpret imagery jointly with text • Describe, answer and reason about a scene
Modalities spanned	• Whatever the sensors provide; no shared representation	• Vision only, plus a spatial or textual prompt	• Language only; perception arrives pre-abstracted	• Vision and language jointly; optical, radar and temporal streams
Contribution to the system	• Any layer, up to complete fielded systems	• A perception front end, or synthetic supervision	• A planning and interface layer above a conventional stack	• An interpretation and reporting layer; rarely, navigation decisions
Adaptation regime	• Designed, or trained for one task and label set	• Mostly frozen and prompted; occasionally composed or fine-tuned	• Frozen and prompted; no disaster fine-tuning observed	• Instruction-tuned on Earth-observation or disaster corpora, or prompted without tuning
Verifiability of the output	• High for designed components, lower for learned perception	• Moderate; masks are inspectable but scored on few disaster sets	• Low, unless the output is a checkable structure	• Low for free text; masks and multiple-choice answers scorable
Key strengths	• Sole class deployed at real incidents • Tailored to its sensors and platform	• Generalizes to unseen change types and regions • Task-agnostic prior from billion-scale data	• Takes missions in the responders' own terms • Copes with incomplete mission specifications	• Long-form disaster reports in natural language • Dedicated disaster-reasoning benchmarks
Critical limitations	• Every capability must be built separately for its task • Brittle outside its design envelope	• Class-agnostic masks; damage semantics added separately • Zero-shot accuracy trails supervised models	• Weak guarantees of correct output • Mostly cloud-hosted, so a link is assumed	• Off-the-shelf models struggle on disaster tasks • Cross-sensor gaps; weak damage counting
Typical validation maturity	• Simulation and benchmarks to field trials, competitions and incidents	• Benchmark evaluation, offline	• Simulation to field trials; an isolated rescue exercise	• Mostly offline evaluation; isolated controlled field trials
Indicative works	• CERBERUS (Tranzatto et al. 2024), X-ICP (Tuna et al. 2024), STEP (Dixit et al. 2025)	• Changen2 (Zheng et al. 2025), Segment Any Change (Zheng et al. 2024b), DAVI (Ahn et al. 2025)	• SPINE (Ravichandran et al. 2025), DEXTER-LLM (Zhu et al. 2025), voice-commanded UGV (Jarabo-Peñas et al. 2025)	• DisasterM3 (Wang et al. 2025), TEOChat (Irvin et al. 2025), GEOBench-VLM (Danish et al. 2025)

is queried through a semantic configuration that a GPT-4o-based planner sets for the mission (Cladera et al. 2025).

10.5 Comparative analysis and key insights

Table 9 compares the four foundation model types. The main insights are that: a) The pre-trained models occupy mainly the two ends of the autonomy stack, entering as perception front ends or as planning and reporting layers. In contrast, the layers between them, namely estimation, control and physical interaction, stay almost wholly conventional. b) Adoption is largely shallow. No large language model is fine-tuned on disaster data, while vision foundation models are mostly frozen, prompted or training-free. Vision-language models, however, have been tuned on Earth-observation and disaster-specific corpora. c) Verifiability decreases as output abstraction rises. Hence, several language-based systems interpose an automatic plan validator or an operator confirming each output before execution. d) Off-board dependence lies mainly with the models that plan or select actions. This ties adoption to site connectivity, unless the models are hosted locally. e) Given the absence of any vision-language-action model or generalist robot policy, the pre-trained models present in this field inform decisions or issue goals, plans or behavior-tree calls that conventional planners and controllers execute. Apart from waypoint and velocity selection for tunnel navigation (Yamanaka et al. 2025), they do not set the motion of the platform themselves.

11 Human-robot interaction and teaming

Human-robot interaction and teaming concerns the modality by which humans and robots trade intent and information. In this respect, human error accounts for a large share of robot failures, while user interfaces are key to situational awareness (Murphy et al. 2016). Moreover, robots are accepted as teammates only once personnel trust them (Adams et al. 2024). The main modalities, as defined in Section 4, are: a) Teleoperation interfaces. b) Immersive and digital-twin supervision. c) Variable and sliding autonomy. d) Natural-language command. e) Operator state, workload and error. f) Responder requirements and trust. g) Public-facing emergency interaction.

11.1 Teleoperation interfaces

Teleoperation interfaces determine how an operator perceives a remote disaster site and steers a robot through it. Flat-screen camera feeds and telemetry are used at a contaminated facility and in task-force missions (Sakaue et al. 2025; Surmann et al. 2024). In contrast, immersive reconstructions and haptic feedback extend the operator's perception.

According to the sensory channel through which the operator perceives the remote site, the main categories are:

- Conventional screen-based control: Camera feeds and telemetry on a flat display. An example is an operation-room interface exchanging commands and sensor information with a robot in a contaminated facility (Sakaue et al. 2025). Another one is a two-monitor command-vehicle workstation for steering a ground robot, together with a tablet giving its operator a drone's view of the robot (Surmann et al. 2024). In contrast, steering a vine

robot by joystick from its tip-camera image alone proved unintuitive. The reasons were the camera's limited field of view and an image rotating out of alignment with gravity as the tip steered (McFarland et al. 2024).

- Immersive visual interfaces: The operator placed inside a reconstructed view of the site. With such a view, trained robotics operators completing a nuclear inspection and radiation-sampling course collided less, but took longer. According to physiological data, they had a lower cognitive workload, but an increased physical demand (Stedman et al. 2023). Moreover, teleoperators of a response robot with a manipulator finished faster and reported lower mental workload in virtual reality (Illing et al. 2025).
- Haptic and multi-sensory feedback at the effector: Perception extended to the point of contact rather than to the viewpoint. In this way, the operator feels what the robot touches through hand controllers and vibrotactile arm sleeves. Moreover, the operator is informed of the explosive residues that the robot's chemical sensor detects (Xiao et al. 2023).

11.2 Immersive and digital-twin supervision

Unlike an immersive teleoperation interface, a digital twin or shared operational picture persists independently of any one robot. In this way, several platforms are managed within one model of the site (Baniqued et al. 2024; Frering et al. 2023). Examples range from a nuclear facility twin to a wildfire operational picture and the map of an underground mine.

A twin of a nuclear facility built from building information models lets its operator move between observing the autonomous fleet and teleoperating one of its robots, on a flat screen or in virtual reality (Baniqued et al. 2024). On a wildfire common operational picture, a commanding officer creates and prioritizes goals for aerial and ground robots. In addition, the officer exchanges located messages and photos with firefighters (Frering et al. 2023). Moreover, through the 3D map of an operational underground mine, a single non-expert worker tasks a whole aerial fleet by inspection type and location (Dahlquist et al. 2025).

11.3 Variable and sliding autonomy

Variable autonomy raises the question of the evidence on which authority shifts. Adapting to human teammates in real time requires estimating their workload, fatigue and stress (Adams et al. 2024). The shifts are triggered by task difficulty and operator performance, by the communication state, or by operator override of transparent assistance.

According to what triggers the shift of authority, the main categories are:

- Task difficulty and operator performance: Shifts tied to how the current allocation performs, namely: a) A stair-climbing behavior taking over from the operator once the robot's pitch exceeds a threshold. Autonomous local navigation had outperformed teleoperation on extreme terrain (Kottege et al. 2025). b) A copilot verifying task execution and notifying the human when assistance is required (Morrell et al. 2024). c) A single supervisor moving across competition events from directing each robot toward coordinating the group. The context switch of selecting and first querying each robot remained the principal limitation (Tranzatto et al. 2024).

- Communication state and pending reports: The link and the reports awaiting delivery, alongside the supervisor's commands, determining the robot's behavior. For instance, a robot returns to communication range once three artifacts are unreported, or once the first one has waited five minutes (Biggie et al. 2023). In another case, manual teleoperation ran over the only radio covering the range at a chemical plant. There, the operators proceeded with extreme caution, while waypoint navigation would have let them focus more on situational awareness (Lichtenfeld et al. 2025).
- Operator override of transparent assistance: Requirements from responder workshops that deviations of an assistance function from the operator's command be visualized and open to override, and that any risk the robot takes be transparent (Daun and von Stryk 2023). Moreover, systems let the operator's knowledge override autonomous choices. For instance, priority regions push a ground team toward important areas or away from dangerous ones (O'Brien et al. 2023). Commands can also bias, bypass or halt an aerial vehicle's automatic goal selection, until a resume command cancels them (Petrlík et al. 2025). Furthermore, points of interest send a rescue robot toward a missing person's last known location, a concept that still requires an expert operator (Bark et al. 2023).

11.4 Natural-language command

Natural-language command lets non-expert operators assign tasks to a robot without additional training (Jarabo-Peñas et al. 2025). However, ambiguous terms risk hindering the matching of commands to entities (Bark et al. 2025). Language serves as a substitute for manual control, a co-active planning partner, a team-level instruction channel or a prompt-level task interface.

According to the role language plays between operator and robot, the main categories are:

- Speech as a substitute for the manual interface: Voice commands from operators without robotics expertise, relayed through the existing command hierarchy. In a field exercise with over 150 first responders, such commands directed an autonomous ground vehicle in three missions (Jarabo-Peñas et al. 2025).
- Language as a co-active planning partner: Language grounded in a semantic representation of the site. The plan is refined with operator feedback (Bark et al. 2025), or resolved from an under-specified mission. In the latter case, the user is asked for clarification, or invited to re-task the robot when planning fails (Ravichandran et al. 2025).
- Language as a team-level instruction channel: Language addressing several platforms. In one case, a team mission is stated in templated natural language, and every planning output is verified by an operator (Zhu et al. 2025). In another case, a ground robot reports its findings in language, negative ones included, and receives multiple operator updates in a single demonstration (Cladera et al. 2025).
- Language as a prompt-level task interface: Task requirements and constraints written into the prompt of a multi-modal model that chooses navigation and inspection commands. Hence, an appended instruction, such as avoiding a ditch seen in the camera image, changes the robot's behavior. Moreover, each command is returned with its rationale (Yamanaka et al. 2025).

11.5 Operator state, workload and error

Studies of operator state, workload and error make the human, rather than the interface, the subject of examination (Murphy et al. 2024; Mehta et al. 2024). They document errors in real disaster flights and scored competition runs, as well as the degradation of team performance under fatigue. Moreover, they supply instrumented datasets pairing operator behavior with physiological and self-reported measures of cognitive state.

According to whether the work measures the operator, the operational outcome, or supplies the data to model both, the main categories are:

- Errors observed in real operations and competitions: Human and technological failures documented during actual disaster flights (Manzini et al. 2023b). Human error occurs in missions, causing mission failures and delaying data products (Murphy et al. 2024). Moreover, in scored competition runs, the interface and cognitive load of a team's single supervisor at times limited its performance (Chung et al. 2023).
- Performance degradation under fatigue: Naturalistic exercise studies of how a human-robot team's vulnerabilities change under fatigue. These vulnerabilities differ by mission phase and crew role, over a 24-hour replication of real deployment operations (Mehta et al. 2024).
- Instrumented datasets of behavior and cognitive state: Resources pairing human-drone collaboration in simulated response tasks with heart rate, pupillometry and self-reported workload, trust and engagement (Orlov Savko et al. 2024). Another resource records operators supervising one or two simulated search-and-rescue robots, with eye tracking, biosignals and labeled delays in recognizing that a robot needs assistance (Senaratne et al. 2026).

11.6 Responder requirements and trust

Responder studies approach the gap between demonstrated capability and accepted practice from the side of the user. Adoption relies on ongoing partnerships with practitioners, while such studies have recently begun to supply quantified accuracy thresholds (Manzini et al. 2023a; Ruan et al. 2025). They elicit requirements directly from responders, by questionnaire on accuracy, or from established human-animal teamwork.

According to how the requirement is elicited, the main categories are:

- Direct elicitation from responders: Workshops with first responders, deployment reports and an acceptance model, under which assistance functions are accepted only if they offer more than established tools while remaining easy to learn and use (Daun and von Stryk 2023). Moreover, professional and volunteer firefighters evaluated a sniffer robot operated by a pilot and a data spotter, under a new standard operating procedure with a commanding officer. They praised the team structure, while the limited situational awareness of the simplified setup remained a serious issue (Steinbauer-Wagner et al. 2025).

Table 10 Human-robot interaction and teaming: Comparative analysis and key insights.

Interaction modality	Teleoperation interfaces	Immersive & digital-twin supervision	Variable & sliding autonomy	Natural-language command	Operator state, workload & error	Responder requirements & trust	Public-facing emergency interaction
Primary function	• Let an operator perceive and drive a remote platform • Sustain the operator's situational awareness	• Supervise a fleet through a shared representation • Decouple supervision from single video feeds	• Reallocate authority during the mission • Match control to difficulty and link state	• Command robots in the operator's own words • Specify missions at the level of intent	• Measure the human cost of operation • Establish what actually goes wrong	• Establish what responders need and will rely on • Set acceptance thresholds	• Interact with the affected public • Convey robot state to non-operators
Human involved	• A remote operator, often with support roles	• A supervisor over a fleet	• An operator or supervisor sharing authority	• Operators, not only robotics experts	• The operator, as the subject of study	• Responders and handlers as informants	• Evacuees and arriving crews
Interaction channel	• Screens, immersive displays, haptics	• A synchronized virtual model or map	• An authority-allocation mechanism	• Speech and text	• Measurement instruments and logs	• Workshops, questionnaires, interviews	• Motion, signaling, physical presence
Direction of information flow	• Bidirectional and continuous	• Aggregated from the fleet; commands back	• Authority passes both ways	• Human to robot, with clarification back	• From the human to the researcher	• From responders to designers	• Two-way with evacuees, one-way to crews
What is evaluated	• Task time, workload, situational awareness	• Usability via expert heuristics and surveys	• Mission results and supervisor interventions	• Plan validity, mission success, interactions	• Error rate, workload, fatigue effects	• Stated need and reliance thresholds	• Compliance, clarity, perceived safety
Key strengths	• Used in real incident operations • Immersive views cut errors in studies	• One view of a heterogeneous fleet • Persists independently of any robot	• Keeps a human available without constant control • Informed by competition experience	• No training on a bespoke interface • Operator can confirm or correct plans	• Locates errors by mission phase • Grounded in real or replicated operations	• Converts acceptance into a design target • Draws on practitioners' field experience	• Plans for people who may not comply • Brings non-operators into design

Table 10 (continued)

Interaction modality	Teleoperation interfaces	Immersive & digital-twin supervision	Variable & sliding autonomy	Natural-language command	Operator state, workload & error	Responder requirements & trust	Public-facing emergency interaction
Critical limitations	• Onboard view alone often insufficient • Immersive gains not shown at incidents	• Rests on few systems • Usability judged by few experts and firefighters	• Supervisor load still limits performance • Switching cost to the operator rarely measured	• Unverified plans may look valid yet fail • Limited tolerance of varied phrasing	• Proposed countermeasures left unevaluated • Concentrated on aerial operations	• Thresholds not yet tested against systems • Small responder samples	• Few public-facing works; an open problem • No field validation at an incident
Typical validation maturity	• Laboratory studies; field trials; real incident deployments	• Simulated mission; field tests with firefighters and in a mine	• Simulation, field trials, competitions and a real deployment	• Simulation, field trials, an exercise	• Real incidents, exercises, competitions, simulation	• Elicitation without systems; a hazmat exercise	• Virtual environment and interviews
Indicative works	• Immersive teleoperation (Stedman et al. 2023), virtual reality vs. screen control (Illing et al. 2025), EOD haptics (Xiao et al. 2023)	• Nuclear digital twin (Baniqued et al. 2024), wildfire operational picture (Frering et al. 2023), mine inspection interface (Dahlquist et al. 2025)	• Terrain-triggered takeover (Kottege et al. 2025), flexible supervised autonomy (Biggie et al. 2023), single-supervisor role shift (Tranzatto et al. 2024)	• Voice-commanded UGV (Jarabo-Peñas et al. 2025), LLM copilot (Bark et al. 2025), DEXTER-LLM (Zhu et al. 2025)	• Human errors at Hurricane Ian (Murphy et al. 2024), fatigue study (Mehta et al. 2024), RW4T dataset (Orlov Savko et al. 2024)	• Responder thresholds (Ruan et al. 2025), elicited requirements (Daun and von Stryk 2023), rescue dog teams (Kleiser et al. 2026)	• Evacuation strategies (Hu et al. 2023), safe to approach (Lee et al. 2023)

- Elicited accuracy thresholds for trust and use: The accuracy that machine-produced semantics must reach in the eyes of responders, elicited by questionnaire (Ruan et al. 2025).
- Study of established human-animal practice: Interviews with search-and-rescue dog handlers, deriving robot roles and design features from a teamwork that relies on mutual responsiveness and nonverbal communication. Robots are expected to complement this teamwork rather than replace it (Kleiser et al. 2026).

11.7 Public-facing emergency interaction

Rescue robots must interact with victims and other responders in front of the robot, as well as with the operator behind it (Murphy 2014). Yet, few works address the affected public and other people at the scene, namely occupants during an evacuation and emergency crews approaching an autonomous vehicle.

According to who the non-operator human is and what the robot must convey to them, the main categories are:

- Interaction with occupants during an evacuation: Automatically synthesized nudging and waiting strategies for evacuating non-compliant users, compared in a simulated-environment study. It is the single work of the category that addresses the affected public (Hu et al. 2023).
- Signaling intent to arriving emergency crews: Interviews on the protocols by which an autonomous road vehicle signals that it is parked and safe to approach. These protocols concern first responders, who can no longer rely on the procedures used for human-driven vehicles when no driver is present (Lee et al. 2023).

11.8 Comparative analysis and key insights

Table 10 compares the seven interaction and teaming modalities. The main insights are that: a) The modalities divide into those that build an interface and those that study the humans who must use one. The remedies proposed by the operator studies, such as an intelligent assistant for mission configuration (Murphy et al. 2024) and adaptive interfaces for fatigued crews (Mehta et al. 2024), rarely reappear as interface requirements. b) Research thins at both ends of the range of humans involved. Shared supervisory representations rest on few systems, while interaction with the affected public rests on isolated work. c) Acceptance has been made quantitative, reporting the accuracy at which machine-produced semantics would be trusted and used. d) Natural-language command (Jarabo-Peñas et al. 2025; Bark et al. 2025; Ravichandran et al. 2025) addresses the reliability of machine-produced information, on which responders make their trust conditional (Ruan et al. 2025).

12 Coordination, communication and computational deployment

Coordination, communication and computational deployment concern how many platforms act together, how they exchange information and on which hardware their computation runs. These aspects decide whether a capability can be fielded, since disasters typically destroy

the communication infrastructure that research systems often assume (Murphy et al. 2016; Steinbauer-Wagner et al. 2025). The main configurations, as defined in Section 4, are: a) Single-robot operation. b) Heterogeneous teams. c) Centralized coordination and allocation. d) Decentralized swarms. e) Communication-constrained operation. f) Onboard computation. g) Offboard and cloud-dependent inference.

12.1 Single-robot operation

Single-robot operation is a configuration in its own right, rather than the degenerate case of a team. Indeed, few robot teams have been deployed, and only one of them was explicitly coordinated (Murphy et al. 2016). In this configuration, a single platform commits to a localized hazard, surveys a bounded area or answers to one operator. In other works, single-platform operation is documented in operational records of real events.

Regarding the scope of the mission entrusted to a single platform and how its operation is organized or documented, the main categories are:

- Commitment to a single localized hazard: A hazard confined to one location, where access rather than coverage binds. In this case, a single ground platform carries the task, with any second platform in support. Examples are a cable-tethered robot at Fukushima Daiichi carrying no batteries or radio (Sakaue et al. 2025), and a teleoperated valve-opening vehicle reached through two radio mesh repeaters, while a drone observed from outside (Lichtenfeld et al. 2025). Moreover, a sniffer robot processes and serves its readings onboard, without external infrastructure (Steinbauer-Wagner et al. 2025).
- Bounded single-platform survey: One platform surveying a delimited area on its own. Examples are a borehole-deployed drone mapping mine voids on its onboard computer (Brown et al. 2026), and a mine UAV flying its gas survey without communication once its waypoints are received (Nordström et al. 2025). Moreover, the category includes a single helicopter or UAV under radiation-aware coverage planning (Blake and Schubert 2025), a humanoid exploring indoors (Shi et al. 2025) and an autonomous underwater vehicle (AUV) searching (Zhou et al. 2026). Finally, a surface vessel streams sonar data to a shore computer (Rymansaib et al. 2023).
- One platform per operator: One platform under the control of one operator, who teleoperates it or switches it between manual and autonomous modes. An example is an indoor firefighting robot exchanging information with firefighters in a safe area through ultra-wideband beacons (Fernández Talavera et al. 2023). Another one is an EOD robot whose teleoperator receives its stereo video over TCP/IP (Xiao et al. 2023).
- Single-platform operational records: What small aircraft flown by separate squads achieved at a real event. In this event, the lack of centralized control of uncrewed and crewed aircraft caused operational friction (Manzini et al. 2023b). Moreover, data managers received most mission data at once (Murphy et al. 2024).

12.2 Heterogeneous teams

Heterogeneous teams combine platforms of different morphologies, so that each covers what the others cannot reach (Kottege et al. 2025; Petrlík et al. 2025). They rely on unified ground-aerial stacks, aerial agents beyond ground mobility or architectures presenting a team as one

system. Moreover, they are tasked through natural-language missions, or assembled from the assets available.

Regarding how the different platforms of a team are integrated, tasked and linked to one another, the main categories are:

- Unified ground-aerial stacks: Legged, wheeled or tracked, and aerial platforms sharing one mapping, planning and communication architecture. In such stacks, the robots themselves carry and deploy the radio nodes of the mesh linking them to the base station. This deployment takes place on operator command (Kottege et al. 2025; Tranzatto et al. 2024) or autonomously (Morrell et al. 2024).
- Cooperative aerial agents beyond ground mobility: Cooperating aerial agents exploring the flyable space that less mobile ground vehicles cannot reach. Meanwhile, ground robots extend the network with breadcrumbs. In an earlier circuit, one aerial robot also landed at a strategic position to relay for the others (Petrlík et al. 2025).
- Architectures presenting a team as one system: Architectures making a robot team operable as a single system. An example consists of two tracked robots, one exploring and one carrying a manipulator, which share perception and localization. They are commanded from two operator control units, and they release the relays of their own wireless mesh as they advance (Wang et al. 2023). Another one is a fleet of multiple heterogeneous robots relaying data wirelessly to a base-station computer, which hosts the Robot Operating System master node (Baniqued et al. 2024).
- Language-specified team missions: Teams tasked through a natural-language mission. An example is a fleet of aerial and ground robots coordinated centrally (Zhu et al. 2025). In another case, an aerial and a ground robot opportunistically share a sparse semantic graph, kept compact because of their limited bandwidth. The aerial robot also relays messages between the robots and the operators (Cladera et al. 2025).
- Teams assembled from available assets: Team composition varying with the task and the circumstances. For instance, camera copilots were added for multi-robot collaboration in a task force's field missions (Surmann et al. 2024). In contrast, a purpose-designed ground-aerial support team is dispatched by a central task management system over an existing network (Frering et al. 2023).

12.3 Centralized coordination and allocation

Centralized coordination and allocation assign robots to tasks from a single decision point (Maresca et al. 2025; Dahlquist et al. 2025). This presupposes that every robot reaches the decision point often enough. The planner optimizes coverage under an energy budget, reallocates as the mission evolves, pools measurements of a hazard field or delegates to a language model.

Regarding what the central planner optimizes and which resource binds it, the main categories are:

- Allocation under an energy budget: Edge-hosted orchestrators merging the data streamed by every robot into one global occupancy grid. They plan coverage jointly with each robot's battery level. Hence, deferred blind-spot targets are visited only when the remaining operation time and battery suffice (Maresca et al. 2025).

- Reallocation as the mission evolves: The assignment recomputed whenever new tasks or observations invalidate it. For instance, a central auctioneer reallocates operator-added inspection tasks, while exchanging only task broadcasts, bids and assignments over a mobile Wi-Fi mesh (Dahlquist et al. 2025). Moreover, learned multi-agent search policies presuppose a team-wide view. In these policies, each agent's input map marks the positions of all other agents, without specifying how these positions reach the agents (Cao et al. 2024).
- Centralized planning over a hazard field: Distributed measurements of a shared physical field pooled under centralized estimation and planning. An example is multi-source gamma localization, fusing the Compton-camera measurements of multiple aerial vehicles into one real-time onboard estimate. In this case, one drone plans for the team, on the assumption that inter-vehicle communication is available (Werner et al. 2024).
- Language models as the coordinating agent: Model-generated subtasks assigned by one optimizer, which takes the explored map and the condition of every robot. The optimizer revises the assignment upon new features, emergent tasks or robot failures. The framework runs centrally, so that global information is available to all robots, while inter-robot communication constraints are left to future work (Zhu et al. 2025).

12.4 Decentralized swarms

At disaster sites, the terrestrial communication system can be destroyed (Khan et al. 2022). Decentralized operation suits such sites, since decentralized algorithms have no single point of failure, while they handle erratic communication better (Recchiuto and Sgorbissa 2018). The robots coordinate without radio, over sparse connectivity, or through consensus-based allocation without an auctioneer.

Regarding how much information each robot is assumed to receive from its peers, the main categories are:

- Coordination without radio communication: Neighbors localized, and information propagated, through a local ultraviolet visual channel rather than by radio. Hence, the team's behavior depends on no radio link, as shown outdoors with multiple UAVs. In contrast, an earlier outdoor flight still relied on shared GNSS positions (Horyna et al. 2023).
- Coordination over sparse connectivity: Only sporadic connectivity required, the searchers leaving each other's communication range to cover more ground. Meanwhile, their shared search state spreads by best-effort gossip, which tolerates disconnections and delays (Dah-Achinanon et al. 2023).
- Consensus-based allocation without an auctioneer: Every robot broadcasting its bids and applying the same consensus rules, without a designated auctioneer. When a bid changes, only the auction part of a task is sent, while the full task set is sent every 30 seconds. This scheme was fielded on multiple ground robots in the SubT Final Event (O'Brien et al. 2023).

12.5 Communication-constrained operation

Work on communication-constrained operation treats the link between robots, operators and the base station as the object of study, rather than an assumption (Bravo-Arrabal et al. 2023;

Manzini et al. 2023c). It measures what a disaster site supplies and what its data products demand. Moreover, it costs connectivity inside the plan, or builds autonomy that continues once contact is lost.

Regarding whether the work measures the link, plans around it, or dispenses with it, the main categories are:

- Measurement of what the site supplies: Field reports that commercial 5G is not operational inside a tunnel. In contrast, a Wi-Fi mesh extended by two ground robots carries the video of two operators, but not their joint throughput once both raise the resolution (Bravo-Arrabal et al. 2023). Passing observations also report that a ground robot's Wi-Fi inside a fire-damaged building had sufficient quality for most of the time (Surmann et al. 2024). Moreover, each added mesh node roughly halved the available bandwidth underground (Tranzatto et al. 2024).
- Measurement of what the data products demand: The volume of imagery, video and mapping data generated at a real event, compared with the upload rates observed in the field. At Hurricane Ian, data were frequently, and at least once per day, collected at a rate exceeding even the best observed upload speed (Manzini et al. 2023c).
- Connectivity treated as a planned resource: Communication costed inside the plan. An example is cellular-enabled collaborative operation over an assumed mobile 5G infrastructure covering the area. In this case, reception energy and a transmission energy depending on the distance to the base station enter the robots' energy budget explicitly (Romero et al. 2024).
- Autonomy that degrades gracefully: Architectures in which the loss of contact does not terminate the mission. The robots continue autonomously beyond communication range, while exploiting links opportunistically, as in systems of the DARPA SubT Challenge (Biggie et al. 2023; Petrlík et al. 2025; Morrell et al. 2024). In this challenge, the single-supervisor rule ensured that the approaches could operate with degraded and unreliable links (Chung et al. 2023).

12.6 Onboard computation

Onboard computation keeps perception and autonomy on the fielded robot itself, within its size, weight, power and latency budgets (Petrlík et al. 2025; Pal et al. 2025). It relies on latency-bound perception closing a control loop and on models compressed for continuous monitoring. Moreover, it comprises complete autonomy stacks on the platform's processor, as well as pre-trained models moved onto the robot.

Regarding the budget that binds the computation and what is given up to meet it, the main categories are:

- Latency-bound perception inside a control loop: Perception that must complete fast enough for the platform to react. An example is smoke segmentation at 30 Hz on an onboard Jetson Orin Nano, feeding the trajectory controller (Pal et al. 2025). In another case, a CPU-bound inference script caps receding-horizon planning and target tracking at 15 Hz (Yadav et al. 2023). Moreover, a drone detector designed for FPGA implementation completes in under 20 ms on a prototyping board, trading refined accuracy for low latency and limited energy (Amesseghер et al. 2025).

- Compressed models for continuous monitoring: Computation deliberately reduced for the limited computing resources of UAVs, by discarding redundant video clips and frames. In offline tests, this lowered cost more than tenfold, while raising accuracy (Zhao et al. 2024a).
- Complete autonomy stacks running on the platform: All algorithms running on the platform's own processor. For instance, the six-core CPU of SubT UAVs also hosts the detector on its integrated graphics at 5 Hz per camera, without a dedicated GPU, while their SLAM omits global optimization for real-time performance (Petrlík et al. 2025). Moreover, real-time estimation runs on a CPU equivalent to the robot's (Tuna et al. 2024). In addition, a whole autonomy stack runs on the Intel Core i7 CPU of wheeled and legged robots (Dixit et al. 2025). Finally, radiation field estimation and informative path planning run onboard a micro aerial robot (Mascarich et al. 2023).
- Pre-trained models moved onto the robot: Pre-trained models executed onboard only as perception components, with budgets of their own. For instance, open-set detection and segmentation occupy 12 GB of GPU memory, with a latency of 3-5 s, on an aerial robot with a 100 W power budget (Cladera et al. 2025). In another case, a vision-language model runs on a ground robot's onboard GPU, while the planning language model is queried remotely (Ravichandran et al. 2025).

12.7 Offboard and cloud-dependent inference

Offboard and cloud-dependent inference runs computation away from the robot, on workstations or cloud servers. A lost link degrades autonomy only when remote models are queried inside a robot's decision loop (Yamanaka et al. 2025; Jarabo-Peñas et al. 2025). Otherwise, the external computation serves offline evaluation of large models, interactive Earth-observation assistants and offboard analysis of disaster imagery.

Regarding what the external computation is used for and how tightly the robot depends on it, the main categories are:

- Offline evaluation of large models on disaster data: Benchmarks and evaluations conducted wholly away from any platform. They establish what current models can and cannot do on disaster imagery, without asserting that they could be run in the field (Wang et al. 2025; Danish et al. 2025).
- Interactive assistants over Earth-observation data: Conversational systems answering questions about multi-sensor or multi-temporal imagery, with no robot in the loop. Their attention to computation extends to a compact 4B-parameter design (Soni et al. 2025), or to 8-bit quantization of the base language model. With low-rank adaptation, this quantization fits training onto a single 16 GB GPU (Irvin et al. 2025).
- Offboard analysis of disaster imagery: Damage and hazard mapping fed by satellites (Bountos et al. 2024), survey flights (Manzini et al. 2026; Bal et al. 2023) or social-media photographs (Huang et al. 2026), and computed offboard. At most, it runs on a desktop GPU, which processed a 3 mi^2 orthomosaic in 5 minutes and is judged fieldable (Manzini et al. 2026). Hence, a lost link delays analysis rather than degrading autonomy.
- External models inside a robot's decision loop: Remote models queried during the mission. An example is a cloud-hosted multi-modal model reached over Starlink and

Table 11 Coordination, communication and computational deployment: Comparative analysis and key insights.

Configuration	Single-robot operation	Heterogeneous teams	Centralized coordination & allocation	Decentralized swarms	Communication-constrained operation	Onboard computation	Offboard & cloud-dependent inference
Primary function	• Deliver one platform to one hazard • Survey a bounded area alone	• Combine complementary mobility and sensing • Cover what no single morphology reaches	• Assign robots to tasks from one decision point • Optimize a team-level objective	• Distribute the decision across the team • Survive the loss of any single peer	• Establish what the site's link actually provides • Keep the mission valid without contact	• Run the autonomy on the platform itself • Keep behavior independent of the link	• Apply large models to disaster data • Move computation off the platform
Number and type of platforms	• One, of any morphology	• Few, of two or more morphologies	• Few, often homogeneous	• Three to fourteen in field tests	• Any; the link, not the fleet, is the subject	• Any; a property of the individual platform	• Often none; an analysis pipeline
Locus of coordination	• The operator or the robot; no inter-robot coordination	• One or more operators over shared maps or video	• A single global decision point	• Each robot, from local and peer data	• Shifts to the robot as contact is lost	• On the platform, by construction	• Off the platform, at a workstation or server
Link dependence	• High if teleoperated, low if autonomous	• Intermittent in SubT stacks, continuous for teleoperation	• High for allocation; varies for execution	• Low; sparse or local contact suffices	• The dependence is the object of study	• Low; link for oversight and map sharing	• None offline; high during missions
Where inference executes	• Onboard, or on an external computer	• Onboard per platform, shared over the network	• Allocation at one node, sensing onboard	• Onboard, independently per robot	• Onboard, or offboard after data transfer	• Onboard, under power and latency budgets	• On external or cloud compute
Key strengths	• Already used at real incidents • Manageable by a single operator • A tether can carry power and data	• Access and viewpoint no single platform offers • Fielded at SubT and real incidents	• Globally efficient assignments • Explicit treatment of budgets	• Robust to link and platform loss • Designed for scalability	• Supplies measured rather than assumed conditions • Bounds what others may assume	• Tolerates degraded communications • No link delay in the loop	• Access to commercial models such as GPT-4o • No platform constraint on model size

Table 11 (continued)

Configuration	Single-robot operation	Heterogeneous teams	Centralized coordination & allocation	Decentralized swarms	Communication-constrained operation	Onboard computation	Offboard & cloud-dependent inference
Critical limitations	• Coverage bounded by one platform's endurance • A single point of failure	• Sensor and stack integration burden • Supervisor load can limit the team	• Single point of failure • Assumes connectivity a site may lack	• Harder to design • Field-tested with small teams	• Measurements from single sites or events • Planners may assume full 5G coverage	• Model size bounded by the platform • Some timings measured off the robot	• In-mission use needs a link a site may lack • Cloud products wait on data upload
Typical validation maturity	• Real incidents, exercises, field trials	• Competitions, field trials, incident reports	• Simulation, laboratory, field trials	• Simulation, field tests, SubT Final Event	• Field measurements, competitions, simulation	• Offline tests, field trials, competitions	• Benchmarks, field trials, rarely incidents
Indicative works	• Firefighter support UGV (Fernández Talavera et al. 2023), Fukushima remote system (Sakaue et al. 2025), Prometheus (Brown et al. 2026)	• Team CSIRO Data61 (Kottege et al. 2025), CERBERUS (Tranzatto et al. 2024), air-ground missions (Cladera et al. 2025)	• REACT (Maresca et al. 2025), mine task auction (Dahlquist et al. 2025), Compton-camera MAVs (Werner et al. 2024)	• Radio-free swarms (Horyna et al. 2023), sparsely connected swarms (Dah-Achinanon et al. 2023), consensus task allocation (O'Brien et al. 2023)	• Mesh vs. 5G underground (Bravo-Arrabal et al. 2023), bandwidth at Hurricane Ian (Manzini et al. 2023c), cellular-enabled teams (Romero et al. 2024)	• Low-latency onboard detection (Amesseghe r et al. 2025), radiation mapping (Mascarich et al. 2023), SubT aerial autonomy (Petrlík et al. 2025)	• DisasterM3 (Wang et al. 2025), GEOBench-VLM (Danish et al. 2025), EarthDial (Soni et al. 2025)

an access point at the tunnel entrance (Yamanaka et al. 2025). Other examples are cloud-hosted language-model agents, while the ground vehicle executes its paths locally (Jarabo-Peñas et al. 2025). Hence, a lost link degrades autonomy, unless a local model is substituted. In one copilot, such a local model performed similarly, but took about a minute per initial query instead of one to three seconds (Bark et al. 2025).

12.8 Comparative analysis and key insights

Table 11 compares the seven configurations. The main insights are that: a) The configurations can be ordered by the communication infrastructure they assume, with field maturity broadly decreasing along that order. b) The evidence is poorly matched to the conditions it describes. Decentralized operation, which best matches a disconnected site, is fielded only with small teams. In contrast, centralized allocation presupposes the connectivity that is measured to be limited. c) Dedicated communication measurements remain few. Moreover, they have not yet reached the coordination designs presupposing 5G coverage or inter-vehicle communication. d) The onboard-offboard split partly follows the split between conventional and foundation-model components. Pre-trained models run onboard only as components, while planning language models are mostly queried remotely. Local hosting is envisaged for a 70B-parameter model (Jarabo-Peñas et al. 2025). In another case, it lengthens an initial query to about a minute on an RTX 4070 GPU (Bark et al. 2025). e) Robots sharing an incident site have each been run by their own operators under a human command structure (Surmann et al. 2024; Lichtenfeld et al. 2025). This was already the case after the Tōhoku tsunami, when remotely operated vehicles not designed to be coordinated left their operators to coordinate among themselves (Murphy et al. 2012).

13 Disaster domains and application settings

The disaster domain concerns the hazard environment for which a method is built. This environment determines the missions a system performs, as well as the terrain and operating conditions it faces (Murphy 2014; Murphy et al. 2016). The main domains, as defined in Section 4, are: a) Earthquake and structural collapse. b) Nuclear, radiological and CBRN incidents. c) Fire and wildfire. d) Flood and hurricane response. e) Subterranean and mine incidents. f) Maritime and underwater emergencies. g) Landslide and geohazard monitoring. h) Industrial and hazmat emergencies. i) Humanitarian demining and explosive hazards.

13.1 Earthquake and structural collapse

Work on earthquake and structural collapse spans a response in which survivors not extricated within the first two days are unlikely to survive (Murphy et al. 2016). It ranges from assessing which structures have failed and reconnoitering collapsed interiors to detecting and accessing voids. Moreover, it covers handling casualties and organizing the search at mission level.

According to the stage of the collapse response that is addressed, the main categories are:

- Overhead and street-level damage assessment: Which structures have failed and how severely. This is assessed from pre- and post-event satellite imagery, including that of the 2023 Türkiye earthquake, whose geography and disaster type differ from those of the

source data (Ahn et al. 2025). Alternatively, it is assessed from ground-level social media photographs of major earthquakes. These avoid overhead imagery, which is expensive and often obstructed by clouds, fog or rain in the critical early stages of a disaster (Huang et al. 2026).

- Interior reconnaissance of collapsed structures: Damaged buildings entered under the perceptual and mobility degradation that debris, fire, smoke and unstable surfaces impose (Cruz Ulloa et al. 2025). The category also covers cluttered, multi-level interiors without satellite positioning, as in the search-and-rescue activities required after an earthquake (Hickling et al. 2026).
- Void detection and confined access: The spaces that determine survival, identified from external perception or reached with platforms shaped for the purpose (Bal et al. 2023; McFarland et al. 2024).
- Physical intervention and casualty handling: Bringing an unconscious casualty found in a pose unsuitable for extraction into one that permits it. This need was made pressing by the scale of medical evacuation after the 2023 earthquake in Türkiye (Peiros et al. 2023).
- Mission-level assistance for rescue teams: The search of a damaged building organized at mission level. In one case, robots are allocated across a fire- or earthquake-damaged factory whose debris and machinery hide trapped victims (Maresca et al. 2025). In another case, robots open doors, such as one deformed by an earthquake, and clear obstacles. Without this ability, robots reached only limited areas of the collapsed buildings in the 2023 Türkiye response (Shi et al. 2025).

13.2 Nuclear, radiological and CBRN incidents

Work on nuclear, radiological and CBRN incidents addresses a hazard that human senses cannot perceive (Werner et al. 2024). It maps and localizes radiation fields, and plans coverage that respects dose limits. Moreover, it hardens robots for long decommissioning campaigns under cumulative dose, while it evaluates operator interfaces with trained robot operators.

According to whether the work maps the invisible hazard, plans against it, hardens against it, or mediates it to an operator, the main categories are:

- Mapping and localizing radiation fields: Where the activity is, or whether it is present. This question arises across confined GNSS-denied facilities, of which decommissioned nuclear plants are the indicative case (Mascarich et al. 2023). It also arises for weak sources in transit across unknown and sometimes GNSS-denied environments, a non-proliferation concern (Yadav et al. 2023). Moreover, it arises in open terrain holding several sources, a need that emerged after the Fukushima Daiichi disaster (Werner et al. 2024).
- Planning coverage against an unseen hazard: Where a helicopter or UAV should fly to cover a contaminated area while staying within exposure limits. These limits are set as a legally enforced cap on the momentary local dose rate, or as a threshold on the cumulative mission dose. The reason is that excessive dose rates can contaminate equipment and endanger personnel (Blake and Schubert 2025).
- Radiation-tolerant hardware for sustained operation: Platform lifespan extended under cumulative dose, so that a long decommissioning campaign can be completed with fewer replacement robots (Sakaue et al. 2025).

- Operator interfaces evaluated with trained operators: Decommissioning tasks for a legacy site, whose building inventories can be unknown before entry and whose manual monitoring exposes staff to dose (Stedman et al. 2023). In another case, a digitized nuclear environment with radiation hot spots is evaluated by robot operators from two nuclear institutions (Baniqued et al. 2024).

13.3 Fire and wildfire

Fire and wildfire involve a hazard that is an active process modifying the environment while the robot operates. Moreover, heat, dense smoke and thermal distortion degrade onboard sensing (Fernández Talavera et al. 2023; Tiwari et al. 2025). The works engage the fire, track its spread, survey it from the air or benchmark perception and decisions within it.

According to whether the hazard is engaged or inspected, tracked, surveyed or benchmarked, the main categories are:

- Intervention and inspection at the fire scene: Suppression whose jet trajectory depends on wind speed and hydrant water pressure (Chaoxia et al. 2025). Moreover, ground support is provided in indoor fires without GNSS, where the fire may block planned routes and dense smoke can affect onboard sensors (Fernández Talavera et al. 2023). Post-fire inspections are also carried out under entry bans. In these inspections, a drone entered through roof-hatch openings narrowed by fire-bent metal, while a tracked robot met narrow passages, loose rubble and wire meshes from burnt-out couches (Surmann et al. 2024).
- Tracking a propagating hazard: An evolving smoke plume followed rather than suppressed, so that the platform remains within it as it moves (Pal et al. 2025). The category also includes resources for predicting at a daily resolution how active wildfires will spread, from satellite active-fire detections together with fuel, topography and weather variables (Gerard et al. 2023).
- Aerial surveillance at landscape scale: Wildfire areas surveyed from the air. This is done by multi-robot support architectures in mountainous terrain (Frering et al. 2023), as well as by UAV video classification of forest fires. In the latter case, the aerial footage suffers blurred frames and camera vibration from air currents (Zhao et al. 2024a).
- Benchmarks for perception and decision making in fire: Perception and episodic memory in smoke-obscured firefighting video (Tiwari et al. 2025), as well as embodied decision making in simulated, dynamically changing fire environments (Zhou et al. 2024). Moreover, aerial thermal sensing is benchmarked amid prescribed wildland fires, where dense trees and the overbearing thermal signal of the fire hamper crew and vehicle detection (Jong et al. 2023). In addition, forest and urban settings degraded by smoke, rain and darkness are covered (Dhrafani et al. 2025).

13.4 Flood and hurricane response

Work on flood and hurricane response covers large-area events observed from different altitudes (Bountos et al. 2024; Manzini et al. 2023b). It ranges from satellite-scale inundation mapping to sUAS surveys flown during the response. Moreover, it covers the wireless,

procedural and fatigue constraints of real and replicated flight operations, as well as the interpretation of post-event imagery by large pre-trained models.

According to the altitude from which the event is observed and what that altitude can resolve, the main categories are:

- Satellite-scale inundation mapping: Water extent delineated over large areas from spaceborne or airborne imagery, in which tree canopies and clouds make flood pixels ambiguous (Sami et al. 2024). A global multi-temporal radar resource makes the problem learnable at that scale (Bountos et al. 2024). Moreover, a benchmark targets the direct forecasting of flood extent, linking the rarely connected weather prediction and spaceborne flood mapping (Victor et al. 2024).
- sUAS survey at the event: Flights over the affected area during the response itself (Manzini et al. 2023b), and their annotated products. Unlike the low-resolution imagery of some earlier public road damage datasets, these products are fine enough for the phenomena of interest to emergency managers (Manzini et al. 2026). In addition, flights over a European flash flood recorded the extent of flooding and the water flow direction (Surmann et al. 2024).
- Constraints revealed by real and replicated flight operations: What actually limited those operations. Wireless capacity in the hurricane-struck area, whether FirstNet, satellite or consumer cellular, was too limited for the data, which were carried to the nearest intact wired network (Manzini et al. 2023c). Moreover, procedural errors were possibly due to squads adopting new protocols when agencies combine under incident command (Murphy et al. 2024). Furthermore, crew fatigue increased over an overnight exercise replicating recent response tempos (Mehta et al. 2024).
- Model-based interpretation of the resulting imagery: Large pre-trained models applied to post-event overhead imagery. They adapt building-damage mapping to unseen regions, such as hurricane-affected ones (Ahn et al. 2025). Moreover, vision-language models are benchmarked on events including several hurricanes, whose extreme weather hinders optical imaging and calls for radar views of the post-event scene (Wang et al. 2025).

13.5 Subterranean and mine incidents

Subterranean and mine incidents deny satellite positioning, light and reliable communication (Ebadi et al. 2024; Chen et al. 2026). The works compensate for these through degradation-robust localization and mapping, field-scale heterogeneous team autonomy, as well as risk-aware and semantically informed traversal. Moreover, they comprise aerial inspection of mines, together with communication and interfaces specific to the underground.

According to what the underground environment denies and how the system compensates for it, the main categories are:

- Localization and mapping under perceptual degradation: Estimates kept valid where the underground site itself degrades sensing. An example is a mine whose tunnels can leave LiDAR registration degenerate (Tuna et al. 2024). Other examples are enclosed, narrow tunnels laden with dust and water vapor under poor illumination (Zeng et al. 2024), as well as narrow flooded galleries of caves and mines, where conventional surveying and mapping procedures cannot be applied (Villarroel et al. 2024).

- Field-scale heterogeneous team autonomy: Competition-derived heterogeneous fleets fielded across tunnels, urban underground structures and natural caves, amid smoke, narrow passages, stairs, mud and degraded wireless links (Kottege et al. 2025; Tranzatto et al. 2024; Morrell et al. 2024). Their aerial component, in turn, faces the dust of mines and the fog of humid caves (Petrlík et al. 2025).
- Risk-aware and semantically informed traversal: Terrain treated as more than a binary obstacle map. Approaches bound the risk of rubble, sudden drops and mud in caves and mines (Dixit et al. 2025), or weigh the stairs and rubble of the urban section of the competition course (Ginting et al. 2023). Moreover, traversable ground is labeled in unstructured, GNSS-denied mines and caves (Biggie et al. 2023).
- Aerial inspection and monitoring of mines: Mines in routine operation or under stability monitoring, rather than after an accident. In these mines, operator-selected inspection tasks are executed, gas is monitored after blasting and voids accessible only through a borehole are reached (Brown et al. 2026; Dahlquist et al. 2025; Nordström et al. 2025).
- Infrastructure and interfaces specific to the underground: The communication achievable in a tunnel (Bravo-Arrabal et al. 2023), as well as the severe signal attenuation that reinforced concrete walls and corners cause in an underground building (Wang et al. 2023). The category also covers the inspection of abandoned mine tunnels, which operators must still maintain after closure. These comprise muddy and concrete-paved tunnels with high humidity and dripping water (Yamanaka et al. 2025).

13.6 Maritime and underwater emergencies

Work on maritime and underwater emergencies spans drift-informed search for a target carried away from its last known point by currents and wind (Zhou et al. 2026). Moreover, it covers the robotic survey of shallow urban waters, where poor visibility and concealed hazards make diver searches slow and dangerous (Rymansaib et al. 2023).

According to whether the platform searches a volume of open water or surveys a delimited submerged area, the main categories are:

- Drift-informed search of open water: Ocean currents and wind carrying a wrecked target away from its last known point, so that the probable target region shifts over time. The approach was validated around an assumed accident location in a reservoir, with water depths of 1 to 10 m and flow speeds of 0.1-0.2 m/s (Zhou et al. 2026).
- Investigation of a fixed submerged site: Delimited shallow urban waters, surveyed so that divers can be directed to likely targets, as in a harbor about 5 m deep (Rymansaib et al. 2023).

13.7 Landslide and geohazard monitoring

Landslides and other geohazards appear only rarely. They appear either as a secondary setting of aerial and satellite work that observes the site from a vantage avoiding the unstable ground, or in an early deployment of ground robots into structures damaged by a mudslide (Pal et al. 2025; Murphy and Stover 2008).

According to whether the unstable site is surveyed from a vantage that does not touch it or is traversed directly, the main categories are:

- Survey from a vantage that avoids the substrate: Aerial and satellite work in which geohazards appear only as a secondary setting. An example is in-plume drone tracking of particulate dispersion, motivated in part by volcanic ash (Pal et al. 2025). Another one is an exercise replicating the tempo of drone responses to hurricanes and a volcanic eruption (Mehta et al. 2024). Moreover, in a satellite multi-hazard benchmark, all models perform better on the simpler rural mountainous landslide scenes than on earthquake, tornado and explosion scenarios. In this benchmark, volcanic scenes are queried for the lava and for intact buildings within 100 m of it (Wang et al. 2025).
- Traversal of mudslide debris: Ground robots sent into structures collaterally damaged by a mudslide, which rarely leaves human-sized voids. In this setting, a small robot was considered imperative, while waterproofing was essential in the highly wet conditions. Moreover, the risk of a further slide, which ended the response, made slide-stability monitoring a further robot task. This monitoring was done manually by geologists and soil engineers (Murphy and Stover 2008).

13.8 Industrial and hazmat emergencies

Industrial and hazmat emergencies involve hazardous substances at plants and other industrial sites, which responders can approach only at considerable risk (Steinbauer-Wagner et al. 2025; Lichtenfeld et al. 2025). The works detect and identify the substance, localize the source of an airborne release, remove it at the source, or intervene at a real plant emergency.

According to whether the substance is detected, traced to its source, removed, or the site entered outright, the main categories are:

- Detection and identification of hazardous substances: Chemical sensing carried into environments that responders can enter only at considerable risk. It is motivated by an unintended gas release from a deep drilling operation, which escalated into an explosion in a nearby residential building. It is also motivated by an accidental nitric-acid release at an industrial site (Steinbauer-Wagner et al. 2025).
- Source localization of an airborne release: The origin of a dispersing plume sought among buildings, which shape the plume structure and make sampling in their wake less informative (Rhodes et al. 2023). Moreover, the origin is sought in warehouses of increasing complexity, used to compare such algorithms. This work is motivated by gas-leak responses, in which humans currently localize the source, and by the monitoring of large chemical plants (Erwich et al. 2024).
- Removal and containment at the source: Hazardous gas extracted from collapsed sites, long pipelines or underground mines, where effective removal requires navigating complex pathways to the contaminated area (Lee et al. 2025).
- Intervention at a real industrial emergency: Robotic intervention during an actual plant emergency. In this emergency, a shock-sensitive gas in a fire-damaged reactor imposed a safety radius of about 200 m. Moreover, a narrow staircase and an upper floor at risk of collapse further constrained the mission (Lichtenfeld et al. 2025).

13.9 Humanitarian demining and explosive hazards

Work on humanitarian demining and explosive hazards uses robots to mitigate the risks of landmines and of ordnance disposal, among the most hazardous occupations (Malizia et al. 2026; Xiao et al. 2023). The robots either detect landmines at a standoff distance from a ground vehicle, or handle ordnance at contact through teleoperated disposal.

According to whether the hazard is detected at a standoff distance or handled at contact, the main categories are:

- Multi-sensor standoff detection of landmines: Detection from a ground vehicle in off-road terrain. The data come from a single overcast late-autumn campaign, over tracks laid with inert antipersonnel and antitank mines, as well as with common items that might be found in natural environments. Hence, models trained on them may face domain gaps in other seasons, terrains or weather (Malizia et al. 2026).
- Teleoperated disposal at contact: EOD, in which heavy protective equipment constrains the technicians' movement and dexterity. Moreover, robots relying heavily on visual telemetry can miss concealed or buried improvised devices (Xiao et al. 2023). The category also includes heavy dual-arm platforms built for battlefields and disaster environments, for casualty extraction and the acquisition of explosive ordnance (Hong et al. 2023).

14 Validation maturity

Validation maturity concerns the setting in which a method has been shown to work. It is a criterion in its own right, since gaps between theory and practice are identified for disaster robots (Murphy 2014). Moreover, computer vision and machine learning have proved difficult to migrate to real disasters (Manzini et al. 2023a). The main validation levels, as defined in Section 4, are: a) Simulation and benchmark evaluation. b) Laboratory conditions. c) Standardized test arenas and competitions. d) Controlled field trials. e) Exercises with responders. f) Deployment at a real incident.

14.1 Simulation and benchmark evaluation

Simulation and benchmark evaluation measures methods without a physical disaster site (Huang et al. 2026; Zhou et al. 2024). The disaster is replaced by annotated imagery, benchmarks probing large pre-trained models or embodied simulators of changing hazards. Moreover, it is replaced by generated data, simulation-only planning or interaction designs assessed without a physical trial. Offline evaluation on data recorded in field trials, at competitions or at an incident is also coded at this level.

Regarding what is made to stand in for the disaster itself, the main categories are:

- Annotated disaster imagery and sensing benchmarks: Labeled data of disaster scenes and hazards against which a perception method is scored. The modalities range from ground photographs and sUAS imagery to SAR, LiDAR and infrared (Huang et al. 2026; Bountos et al. 2024; Manzini et al. 2026; Malizia et al. 2026).

- Benchmarks probing large pre-trained models: Resources establishing what general-purpose models do and fail to do on disaster and geospatial content, some also supplying training splits (Wang et al. 2025; Danish et al. 2025; Tiwari et al. 2025).
- Embodied simulators of changing hazards: Simulated scenes in which the hazard evolves or constrains an agent, or a human-robot team, while it acts. Hence, decisions and behavior rather than recognition are measured (Zhou et al. 2024; Orlov Savko et al. 2024). Moreover, gas dispersion is simulated over computational fluid dynamics (CFD) wind fields in warehouse settings, in order to compare gas-source localization algorithms (Erwich et al. 2024).
- Generated substitutes for missing data: Labeled change data synthesized in place of an expensive, knowledge-intensive annotation campaign. This shifts the training dependence to a generative model, while evaluation remains on real annotated benchmarks (Zheng et al. 2025).
- Algorithmic evaluation in simulation alone: Planning results demonstrated entirely in simulation. Examples are Monte Carlo source-search runs over an experimentally validated CFD release field, as well as coverage flight paths over terrain and radiation scenarios. Other examples are Webots scenarios with a tracked robot and a factory accident, whose returned plans are labeled optimal, valid or invalid (Rhodes et al. 2023; Blake and Schubert 2025; Bark et al. 2023, 2025).
- Interaction designs assessed without a physical trial: Interfaces evaluated with a virtual robot fleet, or with designs shown in images and video, by nuclear robot operators or first responders (Baniqued et al. 2024; Lee et al. 2023). In another case, participants are guided by a simulated evacuation robot through a smoke-filled virtual building (Hu et al. 2023).

14.2 Laboratory conditions

Laboratory evaluation establishes on physical hardware, in a scene arranged in advance by the researchers (Lee et al. 2025; Cao et al. 2024), that a mechanism works or that a simulation-trained policy transfers to a real robot. Moreover, it establishes how operators perform with alternative interfaces, what onboard processors can execute, or whether hazard sensing works against placed sources.

Regarding what the laboratory is able to reproduce, the main categories are:

- Mechanism and mobility validation: A novel mechanism or physical interaction shown to perform its intended function on real hardware. Examples are pressure and bending measurements of a soft growing robot, followed by its growth through a vertical passage and pipe elbows and gas suction at the end. Another example is a robot arm repositioning the limb of a live human subject (Lee et al. 2025; Peiros et al. 2023).
- Transfer of learned policies to real hardware: Simulation-trained policies run on physical platforms in a controlled setting. In one case, multiple TurtleBots search for walking people, who carry colorful boards as targets. In another case, a firefighting robot at a campus site aims its water cannon at dummy flame targets, used instead of real flames for safety (Cao et al. 2024; Chaoxia et al. 2025).

- Controlled studies of operators and interfaces: The human rather than the robot as the subject, comparing interface designs on a fixed course or task set. For instance, screen-based and virtual reality control are compared by multiple participants with robot control experience, over the same Wi-Fi link as in field use. Moreover, trained operators command a robot in a separate building over a VPN (Illing et al. 2025; Stedman et al. 2023). Another study measures the number of discrete vibrotactile signals that trained participants could reliably identify on a sleeve worn on each arm, under different attentional conditions (Xiao et al. 2023).
- Bench characterization of onboard computation: Prototype measurements of what a platform's own processors can execute within its latency and power budgets. These are complemented by indoor-arena flights of an event camera over toy cars, rather than over a disaster scene (Amessegher et al. 2025).
- Hazard sensing against placed sources: Hazards introduced by the researchers indoors. An example is autonomous aerial mapping of radiation fields from uranium ore and radium sources placed in a machine shop and a laboratory (Mascarich et al. 2023).

14.3 Standardized test arenas and competitions

Standardized test arenas and competitions score robotic systems against the same placed targets and rules (Chung et al. 2023; Kottege et al. 2025). They involve whole-system entries in the DARPA SubT Challenge or a drone competition, as well as components measured under course conditions. Moreover, archival accounts consolidate the results after the event.

Regarding what the standardization is used to establish, the main categories are:

- Whole-system competition entries: Integrated multi-robot systems built for and scored in a common course. They report results from the SubT Final Event, or from the multi-level disaster mock-up of the 2024 NATO-Sapience Autonomous Cooperative Drone Competition. Several of them were also deployed in mines and caves outside the competition (Tranzatto et al. 2024; Kottege et al. 2025; Petrlík et al. 2025; Morrell et al. 2024; Hickling et al. 2026).
- Components evaluated against competition conditions: Task allocation, autonomy, traversability and navigation measured in competition runs, at a competition site or in an arena reconstructed to NIST standards. Hence, a component result is obtained under course conditions rather than in isolation. Several of these works also add field trials outside the competition (O'Brien et al. 2023; Biggie et al. 2023; Dixit et al. 2025; Cruz Ulloa et al. 2025).
- Archival consolidation after the event: Retrospective accounts collating results across competitors and events with measures beyond the scoring criteria. Alternatively, they tabulate the open-source SLAM implementations and datasets that each team released (Chung et al. 2023; Ebadi et al. 2024).

14.4 Controlled field trials

Controlled field trials expose a system to the environmental variability of a real site, without an emergency in progress (Nordström et al. 2025; Schomer and Adams 2026). They comprise platform trials in representative terrain, as well as sensing and estimation in genuinely

degraded conditions. Moreover, they include search and planning over real outdoor sites, and semantic and language-driven systems taken outdoors.

Regarding which aspect of the uncontrolled environment the trial is designed to expose, the main categories are:

- Platform trials in representative terrain: A platform, or the autonomy running on it, shown to operate in the physical conditions of its intended domain. Trials took place in a test mine and in operating ones, either in an active production area or in drifts flown about 40 minutes after a large-scale blast. Moreover, they took place in natural caves and mining galleries, as well as on a mock collapsed-structure rubble pile at an urban search-and-rescue task-force training site (Brown et al. 2026; Dahlquist et al. 2025; Nordström et al. 2025; Villarroel et al. 2024; McFarland et al. 2024).
- Sensing and estimation in genuinely degraded conditions: Environments degrading conventional sensing, through low visibility in tunnels, or geometrically uninformative scenes such as a mine and a construction site. In other cases, radiation that only a specialized instrument perceives comes from sources placed by the researchers, carried by a moving ground robot or spread over an open field (Zeng et al. 2024; Tuna et al. 2024; Yadav et al. 2023; Werner et al. 2024).
- Search and planning over real outdoor sites: Search, coverage and tracking with targets spread over open terrain, water or air. Examples are multiple days of flights over a snow-covered butte where mountain rescuers train, as well as generator-produced smoke plumes over an open agricultural field. Moreover, AUV trials took place in a reservoir, while autonomous searches looked for a clothed mannequin on the muddy floor of a harbor. Finally, multiple drones searched a football field for a single fixed target (Schomer and Adams 2026; Pal et al. 2025; Zhou et al. 2026; Rymansaib et al. 2023; Dah-Achinanon et al. 2023).
- Semantic and language-driven systems taken outdoors: Semantically grounded and language-conditioned systems in the field. They range from a legged robot planning over terrain semantics in a two-level outdoor area (Ginting et al. 2023) to planners querying large language or vision-language models. The latter were tested in an outdoor office park, at rural and urban sites, and in real tunnels (Ravichandran et al. 2025; Cladera et al. 2025; Yamanaka et al. 2025).

14.5 Exercises with responders

Exercises with responders place a system before the operational organizations that would use it (Jarabo-Peñas et al. 2025; Steinbauer-Wagner et al. 2025). They comprise full response exercises under their own command structure, as well as trials of a specific capability inside a responder workflow or at an operational facility. Moreover, they include studies of fatigue, usability and acceptance conducted within exercises.

Regarding what the presence of the response organization tests that the researchers cannot test alone, the main categories are:

- Full response exercises: Responders working within their own command structure and procedures, with researchers alongside. Hence, the robot must fit a response, rather than the response accommodate the robot. For instance, the unified command center of

an exercise with multiple parallel rescue missions prioritized logistical requests for a ground robot (Jarabo-Peñas et al. 2025). In a smaller exercise, a limited wildfire scenario was set up. In it, firefighters handled controlled fires with their standard equipment and procedures, while a commanding officer in their mobile command vehicle tasked the robots (Frering et al. 2023).

- Systems trialed inside a responder workflow: A specific capability placed into the working practice of a crew. This tests whether it is usable by the people it was built for, as well as whether it works. An example is a radiation-source search staged with a responder squad of a robot pilot and a data spotter, in order to try the standard operating procedure written for a legged sniffer robot (Steinbauer-Wagner et al. 2025).
- Trials at operational facilities: Trials with the response organization at an operational facility rather than a research site. Examples are two interventions with a city fire department at its Unified Safety Center, one with actual fire and cold smoke. In these interventions, a ground robot detected fire sources and supplied environmental information to firefighters (Fernández Talavera et al. 2023). Another one is a hazardous-materials exercise at the plant of a cooling-unit manufacturer, a real site without a real emergency. There, the robot team was tasked through the fire-service command, while an independent observer recorded the results (Steinbauer-Wagner et al. 2025).
- Elicitation and acceptance studies: Responders observed, interviewed or surveyed within an exercise, in order to establish how fatigue, errors, usability and acceptance shape robot use. Examples are a 24-hour sUAS exercise flown by emergency-management pilots, and a usability survey of the firefighters in a field test (Mehta et al. 2024; Frering et al. 2023).

14.6 Deployment at a real incident

Deployment at a real incident is the only level at which a system's contribution to an actual response can be observed (Surmann et al. 2024; Manzini et al. 2023b). Its record comprises accumulated task-force practice, quantitative analyses of single events and sustained operation at a persistent hazard. Moreover, it comprises intervention at an industrial emergency and learned models run during a response.

Regarding what the incident record actually reports, the main categories are:

- Accumulated task-force practice: Records of selected deployments by a standing robotics task force, in multiple missions exploring destroyed buildings with ground and aerial robots. These records report lessons learned, rather than the performance of one system (Surmann et al. 2024).
- Quantitative records of a single event: Analyses of the flights, data products and errors of one response. They supply quantitative operational data, such as sortie counts, network demands and human-error counts (Manzini et al. 2023b; Murphy et al. 2024; Manzini et al. 2023c).
- Sustained operation at a persistent hazard: A robot in operational use for the decommissioning of a damaged nuclear power station. Its sandbag-accumulation work is planned over several months, rather than over a response. Moreover, its first month of on-site

Table 12 Validation maturity: Comparative analysis and key insights.

Validation level	Simulation & benchmark evaluation	Laboratory conditions	Standardized test arenas & competitions	Controlled field trials	Exercises with responders	Deployment at a real incident
What the evidence establishes	• That a method scores well against a fixed target • Comparability with prior work	• That the mechanism or policy exists in hardware • That a controlled comparison holds	• That a complete system works under common rules • Ranking against other whole systems	• That the environment does not defeat the system	• That responders can and would use it • That it fits an existing workflow	• That it contributed, or failed, under real conditions
Where the evaluation takes place	• Offline data and simulators	• Prepared settings, mostly indoors	• Reproducible physical courses	• Real unstructured sites, researcher-chosen	• Exercise grounds and operational facilities	• The emergency site itself
Who conducts it	• Researchers alone	• Researchers alone	• Researchers, often under an external organizer	• Researchers, at times with site operators or rescuers	• Researchers jointly with responders	• Responders or site staff, often with researchers
What varies and what is held fixed	• Everything fixed but the method	• Physics real; environment fixed	• Course fixed; system free	• Environment free; task and time fixed	• Users and workflow free; hazard staged	• Nothing is fixed
Cost and repeatability	• Low cost, fully repeatable	• Moderate cost, repeatable	• High cost, repeatable by design	• High cost, weakly repeatable	• High cost, tied to exercise calendars	• Event not chosen, not repeatable
Key strengths	• Scale and statistical power • Strongest support for fair comparison	• Isolates a causal claim • Humans studied under controlled conditions	• Forces whole systems to be built • Produces transferable archival records	• Exposes real environmental degradation • Achievable without an emergency	• Tests acceptance, not only capability • Yields responder feedback and error data	• Direct evidence of operational value • Reveals organizational and human failure
Critical limitations	• Lacks real robot and environment dynamics • Rewards benchmark fit	• Scene arranged in advance • Transfer to the field untested	• Shaped by course rules, not by incidents • Mostly legged and ground platforms	• No time pressure, no unfamiliar site • Site chosen by the researchers	• Very few instances • Robots at times operated by researchers	• No experimental control • No foundation-model-based system

Table 12 (continued)

Validation level	Simulation & benchmark evaluation	Laboratory conditions	Standardized test arenas & competitions	Controlled field trials	Exercises with responders	Deployment at a real incident
Criteria where this level dominates	• Satellite and airborne sensing; damage assessment • Foundation-model-based methods	• Manipulators and intervention systems • Continuum platforms, tied with field trials	• Legged platforms; physical intervention • Nuclear and CBRN domain	• Subterranean domain; marine vehicles • Reconnaissance and mapping; aerial platforms	• None; mostly fire, nuclear and hazmat cases	• Floods and hurricanes
Indicative works	• DisasterM3 (Wang et al. 2025), HAZARD (Zhou et al. 2024), DroneAudioSet (Gupta et al. 2025)	• Gas-suction growing robot (Lee et al. 2025), casualty manipulation (Peiros et al. 2023), virtual reality vs. screen control (Illing et al. 2025)	• NIST-arena quadruped (Cruz Ulloa et al. 2025), team CSIRO Data61 (Kottege et al. 2025), SubT results analysis (Chung et al. 2023)	• Mine multi-agent deployment (Dahlquist et al. 2025), Compton radiation mapping (Werner et al. 2024), SPINE (Ravichandran et al. 2025)	• sUAS fatigue exercise (Mehta et al. 2024), voice-commanded UGV (Jarabo-Peñas et al. 2025), hazmat sniffer robot (Steinbauer-Wagner et al. 2025)	• DRZ task force (Surmann et al. 2024), CRASAR at Hurricane Ian (Manzini et al. 2023b), Fukushima remote system (Sakaue et al. 2025)

operation is reported without control-system errors or radiation-related effects (Sakaue et al. 2025).

- Intervention at an industrial emergency: A robot committed to a real chemical-plant emergency. The record documents the demands of the response and the limitations of research platforms that the mission exposed, such as communication constraints and the need for enhanced operator-assistance functions (Lichtenfeld et al. 2025).
- Learned models run during a response: A road-damage segmentation baseline deployed in Florida in response to two hurricanes, in order to gather the feedback of practitioners. These practitioners distrusted outputs whose road lines did not coincide with the imagery (Manzini et al. 2026).

14.7 Comparative analysis and key insights

Table 12 compares the six validation levels. The main insights are that: a) The levels trade experimental control for realism monotonically. Moreover, they describe the field's division of labor more than a path that individual systems travel to the end, as works pairing simulation or laboratory results with exercise or incident evidence remain isolated. b) The distribution across the levels is bimodal rather than pyramidal. It peaks at simulation and benchmark evaluation, and at controlled field trials, with responder exercises the least populated. c) The level a work attains is strongly associated with its domain and platform. A real incident is the most frequent level for flood and hurricane response, while all satellite and airborne sensing works remain at simulation and benchmark evaluation. Moreover, legged platforms mostly concentrate on arenas, competitions and controlled field trials. d) The foundation-model-based works rest on weaker evidence. Most of them lie at the two lowest levels, while none reaches the incident level. e) The operational evidence on the human-side bottlenecks lies in few works. Human-error and fatigue data come from a real incident and a responder exercise (Murphy et al. 2024; Mehta et al. 2024). In contrast, the accuracy that responders would require before trusting semantic outputs rests on a single questionnaire of first responders (Ruan et al. 2025).

15 Public datasets and benchmarks

Building on the ten criteria of Sections 5-14, this section outlines the main public datasets and benchmarks introduced for developing and evaluating disaster robotics methods. The listed resources date from the last three research phases defined in Section 3. In particular, Tables 13-15 group the identified resources by acquisition source. For each entry, they provide its: a) Year. b) Source/platform. c) Modalities involved. d) Scale. e) Annotation type. f) Targeted task. g) Disaster domain. h) Environment type (Real (R), Simulated (S), Hybrid (H)). i) Short description.

15.1 Key observations

Based on the public disaster robotics datasets and benchmarks summarized in Tables 13-15, the following global observations can be drawn:

- The benchmark drought was specific to damage assessment, and the recent wave is not: After xBD (Gupta et al. 2019) in 2019, no further resource dedicated to building damage

Table 13 Public datasets built from remotely sensed and crowd-sourced disaster imagery. Abbreviations used: Environment type (Real (R), Simulated (S), Hybrid (H)); DEM, digital elevation model; NIR, near-infrared; QA, question answering. Entries marked "preparedness" forecast the occurrence or spread of a hazard instead of describing an event already underway.

Satellite and airborne remote sensing									
Name	**Year**	**Source/ platform**	**Modalities**	**Scale**	**Annotation**	**Task**	**Domain**	**Env.**	**Description**
xBD (Gupta et al. 2019)	2019	Satellite, ∼0.5 m	Optical, 3, 4 or 8 bands, bi-temporal	∼700k building annotations; >5,000 km^2; 15 countries	Building polygons; four-grade joint damage scale; damage-type labels	Building damage assessment	22 events, eight disaster types	R	By its authors' account, the largest satellite building-damage dataset at its release; released with the xView 2.0 challenge and the imagery source of later vision-language corpora (Irvin et al. 2025; Chen et al. 2025)
Global Flood Database (Tellman et al. 2021)	2021	Satellite, 250 m	Optical, daily	913 flood events, 2000-2018; 2.23M km^2 inundated	Threshold-derived pixel-level inundation extent; population-exposure estimates	Flood extent mapping and exposure estimation	Flood	R	An observational archive rather than a learning benchmark; maps the maximum observed extent of each event and estimates exposure trends
Digital Typhoon (Kitamoto et al. 2023)	2023	Geostationary satellite, 5 km	Infrared brightness temperature, hourly	1,099 cyclones; 189,364 images; 1978-2022	Hourly best-track metadata: center, grade, pressure, wind	Intensity analysis, forecasting and reanalysis	Tropical cyclone	R	The longest satellite record of any hazard in the set (1978-2022); inter-satellite recalibrated across nine platforms (preparedness)
Wildfire-SpreadTS (Gerard et al. 2023)	2023	Satellite and gridded weather data, 375 m	Multi-spectral, fuel, topography, weather; 23 channels	13,607 images; 607 fire events; 2018-2021	Binary pixel-level active-fire masks; noisy and highly imbalanced	Next-day wildfire spread prediction	Wildfire, United States	R	Daily time series rather than image pairs; extends the mono-temporal Next Day Wildfire Spread dataset to multi-temporal spread forecasting (preparedness)

Table 13 (continued)

Satellite and airborne remote sensing									
Name	**Year**	**Source/ platform**	**Modalities**	**Scale**	**Annotation**	**Task**	**Domain**	**Env.**	**Description**
Kuro Siwo (Bountos et al. 2024)	2024	Satellite, 10 m	SAR, VV/VH, GRD and SLC, multi-temporal	67,490 annotated time series; 202,470 samples; 1.6M unlabeled	Three-class pixel labels (no water, permanent water, flood) by five SAR experts	Rapid flood mapping	Flood; 43 events, 2015-2022	R	Two pre-event and one post-event acquisition per time series; ships with an accompanying benchmark suite
GFF (Victor et al. 2024)	2024	Satellite and reanalysis, 10 m	SAR, DEM, height above nearest drainage, ERA5	298 flood regions; 164,845 tiles; 2014-2020	Pixel-level flood maps at a given lead time, model-generated except where Kuro Siwo expert labels exist	Flood inundation forecasting	Flood, (near-) coastal	R	The forecasting counterpart to Kuro Siwo's observational mapping; basin-partitioned splits prevent leakage (preparedness)
TEOChatlas (Irvin et al. 2025)	2025	Satellite, eight sensors	Optical; single images and temporal sequences of up to eight images	554,071 instruction examples over 1,244,393 images	Instruction and QA pairs; four-grade building damage from xBD	Temporal Earth-observation instruction tuning	Multi-hazard, inherited from xBD	R	Assembled by converting the labels of existing corpora into instructions; by its authors' account, the first instruction-following Earth-observation corpus with temporal tasks
EarthDial (Soni et al. 2025)	2025	Satellite and aerial, multi-resolution	RGB, SAR, multi-spectral, NIR, infrared, hyper-spectral	>11.11M instruction pairs; evaluated on 44 downstream datasets	Instruction and dialogue pairs generated from existing labels	Multi-task Earth-observation dialogue	Damage (37,563 disaster-assessment pairs), change, earthquake	R	The largest instruction corpus in the set; disaster content is one component of a general Earth-observation resource

Table 13 (continued)

Satellite and airborne remote sensing									
Name	**Year**	**Source/ platform**	**Modalities**	**Scale**	**Annotation**	**Task**	**Domain**	**Env.**	**Description**
GEOBench-VLM (Danish et al. 2025)	2025	Satellite and aerial, multi-source	Optical, multi-spectral, SAR; bi- and multi-temporal	>10,000 manually verified instructions; eight categories, 31 subtasks	Five-option multiple choice, boxes, masks, captions	Benchmarking vision-language models on geospatial tasks	Damage assessment, flood, fire risk, earthquake	R	Evaluation-only; the best of 13 models reaches 41.7% multiple-choice accuracy, slightly above twice the random-guess level
DisasterM3 (Wang et al. 2025)	2025	Satellite, multi-sensor	Optical and SAR, bi-temporal	26,988 bi-temporal images; 123k instruction pairs	Instruction and QA pairs over nine tasks, to long-form reports	Disaster damage assessment and response	36 events, ten categories, five continents	R	The only corpus pairing optical pre-event with SAR post-event imagery, as extreme weather often hinders optical imaging; reports a cross-sensor gap in current models
RSCC (Chen et al. 2025)	2025	Satellite, Maxar Open Data	Optical, bi-temporal	62,351 pre/post pairs; 32.7 billion pixels; 72-word mean caption	Free-text change captions, model-generated with 10% expert verification	Disaster change captioning	31 events; earthquake, flood, tsunami, storm, volcano, wildfire	R	Built by captioning the imagery of xBD and a second building-damage corpus; annotation is hybrid rather than human
MONITRS (Revankar et al. 2025)	2025	Satellite, 10 m	Optical, temporal sequences	9,996 incidents; 4.13 images per event; 54.5k QA pairs	Captions generated from news text; geotags; QA pairs	Disaster monitoring visual question answering	Multi-hazard, United States	R	Pairs imagery with news articles covering each event; captions are generated by a language model from news text, with alignment checked by humans on 144 sampled events

Table 13 (continued)

Satellite and airborne remote sensing									
Name	**Year**	**Source/ platform**	**Modalities**	**Scale**	**Annotation**	**Task**	**Domain**	**Env.**	**Description**
BCWildfire (Xu et al. 2026)	2026	Satellite, reanalysis and map data, 1 km	Reflectance, thermal anomalies, land cover, weather, terrain, infrastructure	2.4M samples; 240M ha; 25 years at daily resolution	Binary daily active-fire masks from high-confidence detections	Next-day wildfire risk prediction	Boreal wildfire, British Columbia and surrounding regions	R	A 25-year daily record (2000-2024); includes anthropogenic proximity factors alongside physical drivers (preparedness)
Ground-level and crowd-sourced imagery									
Name	**Year**	**Source/ platform**	**Modalities**	**Scale**	**Annotation**	**Task**	**Domain**	**Env.**	**Description**
Incidents (Weber et al. 2020)	2020	Web image search	Ground-level optical	446,684 positive and 697,464 negative images	Human-verified positive and negative labels; 43 incidents, 49 places	Incident and place classification	Multi-hazard	R	Establishes that incident imagery can be harvested at scale; data and models released through the project website
Incidents1M (Weber et al. 2023)	2023	Web image search	Ground-level optical	977,088 positive images; 1,787,154 in total	Multi-label positive and hard-negative labels over incidents and places	Multi-label incident detection and place classification	Multi-hazard	R	Extends the above roughly twofold and moves to multi-label annotation; 275,381 images carry more than one incident label
EIDSeg (Huang et al. 2026)	2026	Social media and web search	Ground-level optical	3,266 images	Pixel masks over five building and road classes; >70% inter-annotator agreement	Post-earthquake damage segmentation	Earthquake; nine events, 2008-2023	R	By its authors' account, the first large-scale segmentation corpus of post-earthquake social media images, merging EID images with the smaller DSS set; motivated explicitly by the cost of aerial acquisition and expert labeling

Table 14 Public datasets collected by low-altitude aerial and ground platforms, mostly uncrewed robots (LADI v2 and UMA-SAR come from crewed vehicles). Abbreviations used: Environment type (Real (R), Simulated (S), Hybrid (H)); GPS, Global Positioning System; IMU, inertial measurement unit.

Low-altitude aerial sensing									
Name	**Year**	**Source/ platform**	**Modalities**	**Scale**	**Annotation**	**Task**	**Domain**	**Env.**	**Description**
Harvey datasets (Fernandes et al. 2018)	2018	Nine sUAS models, rotorcraft and fixed-wing	Images, video, orthomosaics	112 flights; 56 missions; ∼150 GB of data products	No label set stated; files annotated by data managers; sorties tabulated by mission type and objective	Analysis of sUAS operations at a disaster	Hurricane and flood, Fort Bend County, Texas	R	Flown at the request of county emergency management from before the incident through recovery
AIDER (Kyrkou and Theocharides 2019)	2019	Web, aerial-image databases and the authors' UAV	Optical	2,545 images across four emergency classes and a normal class	Image-level class labels	Aerial emergency scene classification	Fire, flood, collapse, traffic accident	R	Assembled for onboard inference on embedded UAV hardware; deliberately imbalanced toward the normal class, with further images and classes announced
HERIDAL (Božić-Štulić et al. 2019)	2019	UAV, 30-65 m	Optical	>68,750 training patches; ∼500 labeled full-size test images	Labeled person locations	Victim detection	Wilderness search and rescue	H	The earliest of the aerial victim-detection corpora listed; staged volunteers act as missing persons at several locations, and about 3,000 training patches are synthetic, 1,000 of them pasting people photographed in a stadium into real scenes
WiSARD (Broyles et al. 2022)	2022	UAV, five airframes, 20-120 m	Optical and long-wave infrared	55,942 images: 33,786 visual, 22,156 thermal, 15,453 paired	Bounding boxes, single "human" class	Victim detection	Wilderness search and rescue	R	Twelve locations across day, night, dawn and dusk; the time-synchronized pairs offer a benchmark for sensor fusion and visual-thermal registration

Table 14 (continued)

Low-altitude aerial sensing									
Name	**Year**	**Source/ platform**	**Modalities**	**Scale**	**Annotation**	**Task**	**Domain**	**Env.**	**Description**
WIT-UAS (Jong et al. 2023)	2023	UAS over active prescribed burns	Long-wave infrared, one optical camera	6,951 frames, 2,062 with asset labels, over three burn seasons	Bounding boxes over two classes: humans and vehicles	Detection of crew and assets against the fire signature	Wildfire	R	Inverts the usual framing by protecting the responders rather than detecting the hazard; labeled by, or under the direction of, burn-site experts
FLAME (Shamsoshoara et al. 2021) (used in Zhao et al. 2024a)	2021	Two drones (Matrice 200, Phantom 3)	Optical and thermal video; extracted optical frames	Six videos; 39,375 training/validation and 8,617 test frames, 254×254; 2,003 frames with fire masks	Frame-level fire and no-fire labels; binary fire masks; raw videos unlabeled	Fire classification and segmentation	Wildfire; prescribed slash-pile burn, Arizona	R	The surveyed work re-cuts the original dataset into 749 clips of 64 frames for video classification
FIReStereo (Dhrafani et al. 2025)	2025	UAS rig, flown and hand-carried, 24.6 cm stereo baseline	Stereo long-wave infrared, LiDAR, IMU	204,594 stereo thermal images; 16 trajectories; 58% smoke-affected	Ground-truth depth maps for 35,706 frames	Stereo depth estimation under degraded visibility	Wildfire, wildland-urban interface	R	Thermal stereo depth under smoke from training pots and, in some sequences, prescribed fires; day, night, rain and cloud
DroneAudioSet (Gupta et al. 2025)	2025	Frame-mounted quadcopters, 17 microphones	Audio, 16 kHz	23.5 h; signal-to-noise ratio −57.2 to −2.5 dB	Human-presence labels over three sound categories	Victim detection by sound under rotor ego-noise	Search and rescue, indoor	H	The only acoustic victim-detection resource listed; real rotor noise and room acoustics, with human sounds reproduced through a loudspeaker

Table 14 (continued)

Low-altitude aerial sensing									
Name	**Year**	**Source/ platform**	**Modalities**	**Scale**	**Annotation**	**Task**	**Domain**	**Env.**	**Description**
LADI v2 (Scheele et al. 2025)	2025	Crewed light aircraft, Civil Air Patrol	Optical, oblique and nadir	9,963 images; >50 events, 2015-2023; 161.4 km^2	Multi-label image labels aligned to federal damage-assessment criteria	Multi-label disaster image classification	Multi-hazard, United States	R	Forty-six volunteer annotators with majority voting; train and test are split by year, making distribution shift explicit
Road damage benchmark (Manzini et al. 2026)	2026	sUAS, 1.93-12.7 cm/px	Optical orthomosaics	657.25 km of labeled roads; 9,184 alignment annotations	Spatially aligned polygons over a ten-class condition schema	Road damage assessment and road alignment	Ten federally declared disasters	R	Derived from operational response imagery; one baseline model qualitatively validated in two 2024 hurricane responses
Ground and subterranean robot field data									
Name	**Year**	**Source/ platform**	**Modalities**	**Scale**	**Annotation**	**Task**	**Domain**	**Env.**	**Description**
SubT-Tunnel (Rogers et al. 2020)	2020	Teleoperated tracked and wheeled ground robots	3D LiDAR, stereo, thermal, IMU, radio signal strength	Five runs (four Tunnel Circuit, one STIX); two mine sites; 20 artifacts per Tunnel Circuit course	Professionally surveyed artifact ground truth; fiducial references	Mapping and localization evaluation	Underground mine rescue	R	Introduces a mapping metric based on artifacts localized within five meters of their surveyed positions, rather than accumulated trajectory error
GBPlanner field data (Dang et al. 2020)	2020	Two aerial robots and the ANYmal legged robot	LiDAR point clouds; camera data where applicable	Six sites (five mines, one bunker); exploration paths of ∼180-300 m	No labels stated; planning, localization and mapping data	Reproduction of subterranean exploration experiments	Underground mines and a bunker, United States and Switzerland	R	Active and abandoned longwall and room-and-pillar mines, including runs at the DARPA Tunnel Circuit and its integration exercise; released with the open-source planner

Table 14 (continued)

Ground and subterranean robot field data									
Name	**Year**	**Source/ platform**	**Modalities**	**Scale**	**Annotation**	**Task**	**Domain**	**Env.**	**Description**
UMA-SAR (Morales et al. 2021)	2021	Crewed all-terrain vehicle emulating a UGV	Optical and thermal cameras, 3D LiDAR, IMU, two GPS receivers	Seven sequences (four during exercises, three of the empty site); 10.1 km; 123 min	Raw sensor data with GPS ground truth; no object labels	Multi-modal mapping and scene perception	Earthquake and attack response exercises	R	The only listed ground-vehicle corpus recorded in part during real responder training exercises, containing rubble mounds, crushed vehicles, victim stand-ins and a command post
SubT Urban (Rogers et al. 2021)	2021	Tracked ground robot and UAV	3D LiDAR, stereo, RGB-D (aerial), carbon-dioxide sensor, IMU	Six runs (four ground, two aerial), 356-676 m each; two courses	Surveyed artifact ground truth; fiducial references	Evaluation of mapping and localization	Urban underground disaster recovery	R	Recorded inside a partially completed nuclear reactor building abandoned for decades
NeBula multi-robot dataset (Chang et al. 2022)	2022	Wheeled (Husky) and quadruped (Spot) robots	LiDAR keyed scans in odometry pose graphs	Four datasets; 2-4 robots and 1.2-6 km combined each	Ground-truth odometry from surveyed maps; true and false loop-closure pairs	Multi-robot loop closure detection and pose graph optimization	Underground search and rescue	R	Back-end data from a coal mine, a limestone mine, an abandoned nuclear power plant and a SubT Final Event preliminary run; false loop closures supplied to stress robustness
NeBula odometry dataset (Reinke et al. 2022)	2022	Wheeled and legged robots (Husky, Spot)	LiDAR, IMU, wheel-, kinematic- and visual-inertial odometry, cameras	Ten sequences (A-J); 11 h; >16 km	Ground-truth trajectories registered to survey-grade maps, with manual post-processing	LiDAR odometry and SLAM evaluation	Subterranean urban, tunnel and cave settings	R	A limestone mine, a three-level subway setting and lava tubes, under fog, dust, darkness and geometric degeneracy

Table 14 (continued)

Ground and subterranean robot field data									
Name	**Year**	**Source/ platform**	**Modalities**	**Scale**	**Annotation**	**Task**	**Domain**	**Env.**	**Description**
Team MARBLE Final Event data (Biggie et al. 2023)	2023	Two wheeled Husky and two legged Spot robots	3D LiDAR, RGB cameras, IMU, CO_2, Bluetooth	One competition run; per robot, an input and an output rosbag	Raw sensor inputs and visualization outputs; no labels stated	Autonomous subterranean exploration	Underground search and rescue; tunnel, urban and cave course	R	Recorded during a scored run on the DARPA-designed Final Event course in Louisville
SubT-MRS (Zhao et al. 2024b)	2024	Wheeled, legged, aerial and handheld platforms	LiDAR, fisheye, monocular, depth and thermal cameras, IMU	>30 scenes; >500 h; >100 km; 2019-2023	Laser-scanned ground-truth maps and trajectories	Multi-modal, multi-robot, multi-degraded localization and mapping	Subterranean and degraded environments	R	Spans fog, dust, smoke, snow and structureless corridors, and introduces a robustness metric alongside accuracy evaluation
CTU-CRAS-NORLAB SubT data (Petrlík et al. 2025)	2025	UAVs; UAV and UGV cameras for the detection images	Optical images; unsynchronized LiDAR and IMU	37,820 detection images (6,820 background); aerial SLAM data from cave, tunnel and mine	Manual artifact boxes and classes; surveyed-map ground-truth poses	Artifact detection; SLAM evaluation	Subterranean search and rescue	R	Organizer-specified artifacts imaged by ground and aerial robots in the tunnel and urban circuits, Bull Rock Cave and industrial buildings
MineInsight (Malizia et al. 2026)	2026	Ground robot with a manipulator, dual viewpoint	Two LiDARs, optical, monochrome, visible-shortwave and long-wave infrared	∼38k optical, ∼53k visible-shortwave and ∼108k long-wave infrared frames from the arm sensors	Bounding boxes from an automated pipeline with supervised refinement	Landmine detection	Humanitarian demining	R	Fifteen inert landmines and 20 distractor objects on three off-road tracks, recorded in daylight and, on one track, at twilight

Table 15 Public simulation environments, embodied benchmarks, and human-robot interaction and responder datasets. Abbreviations used: Environment type (Real (R), Simulated (S), Hybrid (H)).

Simulation environments and embodied benchmarks									
Name	**Year**	**Source/ platform**	**Modalities**	**Scale**	**Annotation**	**Task**	**Domain**	**Env.**	**Description**
HAZARD (Zhou et al. 2024)	2024	Physics simulator with a humanoid embodied agent	RGB-D, semantic masks, temperature and water-level fields	Three scenarios; 100 procedurally generated scenes per task	Interactive benchmark rather than an annotated corpus	Embodied decision making under an evolving hazard	Fire, flood, wind	S	Built on ThreeDWorld; the hazard evolves while the agent acts, and a perceptual variant withholds segmentation
GSL-Bench (Erwich et al. 2024)	2024	Simulated multirotor on a photorealistic simulator with CFD	Simulated gas and wind sensors, photorealistic rendering	Six warehouse environments; 1,620 runs	Simulation ground truth: source location, wind field, gas concentration field	Gas source localization	Hazardous-material and gas release	S	Motivated by emergency response to gas leaks; couples procedural environment generation with filament-based dispersion modeling
Nuclear coverage environment (Blake and Schubert 2025)	2025	Gymnasium simulation of a helicopter or UAV	Simulated local dose rates, measured area, elevation data	A limited number of radiation scenarios plus derived variants; flat and mountainous terrain	Interactive environment rather than an annotated corpus	Radiation-aware aerial coverage path planning	Nuclear emergencies, including nuclear explosions	S	Scenarios from the German Federal Office for Radiation Protection over AW3D30 elevation data, with altitude-dependent measurement footprints

Table 15 (continued)

Human-robot interaction and responder data									
Name	**Year**	**Source/ platform**	**Modalities**	**Scale**	**Annotation**	**Task**	**Domain**	**Env.**	**Description**
RW4T (Orlov Savko et al. 2024)	2024	Simulated robot drone teamed with a human participant	Heart rate, pupillometry, interaction logs	20 participants; five tasks each; 100 trajectories	State-action-reward sequences; physiological and self-reported cognitive state	Modeling of operator behavior, workload and intent	Generic disaster response	H	The robot and the disaster are simulated; the participant behavior and physiology are real
Fire360 (Tiwari et al. 2025)	2025	Helmet-mounted, handheld and tripod-mounted omnidirectional camera	Omnidirectional video with aligned audio	228 videos; 50 h; 348 action segments; 100 question-answer items	Temporal action segments, object boxes, smoke grading, question-answer pairs	Perception and episodic memory in degraded video	Firefighting, indoor and outdoor	R	Recorded during professionally conducted training drills rather than incidents
HRT-SA (Senaratne et al. 2026)	2026	Simulated subterranean mission (tunnel and cave) with one or two robots	Eye tracking, electrodermal activity, photoplethysmogram, skin temperature, robot logs	30 participants; two runs for 28 of them; ∼90 min per session	Ground-truth perceptual and comprehension situational-awareness latency at assistance events	Online assessment of operator situational awareness	Subterranean search and rescue	H	By its authors' account, the first public corpus supporting continuous rather than questionnaire-based situational-awareness assessment in remote teaming

appears in the listed set until 2025. The intervening years brought hazard-observation corpora (Tellman et al. 2021; Kitamoto et al. 2023; Gerard et al. 2023) and ground and aerial field data (Rogers et al. 2020; Dang et al. 2020; Morales et al. 2021; Rogers et al. 2021; Broyles et al. 2022; Chang et al. 2022; Reinke et al. 2022; Jong et al. 2023; Biggie et al. 2023). Hence, the drought concerned one task rather than the field's data production as a whole. Moreover, most entries date from 2024-2026.

- Overhead sensing forms the largest group, while embodied disaster data concentrate on subterranean settings: Satellite and airborne remote sensing supplies the largest single group. In contrast, the ground and subterranean resources come mostly from DARPA SubT Challenge courses and sites (Rogers et al. 2020; Dang et al. 2020; Rogers et al. 2021; Chang et al. 2022; Reinke et al. 2022; Biggie et al. 2023; Zhao et al. 2024b; Petrlík et al. 2025). The only other settings are a responder training site recorded by a crewed vehicle emulating a UGV (Morales et al. 2021), and prepared landmine tracks (Malizia et al. 2026).
- Ground robot data originates in exercises, competitions and test sites, and only aerial data comes from incidents: Every ground and subterranean entry was recorded at a responder training site (Morales et al. 2021), on a competition course (Rogers et al. 2020, 2021; Biggie et al. 2023), partly on such a course and partly at selected sites (Dang et al. 2020; Chang et al. 2022; Reinke et al. 2022; Zhao et al. 2024b; Petrlík et al. 2025), or on prepared test tracks (Malizia et al. 2026). The same holds for the thermal aerial corpora, from prescribed burns, staged smoke trials and flights over volunteer subjects (Jong et al. 2023; Dhrafani et al. 2025; Broyles et al. 2022), as well as for the firefighting video, from training drills (Tiwari et al. 2025). In contrast, within the low-altitude and ground-level entries, imagery of actual emergencies appears only in aerial ones, namely LADI v2, the road damage benchmark and the web-harvested part of AIDER, all annotated offline (Scheele et al. 2025; Manzini et al. 2026; Kyrkou and Theocharides 2019). Moreover, the data of a hurricane response were released without stated labels (Fernandes et al. 2018).
- Modality coverage narrows sharply away from color imagery: Thermal imaging appears in a small group of ground and aerial entries (Rogers et al. 2020; Morales et al. 2021; Broyles et al. 2022; Jong et al. 2023; Zhao et al. 2024b; Dhrafani et al. 2025; Malizia et al. 2026), the FLAME pile-burn videos among them, and in a few overhead entries (Kitamoto et al. 2023; Soni et al. 2025; Xu et al. 2026). In contrast, SAR appears only in overhead entries (Bountos et al. 2024; Victor et al. 2024; Soni et al. 2025; Danish et al. 2025; Wang et al. 2025), while acoustic sensing appears in DroneAudioSet alone (Gupta et al. 2025). Fire360 includes temporally aligned audio, which its benchmark tasks do not use (Tiwari et al. 2025). Finally, no listed dataset provides tactile or force information.
- The gaps in the data reproduce the thin categories of the taxonomy: None of the listed resources addresses marine or underwater emergencies or rubble voids. Moreover, only MineInsight addresses humanitarian demining (Malizia et al. 2026). Furthermore, none contains radiological measurements. The nearest are a competition course built inside an unfinished nuclear reactor building (Rogers et al. 2021), and the simulated dose rates of an aerial coverage environment (Blake and Schubert 2025).

- Language-shaped resources lead the 2025 entries and are largely derivative: The instruction, caption and question-answering corpora (Irvin et al. 2025; Soni et al. 2025; Danish et al. 2025; Wang et al. 2025; Chen et al. 2025; Revankar et al. 2025) form the largest group among the 2025 entries. Yet, they are largely built by re-annotating existing imagery. In particular, xBD recurs as the underlying source in several of them, while multiple ones rely, wholly or in part, on model-generated annotation.
- Public simulation resources are far thinner than the field's reliance on simulation: Only HAZARD, GSL-Bench and a nuclear coverage environment are listed among the simulation environments and embodied benchmarks (Zhou et al. 2024; Erwich et al. 2024; Blake and Schubert 2025). None of them reproduces a full disaster site. In particular, a humanoid agent moves through rooms and outdoor regions affected by fire, flood or wind, while a multirotor moves through warehouses with a gas release. Moreover, a helicopter or UAV flies over terrain with simulated dose rates.

16 Current challenges

Despite the progress reviewed in Sections 5-15, significant challenges still limit the reliability, acceptance and adoption of robotic systems at real emergencies. Some of them are structural, as each disaster is different and cannot be accurately modeled (Murphy 2014). Others concern how methods handle uncertainty, degraded sensing and pre-trained models. The challenges are organized into seven groups, namely: a) Data and annotation aspects. b) Sensing and mobility under field conditions. c) Communication and computation aspects. d) Autonomy under safety criticality. e) Foundation model transfer aspects. f) Human-side and organizational bottlenecks. g) Evaluation and reproducibility aspects.

16.1 Data and annotation aspects

Learning-based disaster robotics needs data from rare (Revankar et al. 2025) and non-repetitive (Daun and von Stryk 2023) events, while a dedicated open search-and-rescue dataset is still needed (Chen et al. 2026). The challenges concern where embodied data are recorded, whether incident data return to robots and which modalities are covered. Moreover, they concern how language-shaped corpora grow and which categories remain thin. In this respect, the following specific challenges are present:

- Embodied disaster data concentrate on subterranean settings: The public ground and subterranean resources serve mainly underground mapping, localization, exploration and artifact search (Section 15.1). In contrast, tracked, wheeled and aerial platforms are primary in most works (Section 6.8). Hence, onboard ground perception (Section 7) outside subterranean settings is thinly supported by public data. This holds notably in collapse rubble, where small ground robots have been used most frequently (Murphy et al. 2016).
- Data acquired at incidents do not return to the robot: The ground and subterranean corpora come from training sites, competition courses, researcher-chosen sites or prepared tracks, while incident data return only as aerial imagery (Section 15.1). Hence, these corpora leave supervised strategies (Section 9.1) no ground-robot incident data (Section 14.6).

- Data for darkness, smoke, cloud and contact are scarce: Thermal, orbital radar and audio corpora are few, while no listed resource holds tactile or force data (Section 15.1). Hence, perception through smoke, darkness or cloud, as well as learned contact skills (Sections 5.6 and 6.6), rest on little public data.
- Derivative growth of language-shaped corpora: The language-shaped resources leading the 2025 releases relabel imagery gathered for other purposes (Section 15.1). Hence, given the relative rarity of natural disasters (Revankar et al. 2025), the strategies of Sections 7.6 and 9.6 rest on growth in annotation rather than acquisition.
- Data gaps left by limited physical access: Marine and underwater emergencies, rubble voids and radiological measurements are absent from the listed resources, while demining appears only with inert mines on prepared tracks (Section 15.1). These gaps fall in some of the least populated categories (Sections 5.5, 6.5 and 13.6). Hence, they suggest physical access rather than annotation cost as the binding constraint.

16.2 Sensing and mobility under field conditions

Disaster sites constrain what a robot can enter, carry and still sense, from openings as small as 3 cm to smoke, explosive atmospheres and radiation (Murphy 2014). The challenges span the access and payload trade-off, perceptual degradation and contact and confined-space capability, as well as the reliability and hardening of fielded platforms. In this respect, the following specific challenges are present:

- Access and payload trade-off: The tendency of access and payload to trade off (Section 6.8) constrains the robots meant for collapse voids, where survivors may be located. For instance, growing robots able to enter these voids still require further development of their sensing (McFarland et al. 2024; Lee et al. 2025). In parallel, heterogeneous teaming (Section 12.2) transfers the difficulty to coordination rather than removing it.
- Perceptual degradation as the normal operating condition: Methods designed to survive smoke, dust, darkness, degenerate geometry and the absence of a satellite fix (Tuna et al. 2024; Zeng et al. 2024; Ebadi et al. 2024) remain concentrated in the geometric layer (Section 7.1). In contrast, the semantic and linguistic layers are mostly evaluated on clean imagery. Such imagery is judged unsuitable for low-visibility emergencies by a smoke-degraded firefighting benchmark (Tiwari et al. 2025) (Section 13.3). Hence, robustness does not propagate up the ladder of abstraction (Section 7.7).
- Immaturity of contact and confined-space capability: Recent physical intervention work rests mainly on laboratory and simulation studies (Peiros et al. 2023; Xiao et al. 2023; Chaoxia et al. 2025), with isolated real-incident cases (Lichtenfeld et al. 2025; Sakaue et al. 2025). Similarly, rubble access rests on two field studies on a training rubble pile (McFarland et al. 2024) and on offline analysis of incident imagery (Bal et al. 2023). Moreover, no listed dataset (Section 15) can yet replace physical contact in evaluation. Hence, these tasks (Sections 5.5 and 5.6) cannot be advanced by data alone.
- Mechanical reliability and field hardening: Real-incident evidence (Section 14.6) comes almost entirely from tracked, wheeled and small aerial platforms (Surmann et al. 2024; Manzini et al. 2023b; Sakaue et al. 2025). In contrast, legged platforms, expected to enter rescue use with advances in reliability (Delmerico et al. 2019), have no incident record

in the corpus. Moreover, continuum and marine main platforms have only isolated earlier deployments (Whitman et al. 2018; Murphy et al. 2012) (Section 6.8). Survivability and radiation- or heat-tolerance are seldom quantified, irradiation tests of a Fukushima Daiichi robot excepted (Sakaue et al. 2025). Furthermore, incident use entails decontamination (Lichtenfeld et al. 2025; Surmann et al. 2024), and can require replacing seals and tracks damaged by a corrosive atmosphere (Lichtenfeld et al. 2025). However, few platforms are built to be easily cleaned without exposing personnel (Murphy et al. 2016).

16.3 Communication and computation aspects

Communication and computation act as design variables rather than as assumptions, since first-response robots must be developed for limited cloud connectivity (Adams et al. 2024). The challenges concern the infrastructure that configurations presuppose, as well as the fit and thinness of the communication evidence. Moreover, they concern onboard budgets and the late delivery of data products. In this respect, the following specific challenges are present:

- Field maturity falls as assumed infrastructure rises: Demonstrated field maturity broadly decreases as configurations presuppose more communication infrastructure (Section 12.8). Real incidents (Section 14.6) evidence mainly individually operated platforms (Manzini et al. 2023b; Surmann et al. 2024). This holds although rescue missions usually face challenging network conditions, with signals attenuated through layers of collapsed walls (Surmann et al. 2024).
- The evidence base fits the conditions it describes poorly: Decentralized operation, the closest fit for a disconnected site, has been demonstrated outdoors by drone teams and by ground robots on a competition course (Horyna et al. 2023; Dah-Achinanon et al. 2023; O'Brien et al. 2023). In contrast, centralized allocation, which assumes a connectivity this corpus reports as insufficient, rests on simulations, a mine field trial and a competition entry (Maresca et al. 2025; Dahlquist et al. 2025; Hickling et al. 2026) (Sections 12.3 and 12.4). Neither has been tested at a real incident (Section 14.6).
- Thin and unpropagated communication evidence: What a disaster site supplies is characterized by a few field measurements and analyses (Section 12.5). For instance, commercial 5G was not operational inside the tunnels of a search-and-rescue exercise (Bravo-Arrabal et al. 2023), while multiple networks were unable to carry the Hurricane Ian drone data (Manzini et al. 2023c). Link limitations were also noted at incidents (Surmann et al. 2024; Lichtenfeld et al. 2025). However, these results are not yet reflected in allocation formulations assuming cellular coverage or links between vehicles (Romero et al. 2024; Werner et al. 2024) (Section 8.5).
- Onboard budgets and the offboard dependence of planning models: Complete onboard systems remain conventional, whereas planning language models (Section 8.6) are mostly reached offboard (Sections 12.6 and 12.7). For instance, a planner's Internet queries need a network infrastructure that a disaster site may not provide (Ravichandran et al. 2025). Moreover, a copilot's initial query takes about a minute when run locally, against seconds when hosted (Bark et al. 2025). In contrast, low-latency onboard design has so far been shown for compact quantized detectors (Amessegher et al. 2025), while the cited works report no onboard compression of pre-trained planning models.

- Data products reach incident command late: Timely delivery of drone data to incident command has remained problematic since Hurricane Katrina (Manzini et al. 2023c). At Hurricane Ian, human error delayed data products in most logged missions (Murphy et al. 2024). Hence, reporting (Section 5.7) is bounded by the links (Section 12.5) and the data-handling workflow, not only by perception. However, research often focuses on the robots themselves rather than on the rapid delivery of data to the user (Delmerico et al. 2019).

16.4 Autonomy under safety criticality

An erroneous action by a disaster robot can bring a secondary collapse onto trapped survivors or rescue workers (Murphy 2014). Hence, the autonomy paradigms are assessed by how they represent uncertainty and how their decisions are verified. Moreover, they are assessed by whether incident evidence supports them, and whether generalist pre-trained policies close a control loop. In this respect, the following specific challenges are present:

- Inconsistent treatment of uncertainty: Uncertainty is explicit and settable in risk-aware and belief-space planning (Dixit et al. 2025; Ginting et al. 2023) (Section 8.4). In contrast, language-conditioned planning mostly leaves it to a plan validator or to replanning once postconditions fail (Ravichandran et al. 2025; Shi et al. 2025). Moreover, learned control meets it only through randomized training conditions (Sections 8.8 and 9.4). This gap matters, because inspection after a heavy storm often takes place in partially known environments that must first be explored (Ravichandran et al. 2025).
- Language-produced plans rely on checks around the model: The safeguards of language-derived plans (Sections 8.6 and 11.4) are checks around the model, rather than properties of it. In simulated emergencies, directly prompted models often missed plan steps. Moreover, except for GPT-4o, they returned apparently successful but invalid completions of unsolvable tasks (Bark et al. 2025). Similarly, a multi-robot planner reached satisfactory success with human verification after each module, in contrast to the case without it (Zhu et al. 2025). Hence, automatic plan checks (Ravichandran et al. 2025; Shi et al. 2025; Bark et al. 2025) and operator confirmation of a previewed route (Jarabo-Peñas et al. 2025) carry the reliability on which responder trust depends (Sections 10.5 and 11.6).
- The paradigms that scale operator span lack incident evidence: Supervised autonomy (Section 8.2), decentralized coordination and mission-level language specification are meant to relieve the operator burden of Section 11.5. Yet, none has been used at a real incident (Section 14.6). Moreover, in a competition final, a single supervisor was informed of artifacts that the robots had seen but not reported. However, under high workload, the supervisor still missed them (Biggie et al. 2023).
- No generalist pre-trained policy closes a control loop: No work in the corpus uses vision-language-action models or end-to-end generalist policies (Section 10.5). Hence, learned control stays task-specific (Section 8.7). RT-1, OpenVLA and π_0 (Brohan et al. 2023; Kim et al. 2024; Black et al. 2025) have no disaster counterpart, while RT-2 runs on a multi-TPU cloud service (Zitkovich et al. 2023). However, a search-and-rescue robot cannot rely on network connectivity for time-critical control (Firoozi et al. 2025) (Section 12.7). Whether this reflects unsuitability to safety-critical field conditions or merely data scarcity (Section 16.1) is unresolved.

16.5 Foundation model transfer aspects

Pre-trained vision, language and vision-language models have entered disaster robotics rapidly, but shallowly. Hence, their contribution depends on how they are adapted and how well they transfer to disaster reasoning. Moreover, it depends on where in the autonomy stack they act and which domains they reach. In this respect, the following specific challenges are present:

- Adoption is predominantly shallow: Vision foundation models and large language models are almost always taken frozen and prompted (Zheng et al. 2024b; Bark et al. 2025). Vision foundation models are fine-tuned on building-damage data only in a benchmark (Zheng et al. 2025) or as an oracle upper bound (Zheng et al. 2024b). In contrast, only vision-language models (Section 10.4) have been instruction-tuned on disaster data (Irvin et al. 2025; Soni et al. 2025; Wang et al. 2025; Chen et al. 2025; Revankar et al. 2025) (Section 9.6). Hence, the other two categories carry general-purpose priors unadapted, while even xBD covers only a small part of the world (Ahn et al. 2025).
- Limited transfer to disaster reasoning: Generic and remote sensing vision-language models transfer poorly to disaster interpretation without disaster-specific tuning (Section 7.6). Multiple models struggle on a disaster benchmark, while the best on a geospatial suite scores about twice chance (Wang et al. 2025; Danish et al. 2025). This shortfall is attributed to missing disaster training data, the optical-radar gap and insensitivity to damaged-object counts (Wang et al. 2025). It bears on the frozen-model strategies of Section 9.7.
- Pre-trained models occupy mainly the ends of the autonomy stack: They serve as front-end perception or as planning and reporting layers. In contrast, estimation, control and physical interaction remain almost entirely conventional (Section 10.1). The exceptions are a task-specific learned loop-closure ranker (Ebadi et al. 2024) and pre-trained waypoint and velocity selection in tunnels (Yamanaka et al. 2025). Moreover, learned degeneracy detection was judged data-hungry and too slow (Tuna et al. 2024). In physical intervention (Section 5.6), the models select or decompose subtasks (Zhou et al. 2024; Shi et al. 2025), while model-issued goals, plans or behavior-tree calls are executed by conventional layers (Ravichandran et al. 2025; Shi et al. 2025; Jarabo-Peñas et al. 2025).
- Concentration in the overhead-observed domains: Foundation models are applied most often to imagery observed from above (Irvin et al. 2025; Zheng et al. 2024b; Wang et al. 2025) (Section 6.8). On the ground, they are applied only in simulation, laboratories, field trials or exercises (Bark et al. 2025; Shi et al. 2025; Ravichandran et al. 2025; Jarabo-Peñas et al. 2025). Hence, rubble-void, submerged and radiological settings, as well as humanitarian demining, remain unreached.

16.6 Human-side and organizational bottlenecks

Disaster robots reach an incident through the organization that requests, transports and operates them. Moreover, they are operated by responders who attribute their fatigue and overload to time pressure, non-stop work and heat stress under protective gear (Ruan et al. 2025).

The challenges concern operator findings, acceptance thresholds, the affected public and organizational integration. In this respect, the following specific challenges are present:

- Findings about the operator rarely reach the interface: Operator studies already imply requirements, from where drone-squad errors occur at a hurricane response to fatigue alerts and workload-annotated trajectories (Murphy et al. 2024; Mehta et al. 2024; Orlov Savko et al. 2024). In contrast, interfaces (Sections 11.1 and 11.2) are judged by their own questionnaires or expert panels (Stedman et al. 2023; Baniqued et al. 2024). The two literatures remain largely separate (Section 11.8). Hence, exercise and incident evidence (Sections 14.5 and 14.6) seldom informs interface design. An exception is a requirement analysis grounded in human-error deployment failures (Daun and von Stryk 2023).
- Responder acceptance thresholds remain untested against fielded autonomy: A responder survey specifies how accurate machine-produced semantics must be to be trusted (Ruan et al. 2025). Moreover, a requirement analysis stresses transparency and robustness for tools expected to operate in untested conditions (Daun and von Stryk 2023). Yet, the deployed autonomy of Sections 7 and 8 remains untested against these levels at responder exercises or real incidents. In addition, members of a rescue team called out only a few times each year may lack prior training with a robot (Murphy et al. 2016). Low training requirements are an adoption criterion often left unaddressed by research (Delmerico et al. 2019).
- The affected public is almost absent from the corpus: Beyond the trained operator, responders take part as informants (Daun and von Stryk 2023; Ruan et al. 2025; Kleiser et al. 2026) and exercise participants (Jarabo-Peñas et al. 2025) (Section 14.5). Moreover, shared supervisory representations have been tested with expert evaluators and firefighters (Baniqued et al. 2024; Frering et al. 2023) (Section 11.2). In contrast, the affected public appears in an isolated simulated evacuation study (Hu et al. 2023), while arriving crews appear in an interview study on driverless vehicles (Lee et al. 2023) (Section 11.7).
- Integration into the response organization: An analysis of the 2010 robot deployments found that whether the agency in charge had a robot, or a partner with robots, was the biggest predictor of whether and how quickly one was deployed (Murphy 2014). Moreover, first responders often do not know how to employ robots, while commanders exposed to them in joint exercises are more likely to request one (Surmann et al. 2024). A requirement analysis accordingly places mission integration, logistics and training alongside the robot's capabilities (Daun and von Stryk 2023), as for a fire-service sniffer robot (Steinbauer-Wagner et al. 2025) (Section 11.6). Exercises and incident deployments, rather than laboratory or arena tests, are needed to demonstrate such an integration (Sections 14.5 and 14.6).

16.7 Evaluation and reproducibility aspects

Evidence in disaster robotics ranges from benchmark scores to narrative accounts of single incidents. This heterogeneity is structured rather than incidental. The challenges concern the validation ladder and its dependence on domain and platform, as well as simulator sharing,

common metrics and the reproducibility of field results. In this respect, the following specific challenges are present:

- The validation ladder describes a division of labor rather than maturation: The evidence of Section 14.7 implies that systems rarely travel the whole ladder. Its low and high levels are populated largely by different works, with no established route from simulation to incident evidence (Sections 14.1 and 14.6). Moreover, no simulation predicting robot and human performance in a disaster has been validated (Murphy et al. 2016). Furthermore, the foundation-model-based works of Section 10 sit at the two lowest levels more often than other recent works.
- The level a work attains is associated with the disaster domain and the platform: In the current period, settings whose hazards persist supply most ground-robot incident evidence and in-situ field trials (Sakaue et al. 2025; Lichtenfeld et al. 2025; Dahlquist et al. 2025). In contrast, sudden-onset events supply benchmarks and mainly aerial responses (Manzini et al. 2023b; Surmann et al. 2024). For instance, a tracked ground robot inspected only a fire-damaged apartment building (Surmann et al. 2024), while no ground-robot deployment at a major sudden-onset natural disaster is reported first-hand. The robot deployments after the 2023 Türkiye earthquake enter only as background to a method paper (Shi et al. 2025) (Sections 14.6 and 14.7).
- Simulation is widely used for validation yet rarely shared: The public disaster simulators and benchmarks (Section 15) are limited to HAZARD (Zhou et al. 2024), GSL-Bench (Erwich et al. 2024) and a nuclear aerial coverage environment (Blake and Schubert 2025). GSL-Bench was released because researchers were testing gas-source localization on their own simulators. None of the three models a structurally damaged site, while RW4T and HRT-SA were released mainly for human-robot team data (Orlov Savko et al. 2024; Senaratne et al. 2026). Hence, although simulation is the main or a secondary validation setting for many methods (Sections 14.1 and 14.7), its results are mostly reported in study-specific environments.
- No common protocol or metric set across the criteria: Success rate, coverage, map quality, detection accuracy, mission time and operator workload are not commensurable. Moreover, counts of survivors or remains miss a robot's value in establishing that an area holds no survivors (Murphy et al. 2016). No surveyed work offers a holistic protocol attributing a mission failure to the perception, planning or communication layers (Sections 7-12). Furthermore, standardized test methods and competition scoring (Jacoff et al. 2003; Krotkov et al. 2017; Orekhov and Chung 2022) rate task completion on prepared courses, rather than mission outcome at an incident (Section 14.3).
- Field and incident results are difficult to reproduce: Field trials, responder exercises and incident deployments (Sections 14.4-14.6) differ in site, weather, hardware, personnel and procedure. Moreover, each response brings together a changing mix of organizations (Murphy 2014). Hence, the levels of greatest evidential value appear to be the least repeatable, as the exercise and incident reports of the corpus illustrate (Surmann et al. 2024; Manzini et al. 2023b; Mehta et al. 2024) (Section 13).

17 Future research directions

Building on the reviewed methods, the most promising research directions are in one-to-one correspondence with the challenges of Section 16. They point toward data-grounded, field-hardened, communication-tolerant, verifiable and responder-trusted disaster robotic systems. These directions are organized into seven groups, namely: a) Embodied data acquisition and provenance. b) Field-hardened sensing and physical intervention. c) Communication-tolerant and onboard-feasible autonomy. d) Verifiable autonomy with calibrated uncertainty. e) Disaster-specific foundation models. f) Calibrated deferral, public-facing interaction and organizational integration. g) Standardized and field-realistic evaluation.

17.1 Embodied data acquisition and provenance

The missing data concentrate where robots are hardest to deploy and instrument. Hence, progress depends less on cheaper labeling than on acquiring new data. This involves routine capture at exercises and incidents, coverage of under-represented modalities and physics-grounded synthesis where access is limited, as well as growth by acquisition rather than re-annotation. In this respect, the following promising research directions emerge:

- Routine embodied capture at exercises and incidents: Addressing the subterranean concentration of embodied data and the failure of incident data to return to the robot (Section 16.1), future work could agree on recording, consent and deferred-release protocols with response organizations ahead of call-outs. Such protocols are needed, as direct collection of performance data during a disaster may interfere with the operators' performance and raise concerns of liability (Murphy et al. 2016). In this way, exercise and incident ground-robot records could feed supervised learning (Sections 9.1, 14.5 and 14.6). So far, the listed exercise corpus has been recorded on a research mission independent of the responders' ones (Morales et al. 2021).
- Coverage of the under-represented modalities: Addressing the scarcity of data for darkness, smoke, cloud and contact (Section 16.1), future acquisition could go beyond the few thermal, radar and audio corpora (Broyles et al. 2022; Dhrafani et al. 2025; Bountos et al. 2024; Gupta et al. 2025). In particular, it could cover contact and force signals, as well as non-visual signatures recorded in smoke, dust and darkness (Chen et al. 2026). Moreover, it could cover mixed rubble, where radars have not been reliable in detecting obscured victims (Murphy et al. 2016) (Sections 5.2 and 5.5).
- Physics-grounded synthesis where physical access limits acquisition: Addressing the data gaps left by limited physical access (Section 16.1), future research could extend the existing synthesis to confined and degraded environments, with an explicit account of its reality gap. This synthesis already models heat transfer, rising water and wind, gas dispersion, dose rates and overhead change imagery (Zhou et al. 2024; Erwich et al. 2024; Blake and Schubert 2025; Zheng et al. 2025). The reality gap must be accounted for, since simulated sensors in disaster environments are rarely accurate enough to test perception (Murphy et al. 2016) (Sections 9.3 and 9.4).
- Growth by acquisition rather than by re-annotation: Addressing the derivative growth of language-shaped corpora (Section 16.1), future resources could pair each annotation layer with new acquisition and report provenance. In this way, such growth would not

be read as independent evidence. The reason is that re-annotated resources (Irvin et al. 2025; Soni et al. 2025; Chen et al. 2025) carry the events and sensors of their underlying corpora (Section 15.1).

17.2 Field-hardened sensing and physical intervention

What a robot can enter, carry, sense and survive depends jointly on its design and the hazards it meets. Hence, field capability is treated as a platform property. The directions concern degradation robustness above the geometric layer and instrumented contact, as well as the joint design of access, payload and endurance, and quantified hardening. In this respect, the following promising research directions emerge:

- Degradation robustness above the geometric layer: Addressing perceptual degradation as the normal operating condition (Section 16.2), future work could extend tolerance to the semantic and linguistic layers (Section 7.7). So far, the tolerance of smoke, dust, darkness, geometric degeneracy and missing satellite positioning has been built mainly into localization and mapping (Tuna et al. 2024; Zeng et al. 2024; Ebadi et al. 2024). Degraded video from firefighter training (Tiwari et al. 2025) could serve as a test bed.
- Instrumented contact and confined-space capability: Addressing the immaturity of contact and confined-space capability (Section 16.2), future research could add force and tactile instrumentation to the platforms of Sections 6.3 and 6.6. This would go beyond laboratory and simulation intervention work (Peiros et al. 2023), as well as beyond the few rubble-access (McFarland et al. 2024) and void-detection (Bal et al. 2023) studies. In parallel, repeatable contact-rich apparatus could be developed, for instance at task-force training rubble piles. Such apparatus would give physical contact a means of evaluation that no recorded dataset yet offers.
- Access, payload and endurance as a joint design problem: Addressing the access and payload trade-off (Section 16.2), future work could budget access, payload and endurance across a heterogeneous team, coordination included, rather than per platform (Section 12.2). For instance, a growing robot for gas removal at collapsed sites still lacks integrated sensing (Lee et al. 2025). Moreover, a drone deployed into a mine through a 150 mm borehole flies for about ten minutes (Brown et al. 2026).
- Quantitative reporting of field hardening and design for decontamination: Addressing the mechanical reliability and field hardening that incident use demands (Section 16.2), future studies could report survivability, thermal and radiation tolerance, and mechanical reliability as measured quantities with their exposure conditions. This would go beyond component-level irradiation results (Sakaue et al. 2025). Moreover, decontamination and repair, which a chemical-plant mission required (Lichtenfeld et al. 2025), could be treated as design requirements.

17.3 Communication-tolerant and onboard-feasible autonomy

Disaster sites can remove the links and remote computation that several current systems presuppose. Hence, autonomy should lose capability gradually, rather than fail when the link is lost. The directions comprise intermittent connectivity as the design default

and measurement-driven communication-aware design. Moreover, they comprise onboard-feasible foundation models, graceful degradation across deployment configurations and mission-oriented delivery of data products. In this respect, the following promising research directions emerge:

- Intermittent connectivity as the design default: Addressing the poor fit of coordination evidence to a disconnected site (Section 16.3), future research could make intermittent communication the default assumption. It could build on ultraviolet signaling between neighbors (Horyna et al. 2023), rendezvous points with relay formations (Dah-Achinanon et al. 2023) and consensus-based allocation without a designated auctioneer (O'Brien et al. 2023) (Section 12.4). Moreover, it could add delay-tolerant map exchange and data muling, which was done manually at a hurricane response (Manzini et al. 2023b). Consistency guarantees that survive prolonged disconnection could also be added, and tested at responder exercises (Section 14.5).
- Measurement-driven communication-aware design: Addressing the thin and unpropagated communication evidence (Section 16.3), future work could extend measurement campaigns to further site types, such as collapsed buildings. Examples of such campaigns are those of mesh and 5G performance in exercise tunnels (Bravo-Arrabal et al. 2023), and of drone data workflows at a hurricane response (Manzini et al. 2023c). Moreover, map sharing and allocation (Sections 7.2 and 8.5) could be parameterized by measured rather than assumed link budgets.
- Onboard-feasible foundation models: Addressing the offboard dependence of planning models (Section 16.3), future research could pursue distillation, quantization and robot-base-station partitioning of inference. This would build on low-latency quantized onboard detectors (Amesseghe r et al. 2025), in line with the smaller onboard language models envisaged for a semantic planner (Ravichandran et al. 2025). Hence, a degraded link would yield degraded capability rather than none (Section 12.6).
- Graceful degradation across deployment configurations: Addressing the fall of field maturity as assumed infrastructure rises (Section 16.3), future systems could migrate between centralized, decentralized and single-robot operation within a mission (Sections 12.1 and 12.8). In such systems, fallback behavior would be a first-class requirement. For instance, at a chemical-plant emergency, a range covered only by the manufacturer's radio left a research module's assistance functions unusable (Lichtenfeld et al. 2025).
- Mission-oriented production and delivery of data products: Addressing the late arrival of data products at incident command (Section 16.3), future research could generate and quality-check them on the robot or at the edge, before the squad leaves the flight location. These products could follow formats agreed with incident command. In this respect, a hurricane-response analysis recommends edge computing, given the lack of high-bandwidth links in the first 72 hours (Manzini et al. 2023c). Moreover, the reporting task (Section 5.7) could be judged by when its products reach decision makers.

17.4 Verifiable autonomy with calibrated uncertainty

A wrong action can endanger trapped survivors and rescue workers (Murphy 2014). Hence, autonomy should earn wider scope only once its uncertainty is quantified and its decisions are

verified against site safety. In this context, supervised autonomy is a candidate middle ground, while the vision-language-action question is still to be settled. In this respect, the following promising research directions emerge:

- Uncertainty as a first-class output of the newer paradigms: Addressing the inconsistent treatment of uncertainty (Section 16.4), future work could carry the explicit uncertainty of risk-aware and belief-space planning (Dixit et al. 2025; Ginting et al. 2023) into the newer paradigms of language-conditioned planning and learned control. This uncertainty could take the form of confidence estimates calibrated on responder-exercise data and propagated to the operator. An option is conformal prediction, with which a language-instructed robot asks for help whenever more than one candidate action remains (Firoozi et al. 2025).
- Verification against site safety rather than plan validity: Addressing plan checks that target syntactic, spatial or task-level validity (Ravichandran et al. 2025; Shi et al. 2025; Bark et al. 2025), while site safety is judged, if at all, by human confirmation or verification (Jarabo-Peñas et al. 2025; Zhu et al. 2025) (Section 16.4), future research could adapt the safety filters, shielding, Hamilton-Jacobi reachability and control barrier functions of safe learning-based control (Brunke et al. 2022). These tools could intercept the outputs of pre-trained planners before execution. Their safe set would exclude what a rescue robot must avoid, such as secondary collapses or dislodged debris (Murphy 2014).
- Supervised autonomy as a candidate middle ground: Addressing the missing incident evidence of the paradigms that scale operator span (Section 16.4), future work could establish in exercises with responders how many robots one operator can supervise, and under what abstraction of intent. This would go beyond a competition final with one supervisor of four ground robots (Biggie et al. 2023). It would also test against the field-derived baseline of one person per vehicle and per payload plus one (Murphy et al. 2016) (Section 8.2).
- Settling the vision-language-action question: Addressing the absence of a generalist pre-trained policy closing a control loop (Section 16.4), future research could evaluate vision-language-action models and end-to-end generalist policies, such as OpenVLA and π_0 (Kim et al. 2024; Black et al. 2025), against task-specific policies (Section 8.7). The evaluation could use disaster-relevant embodiments and tasks, with a safe fallback to classical autonomy on local computation when the cloud link is lost (Firoozi et al. 2025). Such an evaluation would reveal whether unsuitability to safety-critical conditions or data scarcity explains their absence.

17.5 Disaster-specific foundation models

Pre-trained models enter disaster robotics with priors largely unadapted to the domain. Hence, their value depends on adaptation beyond the vision-language category and on grounding in damage appearance and response procedure. Moreover, it depends on extension past the two ends of the autonomy stack, and on reaching ground-level and hazardous settings. In this respect, the following promising research directions emerge:

- Adaptation beyond the vision-language category: Addressing the predominantly shallow adoption of pre-trained models (Section 16.5), future work could apply parameter-efficient adaptation to the mostly unadapted vision foundation model and large language model categories (Sections 10.2 and 10.3), favoring unseen regions and disaster types. It could build on low-rank adaptation (Irvin et al. 2025; Revankar et al. 2025) and full-parameter fine-tuning on disaster change captions (Chen et al. 2025), as well as on pseudo labels generated with a segmentation model (Ahn et al. 2025).
- Grounding in damage appearance and response procedure: Addressing the limited transfer to disaster reasoning (Section 16.5), future research could train and evaluate against responder-defined semantics and reporting doctrine. This transfer is currently measured with multiple-choice answers (Danish et al. 2025; Wang et al. 2025), a format credited with also preventing hallucinations (Danish et al. 2025), or with open-ended reports scored by GPT-4.1 (Wang et al. 2025). Instead, free-form reports could be judged by responders, with emphasis on hallucination-resistant, evidence-linked outputs for decision support (Section 5.7).
- Extension past the two ends of the autonomy stack: Addressing the concentration of pre-trained models at the ends of the autonomy stack (Section 16.5), future work could test pre-trained representations for estimation in degraded conditions. Moreover, it could test whether the affordance and manipulation formulations of foundation-model robotics (Firoozi et al. 2025) transfer to contact-rich intervention (Section 5.6). A starting point would be the door opening and obstacle clearance that a substantial proportion of inaccessible areas of simulated disaster zones require (Shi et al. 2025).
- Reaching the domains that are not observed from above: Addressing the concentration of foundation models on overhead imagery (Irvin et al. 2025; Zheng et al. 2024b) (Section 16.5), future research could tune and evaluate models on data captured inside rubble voids, at submerged and radiological sites, and in demining settings. This would proceed jointly with the acquisition and onboard-inference directions of Sections 17.1 and 17.3.

17.6 Calibrated deferral, public-facing interaction and organizational integration

Technology can partially alleviate, but not solve, responders' fatigue and cognitive overload (Ruan et al. 2025). Hence, a robot helps a response only when its interface follows operator-failure evidence and it signals the limits of its competence. Moreover, it must address the affected public, and be developed with the deploying organization. In this respect, the following promising research directions emerge:

- Interfaces derived from operator-failure evidence: Addressing operator findings that rarely reach the interface (Section 16.6), future work could turn the remedies and workload data of operator-failure research (Murphy et al. 2024; Mehta et al. 2024; Orlov Savko et al. 2024), as well as the low training demands (Delmerico et al. 2019), into design requirements. Interfaces (Section 11.1) could then be evaluated against them with rescue teams. Called out only a few times a year, such teams may first meet the robot at the disaster (Murphy et al. 2016).

- Calibrated confidence and explicit deferral: Addressing the untested responder thresholds, which separate the accuracy that machine-produced semantics need to earn trust from that needed to be judged useful (Ruan et al. 2025) (Sections 11.6 and 16.6), future systems could quantify the uncertainty of each output. A requirement analysis asks the same of maps for responders (Daun and von Stryk 2023). Moreover, outputs between the two levels could be treated as advisory, while the system defers to the operator below the usefulness level. Both levels could be set per task since, as one respondent noted, requirements vary with the occasion (Ruan et al. 2025).
- The affected public and shared supervisory representations: Addressing the near absence of the affected public (Section 16.6), future research could build on robot guidance of evacuees, a parked autonomous vehicle signaling responders, a facility digital twin and a wildfire common operational picture (Hu et al. 2023; Lee et al. 2023; Baniqued et al. 2024; Frering et al. 2023). On this basis, it could address non-responders, such as trapped victims in voids too dangerous for human or canine entry (Murphy et al. 2016) (Section 11.7). Moreover, it could design shared representations for incident command roles beyond the robot team supervisor (Adams et al. 2024) (Section 11.2).
- Robots developed with and inside response organizations: Addressing the integration of robots into the response organization (Section 16.6), future work could co-develop the receiving organization's procedures, roles and training with the robot. These roles could be integrated into the incident command structure, team training and validation (Adams et al. 2024). Moreover, logistics such as transport and charging could be planned with the receiving organization (Daun and von Stryk 2023). Joint exercises, held every six months by one robotic task force (Surmann et al. 2024), could serve as a route to adoption (Section 14.5). In support of this, robots were put to use in half a day on average in four 2010 deployments, whose incident-command agency or company already used robots day to day or had prior lines of authority in place (Murphy 2014). Furthermore, the operating procedure written for a fire-service sniffer robot also defined its user requirements (Steinbauer-Wagner et al. 2025).

17.7 Standardized and field-realistic evaluation

Each disaster is different (Murphy et al. 2016), while the evidence closest to an actual response is the hardest to repeat. Hence, evaluation should become comparable across works. The directions comprise defined routes across the validation ladder and shared simulators with a minimum field-reporting schema, as well as mission-level protocols co-designed with practitioners. In this respect, the following promising research directions emerge:

- Defined routes across the validation ladder: Addressing the division of labor along the validation ladder and its link to domain and platform (Section 16.7), future work could define staged protocols. These would carry a single method from simulation through standardized arenas (Jacoff et al. 2003) to responder exercises (Fernández Talavera et al. 2023), with live-virtual trials in between. Every stage would be reported, while ground robots meant for sudden-onset events would be prioritized. This is in line with calls for early formative and later summative evaluations of human-robot teams (Adams et al. 2024).

- Shared simulation environments and a minimum field-reporting schema: Addressing the rarely shared simulation and the poor reproducibility of field results (Section 16.7), future research could consolidate shared environments of damaged structures, with simulated sensing checked against field data. No listed public simulator models such structures (Zhou et al. 2024; Erwich et al. 2024; Blake and Schubert 2025). Moreover, a minimum reporting schema could be adopted for field, exercise and incident results. This schema would cover site, weather, hardware, personnel and procedure. It would also record the changing mix of organizations that a response brings together (Murphy 2014).
- Mission-level protocols co-designed with practitioners: Addressing the lack of a common protocol or metric set (Section 16.7), future work could extend standardized test methods and competition scoring (Jacoff et al. 2003; Orekhov and Chung 2022; Krotkov et al. 2017) into holistic mission-level protocols. These protocols would target scenarios co-defined with response organizations (Daun and von Stryk 2023; Ruan et al. 2025), shaped by incidents rather than competitions (Section 14.3). Moreover, they would credit a finding that an area holds no survivors (Murphy et al. 2016).

18 Conclusion

Disaster robotics seeks to extend the reach of responders into voids and sites that are too small, too hot, too toxic or too dangerous for them to enter. Its evidence base now combines more than two decades of incident deployments with a rapidly growing body of learning-based methods. This paper reviewed the field systematically, restricted to works with a concrete connection to disasters and an evaluation on disaster data or in emergency settings, whether real, replicated or simulated. In particular, it delineated the evolution of the field through five research phases, and analyzed the literature along ten criteria, with a comparative analysis and critical insights for each. Moreover, it reviewed the public datasets and benchmarks, and discussed the current challenges together with the corresponding future research directions.

Four overarching observations emerge. First, deep neural networks have spread rapidly through the recent literature. Yet, the systems that have reached real incidents rely predominantly on model-based perception, estimation, planning and control, with the deep-network literature and the deployed systems remaining largely distinct. Second, the evidence base has shifted toward simulation and benchmark evaluation. Real-incident deployments account for a smaller share of the recent works than of the earlier ones. Third, foundation models have entered the field mainly as perception front ends and mission-level planners layered over existing autonomy stacks, rather than as embodied policies. Most of the corresponding works remain at simulation and benchmark evaluation, while none has yet been deployed at a real incident. Fourth, response tasks and disaster domains are largely bound to specific embodiments. Meanwhile, ground-level damage assessment, flood response with ground or marine robots, as well as landslide and maritime emergencies remain thinly covered. The robot-borne public datasets reproduce this pattern.

Overall, several obstacles to the operational adoption of disaster robots are structural rather than algorithmic. They concern where data and evidence on disasters can be obtained, as well as what a disaster site leaves available for communication and computation. Other obstacles concern the integration of robots into response organizations. By consolidating the

phases, criteria, datasets, challenges and future directions of the field into a single reference, this review aims to provide a structured map of disaster robotics. Moreover, it aims to provide a roadmap toward robotic systems that are grounded in disaster data, hardened for the field, tolerant of degraded communication, verifiable, and trusted by the responders who operate them.

Declarations

Funding. This work has received funding from the European Union's Horizon Europe research and innovation program under Grant Agreement No. 101168042 project TRIFFID (auTonomous Robotic aId For increasing First responders Efficiency).

Competing interests. The authors declare no competing interests.

Ethics approval and consent to participate. Not applicable.

Data availability. Not applicable.

Author contributions. G.A.: Conceptualization, Methodology, Investigation, Writing - original draft, Writing - review and editing. A.P.: Conceptualization, Methodology, Investigation, Writing - original draft, Writing - review and editing. T.L.: Writing - review and editing, Supervision. V.A.: Writing - review and editing, Supervision. P.S.: Writing - review and editing, Supervision. I.V.: Methodology, Investigation, Writing - original draft, Writing - review and editing, Supervision. G.Th.P.: Conceptualization, Methodology, Investigation, Writing - original draft, Writing - review and editing, Supervision, Funding acquisition.

References

Adams JA, Scholtz J, Sciarretta A (2024) Human-robot teaming challenges for the military and first response. Annual Review of Control, Robotics, and Autonomous Systems 7(1):149–173. https://doi.org/10.1146/annurev-control-061223-124431

Agha A, Otsu K, Morrell B, et al (2022) NeBula: TEAM CoSTAR's robotic autonomy solution that won phase II of DARPA subterranean challenge. Field Robotics 2:1432–1506. https://doi.org/10.55417/fr.2022047

Ahn K, Han S, Park S, et al (2025) Generalizable disaster damage assessment via change detection with vision foundation model. In: Proc. AAAI Conf. Artificial Intelligence, pp 27784–27792, https://doi.org/10.1609/aaai.v39i27.34994

Amessegher I, Gaudard A, Anyinam-Boateng KN, et al (2025) Towards low-latency object detection on board reactive search-and-rescue drones. In: Proc. IEEE Int. Symp. Safety, Security, and Rescue Robotics (SSRR), pp 19–24, https://doi.org/10.1109/SSRR68451.2025.11391298

Arai M, Tanaka Y, Hirose S, et al (2008) Development of "Souryu-IV" and "Souryu-V": Serially connected crawler vehicles for in-rubble searching operations. Journal of Field Robotics 25(1–2):31–65. https://doi.org/10.1002/rob.20229

Bachrach A, Prentice S, He R, et al (2012) Estimation, planning, and mapping for autonomous flight using an RGB-D camera in GPS-denied environments. The International Journal of Robotics Research 31(11):1320–1343. https://doi.org/10.1177/0278364912455256

Bal A, Gupta A, Goyal P, et al (2023) Towards automated void detection for search and rescue with 3D perception. In: Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS), pp 3290–3297, https://doi.org/10.1109/IROS55552.2023.10341454

Baniqued PDE, Bremner P, Sandison M, et al (2024) Multimodal immersive digital twin platform for cyber-physical robot fleets in nuclear environments. Journal of Field Robotics 41(5):1521–1540. https://doi.org/10.1002/rob.22329

Bark F, Daun K, von Stryk O (2023) Affordance-based actionable semantic mapping and planning for mobile rescue robots. In: Proc. IEEE Int. Symp. Safety, Security, and Rescue Robotics (SSRR), pp 53–60, https://doi.org/10.1109/SSRR59696.2023.10499938

Bark F, Daun K, von Stryk O (2025) Large language model-based copilot for planning and executing of co-active USAR missions using actionable semantic scene graphs. In: Proc. IEEE Int. Symp. Safety, Security, and Rescue Robotics (SSRR), pp 146–153, https://doi.org/10.1109/SSRR68451.2025.11391310

Bellicoso CD, Bjelonic M, Wellhausen L, et al (2018) Advances in real-world applications for legged robots. Journal of Field Robotics 35(8):1311–1326. https://doi.org/10.1002/rob.21839

Biggie H, Rush ER, Riley DG, et al (2023) Flexible supervised autonomy for exploration in subterranean environments. Field Robotics 3:125–189. https://doi.org/10.55417/fr.2023004

Bircher A, Kamel M, Alexis K, et al (2016) Receding horizon "next-best-view" planner for 3D exploration. In: Proc. IEEE Int. Conf. Robotics and Automation (ICRA), pp 1462–1468, https://doi.org/10.1109/ICRA.2016.7487281

Black K, Brown N, Driess D, et al (2025) π_0: A vision-language-action flow model for general robot control. In: Robotics: Science and Systems (RSS) XXI, https://doi.org/10.15607/RSS.2025.XXI.010

Blake J, Schubert M (2025) Aerial coverage path planning in nuclear emergencies. In: Proc. Int. Joint Conf. Artificial Intelligence (IJCAI), pp 11009–11012, https://doi.org/10.24963/ijcai.2025/1251

Bountos NI, Sdraka M, Zavras A, et al (2024) Kuro Siwo: 33 billion m^2 under the water. A global multi-temporal satellite dataset for rapid flood mapping. In: Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track, pp 38105–38121, https://doi.org/10.52202/079017-1204

Božić-Štulić D, Marušić Ž, Gotovac S (2019) Deep learning approach in aerial imagery for supporting land search and rescue missions. International Journal of Computer Vision 127(9):1256–1278. https://doi.org/10.1007/s11263-019-01177-1

Bravo-Arrabal J, Álvarez-Merino CS, Fernández Lozano JJ, et al (2023) Field report on experimental comparison of a WiFi mesh network against commercial 5G in an underground disaster environment. In: Proc. IEEE Int. Symp. Safety, Security, and Rescue Robotics (SSRR), pp 186–192, https://doi.org/10.1109/SSRR59696.2023.10499931

Brohan A, Brown N, Carbajal J, et al (2023) RT-1: Robotics transformer for real-world control at scale. In: Robotics: Science and Systems (RSS) XIX, https://doi.org/10.15607/RSS.2023.XIX.025

Brown L, Clarke R, Akbari A, et al (2026) Subterranean mine inspection using the Prometheus re-configurable drone. Journal of Field Robotics 43(3):1646–1660. https://doi.org/10.1002/rob.70112

Broyles D, Hayner CR, Leung K (2022) WiSARD: A labeled visual and thermal image dataset for wilderness search and rescue. In: Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS), pp 9467–9474, https://doi.org/10.1109/IROS47612.2022.9981298

Brunke L, Greeff M, Hall AW, et al (2022) Safe learning in robotics: From learning-based control to safe reinforcement learning. Annual Review of Control, Robotics, and Autonomous Systems 5(1):411–444. https://doi.org/10.1146/annurev-control-042920-020211

Cadena C, Carlone L, Carrillo H, et al (2016) Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. IEEE Transactions on Robotics 32(6):1309–1332. https://doi.org/10.1109/TRO.2016.2624754

Cao C, Zhu H, Choset H, et al (2021) TARE: A hierarchical framework for efficiently exploring complex 3D environments. In: Robotics: Science and Systems (RSS) XVII, https://doi.org/10.15607/RSS.2021.XVII.018

Cao X, Li M, Tao Y, et al (2024) HMA-SAR: Multi-agent search and rescue for unknown located dynamic targets in completely unknown environments. IEEE Robotics and Automation Letters 9(6):5567–5574. https://doi.org/10.1109/LRA.2024.3396097

Casper J, Murphy RR (2003) Human-robot interactions during the robot-assisted urban search and rescue response at the World Trade Center. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 33(3):367–385. https://doi.org/10.1109/TSMCB.2003.811794

Chang Y, Ebadi K, Denniston CE, et al (2022) LAMP 2.0: A robust multi-robot SLAM system for operation in challenging large-scale underground environments. IEEE Robotics and Automation Letters 7(4):9175–9182. https://doi.org/10.1109/LRA.2022.3191204

Chaoxia C, Shang W, Zhou J, et al (2025) Sim-to-real transfer of automatic extinguishing strategy for firefighting robots. IEEE Robotics and Automation Letters 10(1):1–8. https://doi.org/10.1109/LRA.2024.3502059

Chen T, Rushworth A, Jia F (2026) A review on search and rescue robots in complex scenarios: Key technologies of simultaneous localisation and mapping. Journal of Field Robotics 43(7):4783–4820. https://doi.org/10.1002/rob.70256

Chen Z, Wang C, Zhang N, et al (2025) RSCC: A large-scale remote sensing change caption dataset for disaster events. In: Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track, pp 75999–76030, https://doi.org/10.52202/085713-2290

Chiou M, Epsimos GT, Nikolaou G, et al (2022) Robot-assisted nuclear disaster response: Report and insights from a field exercise. In: Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS), pp 4545–4552, https://doi.org/10.1109/IROS47612.2022.9981881

Chung TH, Orekhov V, Maio A (2023) Into the robotic depths: Analysis and insights from the DARPA subterranean challenge. Annual Review of Control, Robotics, and Autonomous Systems 6(1):477–502. https://doi.org/10.1146/annurev-control-062722-100728

Cladera F, Ravichandran Z, Hughes J, et al (2025) Air-ground collaboration for language-specified missions in unknown environments. IEEE Transactions on Field Robotics 2:626–642. https://doi.org/10.1109/TFR.2025.3584019

Cruz Ulloa C, Tolón J, del Cerro J, et al (2025) Efficient navigation for quadruped robots in post-disaster scenarios. In: Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS), pp 6742–6748, https://doi.org/10.1109/IROS60139.2025.11246824

Dah-Achinanon U, Marjani Bajestani SE, Lajoie PY, et al (2023) Search and rescue with sparsely connected swarms. Autonomous Robots 47(7):849–863. https://doi.org/10.1007/s10514-022-10080-7

Dahlquist N, Nordström S, Stathoulopoulos N, et al (2025) Deployment of an aerial multiagent system for automated task execution in large-scale underground mining environments. IEEE Transactions on Field Robotics 2:465–484. https://doi.org/10.1109/TFR.2025.3580068

Dang T, Mascarich F, Khattak S, et al (2019) Graph-based path planning for autonomous robotic exploration in subterranean environments. In: Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS), pp 3105–3112, https://doi.org/10.1109/IROS40897.2019.8968151

Dang T, Tranzatto M, Khattak S, et al (2020) Graph-based subterranean exploration path planning using aerial and legged robots. Journal of Field Robotics 37(8):1363–1388. https://doi.org/10.1002/rob.21993

Danish MS, Munir MA, Shah SRA, et al (2025) GEOBench-VLM: Benchmarking vision-language models for geospatial tasks. In: Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), pp 7132–7142, https://doi.org/10.1109/ICCV51701.2025.00670

Daun K, von Stryk O (2023) Requirements and challenges for autonomy and assistance functions for ground rescue robots in reconnaissance missions. In: Proc. IEEE Int. Symp. Safety, Security, and Rescue Robotics (SSRR), pp 83–90, https://doi.org/10.1109/SSRR59696.2023.10499930

Delmerico J, Mintchev S, Giusti A, et al (2019) The current state and future outlook of rescue robotics. Journal of Field Robotics 36(7):1171–1191. https://doi.org/10.1002/rob.21887

Dhrafani D, Liu Y, Jong A, et al (2025) FIReStereo: Forest infrared stereo dataset for UAS depth perception in visually degraded environments. IEEE Robotics and Automation Letters 10(4):3302–3309. https://doi.org/10.1109/LRA.2025.3536278

Dixit A, Fan DD, Otsu K, et al (2025) STEP: Stochastic traversability evaluation and planning for risk-aware navigation; results from the DARPA subterranean challenge. IEEE Transactions on Field Robotics 2:81–99. https://doi.org/10.1109/TFR.2024.3512433

Ebadi K, Chang Y, Palieri M, et al (2020) LAMP: Large-scale autonomous mapping and positioning for exploration of perceptually-degraded subterranean environments. In: Proc. IEEE Int. Conf. Robotics and Automation (ICRA), pp 80–86, https://doi.org/10.1109/ICRA40945.2020.9197082

Ebadi K, Bernreiter L, Biggie H, et al (2024) Present and future of SLAM in extreme environments: The DARPA SubT challenge. IEEE Transactions on Robotics 40:936–959. https://doi.org/10.1109/TRO.2023.3323938

Endres F, Hess J, Sturm J, et al (2014) 3-D mapping with an RGB-D camera. IEEE Transactions on Robotics 30(1):177–187. https://doi.org/10.1109/TRO.2013.2279412

Erwich HH, Duisterhof BP, de Croon GCHE (2024) GSL-Bench: High fidelity gas source localization benchmarking tool. In: Proc. IEEE Int. Conf. Robotics and Automation (ICRA), pp 17672–17678, https://doi.org/10.1109/ICRA57147.2024.10610755

Fallon M, Kuindersma S, Karumanchi S, et al (2015) An architecture for online affordance-based perception and whole-body planning. Journal of Field Robotics 32(2):229–254. https://doi.org/10.1002/rob.21546

Feng S, Whitman E, Xinjilefu X, et al (2015) Optimization-based full body control for the DARPA robotics challenge. Journal of Field Robotics 32(2):293–312. https://doi.org/10.1002/rob.21559

Fernandes O, Murphy R, Adams J, et al (2018) Quantitative data analysis: CRASAR small unmanned aerial systems at Hurricane Harvey. In: Proc. IEEE Int. Symp. Safety, Security, and Rescue Robotics (SSRR), pp 1–6, https://doi.org/10.1109/SSRR.2018.8468647

Fernández Talavera N, Roldán-Gómez JJ, Martín F, et al (2023) An autonomous ground robot to support firefighters' interventions in indoor emergencies. Journal of Field Robotics 40(3):614–625. https://doi.org/10.1002/rob.22150

Firoozi R, Tucker J, Tian S, et al (2025) Foundation models in robotics: Applications, challenges, and the future. The International Journal of Robotics Research 44(5):701–739. https://doi.org/10.1177/02783649241281508

Frering L, Koefler A, Huber M, et al (2023) Multi-robot support system for fighting wildfires in challenging environments: System design and field test report. In: Proc. IEEE Int. Symp. Safety, Security, and Rescue Robotics (SSRR), pp 32–38, https://doi.org/10.1109/SSRR59696.2023.10499951

Gao A, Murphy RR, Chen W, et al (2021) Progress in robotics for combating infectious diseases. Science Robotics 6(52):eabf1462. https://doi.org/10.1126/scirobotics.abf1462

Gehring C, Fankhauser P, Isler L, et al (2021) ANYmal in the field: Solving industrial inspection of an offshore HVDC platform with a quadrupedal robot. In: Field and Service Robotics (FSR), pp 247–260, https://doi.org/10.1007/978-981-15-9460-1_18

Gerard S, Zhao Y, Sullivan J (2023) WildfireSpreadTS: A dataset of multi-modal time series for wildfire spread prediction. In: Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track, pp 74515–74529, https://doi.org/10.52202/075280-3258

Ginting MF, Kim SK, Peltzer O, et al (2023) Safe and efficient navigation in extreme environments using semantic belief graphs. In: Proc. IEEE Int. Conf. Robotics and Automation (ICRA), pp 5653–5658, https://doi.org/10.1109/ICRA48891.2023.10161056

Gupta C, Ramesh S, Sasikumar P, et al (2025) DroneAudioSet: An audio dataset for drone-based search and rescue. In: Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track, pp 133021–133050, https://doi.org/10.52202/085713-4007

Gupta R, Goodman B, Patel N, et al (2019) Creating xBD: A dataset for assessing building damage from satellite imagery. In: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR) Workshops, pp 10–17

Hatazaki K, Konyo M, Isaki K, et al (2007) Active scope camera for urban search and rescue. In: Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS), pp 2596–2602, https://doi.org/10.1109/IROS.2007.4399386

Hickling T, Hogan M, Tammam A, et al (2026) Deep reinforcement learning based autonomous decision-making for cooperative uncrewed aerial vehicles: A search and rescue real world application. Journal of Field Robotics 43(7):4382–4401. https://doi.org/10.1002/rob.70248

Hirose S, Fukushima EF (2004) Snakes and strings: New robotic components for rescue operations. The International Journal of Robotics Research 23(4–5):341–349. https://doi.org/10.1177/0278364904042202

Hong S, Park G, Lee Y (2023) Robust design and task-priority control for rescue robot HURCULES. Journal of Field Robotics 40(5):1151–1171. https://doi.org/10.1002/rob.22176

Horyna J, Báča T, Walter V, et al (2023) Decentralized swarms of unmanned aerial vehicles for search and rescue operations without explicit communication. Autonomous Robots 47(1):77–93. https://doi.org/10.1007/s10514-022-10066-5

Hu Q, Komagata T, Kanbara S, et al (2026) The shifting paradigms of disaster robotics three decades of research. Journal of Field Robotics 43(6):3783–3793. https://doi.org/10.1002/rob.70223

Hu Y, Ryu J, Gundana D, et al (2023) Nudging or waiting?: Automatically synthesized robot strategies for evacuating noncompliant users in an emergency situation. In: Proc. ACM/IEEE Int. Conf. Human-Robot Interaction (HRI), pp 603–611, https://doi.org/10.1145/3568162.3576955

Huang H, Liu C, Zhang D, et al (2026) EIDSeg: A pixel-level semantic segmentation dataset for post-earthquake damage assessment from social media images. In: Proc. AAAI Conf. Artificial Intelligence, pp 38589–38597, https://doi.org/10.1609/aaai.v40i45.41201

Huang Y, Shan T, Chen F, et al (2022) DiSCo-SLAM: Distributed scan context-enabled multi-robot LiDAR SLAM with two-stage global-local graph optimization. IEEE Robotics and Automation Letters 7(2):1150–1157. https://doi.org/10.1109/LRA.2021.3138156

Hudson N, Talbot F, Cox M, et al (2022) Heterogeneous ground and air platforms, homogeneous sensing: Team CSIRO Data61's approach to the DARPA subterranean challenge. Field Robotics 2:595–636. https://doi.org/10.55417/fr.2022021

Hutter M, Gehring C, Jud D, et al (2016) ANYmal - a highly mobile and dynamic quadrupedal robot. In: Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS), pp 38–44, https://doi.org/10.1109/IROS.2016.7758092

Hwangbo J, Lee J, Dosovitskiy A, et al (2019) Learning agile and dynamic motor skills for legged robots. Science Robotics 4(26):eaau5872. https://doi.org/10.1126/scirobotics.aau5872

Illing B, Gaspers B, Schulz D (2025) Comparing intuitive virtual reality control with screen-based control for response robots with a manipulator. In: Proc. IEEE Int. Symp. Safety, Security, and Rescue Robotics (SSRR), pp 64–69, https://doi.org/10.1109/SSRR68451.2025.11391297

Irvin JA, Liu ER, Chen JC, et al (2025) TEOChat: A large vision-language assistant for temporal Earth observation data. In: Proc. Int. Conf. Learning Representations (ICLR), pp 68883–68911

Jacoff A, Messina E, Weiss BA, et al (2003) Test arenas and performance metrics for urban search and rescue robots. In: Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS), pp 3396–3403, https://doi.org/10.1109/IROS.2003.1249681

Jarabo-Peñas A, Bravo-Arrabal J, Lin-Yang Dh, et al (2025) Real-world deployment of an LLM-enabled voice-commanded UGV for logistics in SAR missions. In: Proc. IEEE Int. Symp. Safety, Security, and Rescue Robotics (SSRR), pp 57–63, https://doi.org/10.1109/SSRR68451.2025.11391314

Johnson M, Shrewsbury B, Bertrand S, et al (2015) Team IHMC's lessons learned from the DARPA robotics challenge trials. Journal of Field Robotics 32(2):192–208. https://doi.org/10.1002/rob.21571

Jong A, Yu M, Dhrafani D, et al (2023) WIT-UAS: A wildland-fire infrared thermal dataset to detect crew assets from aerial views. In: Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS), pp 11464–11471, https://doi.org/10.1109/IROS55552.2023.10341683

Khan A, Gupta S, Gupta SK (2022) Emerging UAV technology for disaster detection, mitigation, response, and preparedness. Journal of Field Robotics 39(6):905–955. https://doi.org/10.1002/rob.22075

Kim MJ, Pertsch K, Karamcheti S, et al (2024) OpenVLA: An open-source vision-language-action model. In: Proc. Conf. Robot Learning (CoRL), pp 2679–2713

Kirillov A, Mintun E, Ravi N, et al (2023) Segment anything. In: Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), pp 3992–4003, https://doi.org/10.1109/ICCV51070.2023.00371

Kitamoto A, Hwang J, Vuillod B, et al (2023) Digital Typhoon: Long-term satellite image dataset for the spatio-temporal modeling of tropical cyclones. In: Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track, pp 40623–40636, https://doi.org/10.52202/075280-1768

Kitano H, Tadokoro S (2001) RoboCup Rescue: A grand challenge for multiagent and intelligent systems. AI Magazine 22(1):39–52. https://doi.org/10.1609/aimag.v22i1.1542

Kitano H, Tadokoro S, Noda I, et al (1999) RoboCup Rescue: Search and rescue in large-scale disasters as a domain for autonomous agents research. In: Proc. IEEE Int. Conf. Systems, Man, and Cybernetics (SMC), pp 739–743, https://doi.org/10.1109/ICSMC.1999.816643

Kleiner A, Dornhege C (2007) Real-time localization and elevation mapping within urban search and rescue scenarios. Journal of Field Robotics 24(8–9):723–745. https://doi.org/10.1002/rob.20208

Kleiner A, Kümmerle R (2007) Genetic MRF model optimization for real-time victim detection in search and rescue. In: Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS), pp 3025–3030, https://doi.org/10.1109/IROS.2007.4399006

Kleiser KL, Buntsma V, Schneider S (2026) Learning from human-dog collaboration: Towards the design of search and rescue robots. In: Companion Proc. ACM/IEEE Int. Conf. Human-Robot Interaction (HRI), pp 1004–1008, https://doi.org/10.1145/3776734.3794545

Kottege N, Williams J, Tidd B, et al (2025) Heterogeneous robot teams with unified perception and autonomy: How Team CSIRO Data61 tied for the top score at the DARPA subterranean challenge. IEEE Transactions on Field Robotics 2:100–130. https://doi.org/10.1109/TFR.2024.3522063

Krotkov E, Hackett D, Jackel L, et al (2017) The DARPA robotics challenge finals: Results and perspectives. Journal of Field Robotics 34(2):229–240. https://doi.org/10.1002/rob.21683

Kruijff-Korbayová I, Freda L, Gianni M, et al (2016) Deployment of ground and aerial robots in earthquake-struck Amatrice in Italy (brief report). In: Proc. IEEE Int. Symp. Safety, Security, and Rescue Robotics (SSRR), pp 278–279, https://doi.org/10.1109/SSRR.2016.7784314

Kuindersma S, Deits R, Fallon M, et al (2016) Optimization-based locomotion planning, estimation, and control design for the Atlas humanoid robot. Autonomous Robots 40(3):429–455. https://doi.org/10.1007/s10514-015-9479-3

Kulkarni M, Dharmadhikari M, Tranzatto M, et al (2022) Autonomous teamed exploration of subterranean environments using legged and aerial robots. In: Proc. IEEE Int. Conf. Robotics and Automation (ICRA), pp 3306–3313, https://doi.org/10.1109/ICRA46639.2022.9812401

Kumar A, Fu Z, Pathak D, et al (2021) RMA: Rapid motor adaptation for legged robots. In: Robotics: Science and Systems (RSS) XVII, https://doi.org/10.15607/RSS.2021.XVII.011

Kyrkou C, Theocharides T (2019) Deep-learning-based aerial image classification for emergency response applications using unmanned aerial vehicles. In: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR) Workshops, pp 517–525, https://doi.org/10.1109/CVPRW.2019.00077

Lee J, Hwangbo J, Wellhausen L, et al (2020) Learning quadrupedal locomotion over challenging terrain. Science Robotics 5(47):eabc5986. https://doi.org/10.1126/scirobotics.abc5986

Lee S, Kim NG, Seo D, et al (2025) Helical structured soft growing robot for hazardous gas suction in inaccessible environments. In: Proc. IEEE Int. Conf. Robotics and Automation (ICRA), pp 12907–12913, https://doi.org/10.1109/ICRA55743.2025.11128325

Lee SH, Patil V, Britten N, et al (2023) Safe to approach: Insights on autonomous vehicle interaction protocols with first responders. In: Companion Proc. ACM/IEEE Int. Conf. Human-Robot Interaction (HRI), pp 399–402

Lichtenfeld J, Bark F, Grafe R, et al (2025) Lessons from the field: A case study of robotic intervention in an industrial emergency. In: Proc. IEEE Int. Symp. Safety, Security, and Rescue Robotics (SSRR), pp 160–165, https://doi.org/10.1109/SSRR68451.2025.11391312

Lim J, Lee I, Shim I, et al (2017) Robot system of DRC-HUBO+ and control strategy of team KAIST in DARPA robotics challenge finals. Journal of Field Robotics 34(4):802–829. https://doi.org/10.1002/rob.21673

Liu Y, Nejat G (2013) Robotic urban search and rescue: A survey from the control perspective. Journal of Intelligent & Robotic Systems 72(2):147–165. https://doi.org/10.1007/s10846-013-9822-x

Loquercio A, Kaufmann E, Ranftl R, et al (2021) Learning high-speed flight in the wild. Science Robotics 6(59):eabg5810. https://doi.org/10.1126/scirobotics.abg5810

Malizia M, Hamesse C, Hasselmann K, et al (2026) MineInsight: A multi-sensor dataset for humanitarian demining robotics in off-road environments. IEEE Robotics and Automation Letters 11(2):1650–1657. https://doi.org/10.1109/LRA.2025.3643265

Manzini T, Murphy RR, Heim E, et al (2023a) Harnessing AI and robotics in humanitarian assistance and disaster response. Science Robotics 8(80):eadj2767. https://doi.org/10.1126/scirobotics.adj2767

Manzini T, Murphy RR, Merrick DF (2023b) Quantitative data analysis: CRASAR small unmanned aerial systems at Hurricane Ian. In: Proc. IEEE Int. Symp. Safety, Security, and Rescue Robotics (SSRR), pp 7–12, https://doi.org/10.1109/SSRR59696.2023.10499943

Manzini T, Murphy RR, Merrick DF, et al (2023c) Wireless network demands of data products from small uncrewed aerial systems at Hurricane Ian. In: Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS), pp 9941–9946, https://doi.org/10.1109/IROS55552.2023.10342413

Manzini T, Perali P, Karnik R, et al (2026) A benchmark dataset for spatially aligned road damage assessment in small uncrewed aerial systems disaster imagery. In: Proc. AAAI Conf. Artificial Intelligence, pp 39016–39024, https://doi.org/10.1609/aaai.v40i46.41248

Maresca F, Romero A, Delgado C, et al (2025) REACT: Multi robot energy-aware orchestrator for indoor search and rescue critical tasks. In: Proc. IEEE Int. Conf. Robotics and Automation (ICRA), pp 5445–5451, https://doi.org/10.1109/ICRA55743.2025.11127906

Mascarich F, Kulkarni M, De Petris P, et al (2023) Autonomous mapping and spectroscopic analysis of distributed radiation fields using aerial robots. Autonomous Robots 47(2):139–160. https://doi.org/10.1007/s10514-022-10064-7

McCormac J, Handa A, Davison AJ, et al (2017) SemanticFusion: Dense 3D semantic mapping with convolutional neural networks. In: Proc. IEEE Int. Conf. Robotics and Automation (ICRA), pp 4628–4635, https://doi.org/10.1109/ICRA.2017.7989538

McFarland C, Dhawan A, Kumari R, et al (2024) Field insights for portable vine robots in urban search and rescue. In: Proc. IEEE Int. Symp. Safety, Security, and Rescue Robotics (SSRR), pp 190–197, https://doi.org/10.1109/SSRR62954.2024.10770046

Mehta RK, Yadav A, Murphy RR, et al (2024) Human-robot team vulnerabilities under fatigue during sUAS disaster response simulations. In: Proc. IEEE Int. Symp. Safety, Security, and Rescue Robotics (SSRR), pp 98–103, https://doi.org/10.1109/SSRR62954.2024.10770061

Michael N, Shen S, Mohta K, et al (2012) Collaborative mapping of an earthquake-damaged building via ground and aerial robots. Journal of Field Robotics 29(5):832–841. https://doi.org/10.1002/rob.21436

Micire MJ (2008) Evolution and field performance of a rescue robot. Journal of Field Robotics 25(1–2):17–30. https://doi.org/10.1002/rob.20218

Miki T, Lee J, Hwangbo J, et al (2022) Learning robust perceptive locomotion for quadrupedal robots in the wild. Science Robotics 7(62):eabk2822. https://doi.org/10.1126/scirobotics.abk2822

Mintchev S, de Rivaz S, Floreano D (2017) Insect-inspired mechanical resilience for multicopters. IEEE Robotics and Automation Letters 2(3):1248–1255. https://doi.org/10.1109/LRA.2017.2658946

Mohta K, Watterson M, Mulgaonkar Y, et al (2018) Fast, autonomous flight in GPS-denied and cluttered environments. Journal of Field Robotics 35(1):101–120. https://doi.org/10.1002/rob.21774

Morales J, Vázquez-Martín R, Mandow A, et al (2021) The UMA-SAR dataset: Multimodal data collection from a ground vehicle during outdoor disaster response training exercises. The International Journal of Robotics Research 40(6–7):835–847. https://doi.org/10.1177/02783649211004959

Morrell B, Otsu K, Agha A, et al (2024) An addendum to NeBula: Toward extending Team CoSTAR's solution to larger scale environments. IEEE Transactions on Field Robotics 1:476–526. https://doi.org/10.1109/TFR.2024.3430891

Murphy RR (2000) Marsupial and shape-shifting robots for urban search and rescue. IEEE Intelligent Systems 15(2):14–19. https://doi.org/10.1109/5254.850822

Murphy RR (2004a) Human-robot interaction in rescue robotics. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 34(2):138–153. https://doi.org/10.1109/TSMCC.2004.826267

Murphy RR (2004b) Trial by fire [rescue robots]. IEEE Robotics & Automation Magazine 11(3):50–61. https://doi.org/10.1109/MRA.2004.1337826

Murphy RR (2014) Disaster Robotics. Intelligent Robotics and Autonomous Agents, MIT Press, Cambridge, MA, USA, https://doi.org/10.7551/mitpress/9407.001.0001

Murphy RR, Stover S (2008) Rescue robots for mudslides: A descriptive study of the 2005 La Conchita mudslide response. Journal of Field Robotics 25(1–2):3–16. https://doi.org/10.1002/rob.20207

Murphy RR, Steimle E, Griffin C, et al (2008) Cooperative use of unmanned sea surface and micro aerial vehicles at Hurricane Wilma. Journal of Field Robotics 25(3):164–180. https://doi.org/10.1002/rob.20235

Murphy RR, Kravitz J, Stover SL, et al (2009) Mobile robots in mine rescue and recovery. IEEE Robotics & Automation Magazine 16(2):91–103. https://doi.org/10.1109/MRA.2009.932521

Murphy RR, Dreger KL, Newsome S, et al (2012) Marine heterogeneous multirobot systems at the Great Eastern Japan Tsunami recovery. Journal of Field Robotics 29(5):819–831. https://doi.org/10.1002/rob.21435

Murphy RR, Tadokoro S, Kleiner A (2016) Disaster robotics. In: Siciliano B, Khatib O (eds) Springer Handbook of Robotics, 2nd edn. Springer, Cham, Switzerland, p 1577–1604, https://doi.org/10.1007/978-3-319-32552-1_60

Murphy RR, Manzini T, Peres SC (2024) Analysis of human errors in small uncrewed aerial systems used during Hurricane Ian. In: Proc. IEEE Int. Symp. Safety, Security, and Rescue Robotics (SSRR), pp 64–71, https://doi.org/10.1109/SSRR62954.2024.10770053

Nagatani K, Kiribayashi S, Okada Y, et al (2011) Redesign of rescue mobile robot Quince. In: Proc. IEEE Int. Symp. Safety, Security, and Rescue Robotics (SSRR), pp 13–18, https://doi.org/10.1109/SSRR.2011.6106794

Nagatani K, Kiribayashi S, Okada Y, et al (2013) Emergency response to the nuclear accident at the Fukushima Daiichi nuclear power plants using mobile rescue robots. Journal of Field Robotics 30(1):44–63. https://doi.org/10.1002/rob.21439

Niroui F, Zhang K, Kashino Z, et al (2019) Deep reinforcement learning robot for search and rescue applications: Exploration in unknown cluttered environments. IEEE Robotics and Automation Letters 4(2):610–617. https://doi.org/10.1109/LRA.2019.2891991

Nordström S, Stathoulopoulos N, Dahlquist N, et al (2025) Safety inspections and gas monitoring in hazardous mining areas shortly after blasting using autonomous UAVs. Journal of Field Robotics 42(5):2076–2094. https://doi.org/10.1002/rob.22500

Norton A, Ober W, Baraniecki L, et al (2017) Analysis of human-robot interaction at the DARPA robotics challenge finals. The International Journal of Robotics Research 36(5–7):483–513. https://doi.org/10.1177/0278364916688254

O'Brien M, Williams J, Chen S, et al (2023) Dynamic task allocation approaches for coordinated exploration of subterranean environments. Autonomous Robots 47(8):1559–1577. https://doi.org/10.1007/s10514-023-10142-4

Orekhov VL, Chung TH (2022) The DARPA subterranean challenge: A synopsis of the circuits stage. Field Robotics 2:735–747. https://doi.org/10.55417/fr.2022024

Orlov Savko L, Qian Z, Gremillion G, et al (2024) RW4T dataset: Data of human-robot behavior and cognitive states in simulated disaster response tasks. In: Proc. ACM/IEEE Int. Conf. Human-Robot Interaction (HRI), pp 924–928, https://doi.org/10.1145/3610977.3637481

Pal SK, Sharma S, Krishnakumar N, et al (2025) Autonomous drone for dynamic smoke plume tracking. In: Proc. IEEE Int. Conf. Robotics and Automation (ICRA), pp 303–309, https://doi.org/10.1109/ICRA55743.2025.11128144

Papachristos C, Mascarich F, Khattak S, et al (2019) Localization uncertainty-aware autonomous exploration and mapping with aerial robots using receding horizon path-planning. Autonomous Robots 43(8):2131–2161. https://doi.org/10.1007/s10514-019-09864-1

Peiros E, Chiu ZY, Zhi Y, et al (2023) Finding biomechanically safe trajectories for robot manipulation of the human body in a search and rescue scenario. In: Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS), pp 167–173, https://doi.org/10.1109/IROS55552.2023.10342353

Petrlík M, Petráček P, Krátký V, et al (2025) UAVs beneath the surface: Cooperative autonomy for subterranean search and rescue in DARPA SubT. IEEE Transactions on Field Robotics 2:643–689. https://doi.org/10.1109/TFR.2024.3492160

Pham HX, La HM, Feil-Seifer D, et al (2017) A distributed control framework for a team of unmanned aerial vehicles for dynamic wildfire tracking. In: Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS), pp 6648–6653, https://doi.org/10.1109/IROS.2017.8206579

Pratt G, Manzo J (2013) The DARPA robotics challenge [competitions]. IEEE Robotics & Automation Magazine 20(2):10–12. https://doi.org/10.1109/MRA.2013.2255424

Ravichandran Z, Murali V, Tzes M, et al (2025) SPINE: Online semantic planning for missions with incomplete natural language specifications in unstructured environments. In: Proc. IEEE Int. Conf. Robotics and Automation (ICRA), pp 13714–13721, https://doi.org/10.1109/ICRA55743.2025.11128238

Recchiuto CT, Sgorbissa A (2018) Post-disaster assessment with unmanned aerial vehicles: A survey on practical implementations and research approaches. Journal of Field Robotics 35(4):459–490. https://doi.org/10.1002/rob.21756

Reinke A, Palieri M, Morrell B, et al (2022) LOCUS 2.0: Robust and computationally efficient lidar odometry for real-time 3D mapping. IEEE Robotics and Automation Letters 7(4):9043–9050. https://doi.org/10.1109/LRA.2022.3181357

Revankar S, Mall U, Phoo CP, et al (2025) MONITRS: Multimodal observations of natural incidents through remote sensing. In: Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track, pp 144930–144955, https://doi.org/10.52202/085713-4355

Rhodes C, Liu C, Westoby P, et al (2023) Autonomous search of an airborne release in urban environments using informed tree planning. Autonomous Robots 47(1):1–18. https://doi.org/10.1007/s10514-022-10063-8

Rogers JG, Gregory JM, Fink J, et al (2020) Test your SLAM! The SubT-Tunnel dataset and metric for mapping. In: Proc. IEEE Int. Conf. Robotics and Automation (ICRA), pp 955–961, https://doi.org/10.1109/ICRA40945.2020.9197156

Rogers JG, Schang A, Nieto-Granda C, et al (2021) The DARPA SubT urban circuit mapping dataset and evaluation metric. In: Experimental Robotics (ISER), Springer Proceedings in Advanced Robotics, vol 19. Springer, Cham, p 391–401, https://doi.org/10.1007/978-3-030-71151-1_35

Romero A, Delgado C, Zanzi L, et al (2024) Cellular-enabled collaborative robots planning and operations for search-and-rescue scenarios. In: Proc. IEEE Int. Conf. Robotics and Automation (ICRA), pp 5942–5948, https://doi.org/10.1109/ICRA57147.2024.10611179

Rosinol A, Abate M, Chang Y, et al (2020) Kimera: An open-source library for real-time metric-semantic localization and mapping. In: Proc. IEEE Int. Conf. Robotics and Automation (ICRA), pp 1689–1696, https://doi.org/10.1109/ICRA40945.2020.9196885

Rouček T, Pecka M, Čížek P, et al (2022) System for multi-robotic exploration of underground environments CTU-CRAS-NORLAB in the DARPA subterranean challenge. Field Robotics 2:1779–1818. https://doi.org/10.55417/fr.2022055

Ruan T, Betta Z, Tzoumas G, et al (2025) First responders' perceptions of semantic information for situational awareness in robot-assisted emergency response. In: Proc. IEEE Int. Symp. Safety, Security, and Rescue Robotics (SSRR), pp 37–44, https://doi.org/10.1109/SSRR68451.2025.11391296

Rymansaib Z, Thomas B, Treloar AA, et al (2023) A prototype autonomous robot for underwater crime scene investigation and emergency response. Journal of Field Robotics 40(5):983–1002. https://doi.org/10.1002/rob.22164

Sakaue T, Nishizawa K, Takahashi T, et al (2025) Development and deployment of radiation-tolerant remotely-control system in Fukushima Daiichi NPS. In: Proc. IEEE Int. Symp. Safety, Security, and Rescue Robotics (SSRR), pp 121–126, https://doi.org/10.1109/SSRR68451.2025.11391270

Salas-Espinales A, Vázquez-Martín R, García-Cerezo A, et al (2023) SAR nets: An evaluation of semantic segmentation networks with attention mechanisms for search and rescue scenes. In: Proc. IEEE Int. Symp. Safety, Security, and Rescue Robotics (SSRR), pp 139–144, https://doi.org/10.1109/SSRR59696.2023.10499929

Sami MT, Yan D, Adhikari S, et al (2024) EvaNet: Elevation-guided flood extent mapping on Earth imagery. In: Proc. Int. Joint Conf. Artificial Intelligence (IJCAI), pp 1200–1208, https://doi.org/10.24963/ijcai.2024/133

Scheele S, Picchione K, Liu J (2025) LADI v2: Multi-label dataset and classifiers for low-altitude disaster imagery. In: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR) Workshops, pp 2226–2234, https://doi.org/10.1109/CVPRW67362.2025.00210

Scherer S, Agrawal V, Best G, et al (2022) Resilient and modular subterranean exploration with a team of roving and flying robots. Field Robotics 2:678–734. https://doi.org/10.55417/fr.2022023

Schomer NL, Adams JA (2026) Terrain-aware probabilistic search planning for uncrewed aerial vehicles. IEEE Transactions on Field Robotics 3:40–58. https://doi.org/10.1109/TFR.2025.3641875

Schwarz M, Lenz C, Rochow A, et al (2021) NimbRo avatar: Interactive immersive telepresence with force-feedback telemanipulation. In: Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS), pp 5312–5319, https://doi.org/10.1109/IROS51168.2021.9636191

Senaratne H, Attfield R, Widhanapathirana SRJ, et al (2026) A multimodal dataset for online assessment of human situational awareness during remote human-robot teaming. IEEE Robotics and Automation Letters 11(10):11235–11242. https://doi.org/10.1109/LRA.2026.3723283

Shamsoshoara A, Afghah F, Razi A, et al (2021) Aerial imagery pile burn detection using deep learning: The FLAME dataset. Computer Networks 193:108001. https://doi.org/10.1016/j.comnet.2021.108001

Shen S, Michael N, Kumar V (2011) Autonomous multi-floor indoor navigation with a computationally constrained MAV. In: Proc. IEEE Int. Conf. Robotics and Automation (ICRA),

pp 20–25, https://doi.org/10.1109/ICRA.2011.5980357

Shi J, Huang K, Pan H, et al (2025) Autonomous subtask generation for indoor search and rescue mission via large-language-model and behavior-tree integration. In: Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS), pp 16710–16716, https://doi.org/10.1109/IROS60139.2025.11245939

Song Y, Steinweg M, Kaufmann E, et al (2021) Autonomous drone racing with deep reinforcement learning. In: Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS), pp 1205–1212, https://doi.org/10.1109/IROS51168.2021.9636053

Soni S, Dudhane A, Debary H, et al (2025) EarthDial: Turning multi-sensory Earth observations to interactive dialogues. In: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), pp 14303–14313

Stedman H, Kocer BB, van Zalk N, et al (2023) Evaluating immersive teleoperation interfaces: Coordinating robot radiation monitoring tasks in nuclear facilities. In: Proc. IEEE Int. Conf. Robotics and Automation (ICRA), pp 11972–11978, https://doi.org/10.1109/ICRA48891.2023.10161011

Steinbauer-Wagner G, Rihan YMB, Edlinger R, et al (2025) Close the valley of death - a deployable sniffer robot for hazardous substances. In: Proc. IEEE Int. Symp. Safety, Security, and Rescue Robotics (SSRR), pp 31–36, https://doi.org/10.1109/SSRR68451.2025.11391259

Surmann H, Daun K, Schnaubelt M, et al (2024) Lessons from robot-assisted disaster response deployments by the German Rescue Robotics Center task force. Journal of Field Robotics 41(3):782–797. https://doi.org/10.1002/rob.22275

Tellman B, Sullivan JA, Kuhn C, et al (2021) Satellite imaging reveals increased proportion of population exposed to floods. Nature 596(7870):80–86. https://doi.org/10.1038/s41586-021-03695-w

Tian Y, Chang Y, Herrera Arias F, et al (2022) Kimera-Multi: Robust, distributed, dense metric-semantic SLAM for multi-robot systems. IEEE Transactions on Robotics 38(4):2022–2038. https://doi.org/10.1109/TRO.2021.3137751

Tiwari A, Masoud F, Nguyen DT, et al (2025) Fire360: A benchmark for robust perception and episodic memory in degraded 360° firefighting video. In: Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track, pp 33390–33420, https://doi.org/10.52202/085713-0990

Tomić T, Schmid K, Lutz P, et al (2012) Toward a fully autonomous UAV: Research platform for indoor and outdoor urban search and rescue. IEEE Robotics & Automation Magazine 19(3):46–56. https://doi.org/10.1109/MRA.2012.2206473

Tranzatto M, Miki T, Dharmadhikari M, et al (2022) CERBERUS in the DARPA subterranean challenge. Science Robotics 7(66):eabp9742. https://doi.org/10.1126/scirobotics.abp9742

Tranzatto M, Dharmadhikari M, Bernreiter L, et al (2024) Team CERBERUS wins the DARPA subterranean challenge: Technical overview and lessons learned. Field Robotics 4:249–312. https://doi.org/10.55417/fr.2024009

Tuna T, Nubert J, Nava Y, et al (2024) X-ICP: Localizability-aware LiDAR registration for robust localization in extreme environments. IEEE Transactions on Robotics 40:452–471. https://doi.org/10.1109/TRO.2023.3335691

Victor B, Letard M, Naylor P, et al (2024) Off to new shores: A dataset and benchmark for (near-)coastal flood inundation forecasting. In: Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track, pp 114797–114811, https://doi.org/10.52202/079017-3644

Villarroel JL, Lera FJ, Tardioli D, et al (2024) RoboBoat: A robotic boat for 3D mapping of partially flooded underground sites. Journal of Field Robotics 41(4):1029–1053. https://doi.org/10.1002/rob.22303

Wang G, Wang W, Ding P, et al (2023) Development of a search and rescue robot system for the underground building environment. Journal of Field Robotics 40(3):655–683. https://doi.org/10.1002/rob.22152

Wang J, Xuan W, Qi H, et al (2025) DisasterM3: A remote sensing vision-language dataset for disaster damage assessment and response. In: Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track, pp 179475–179489, https://doi.org/10.52202/085713-5398

Weber E, Marzo N, Papadopoulos DP, et al (2020) Detecting natural disasters, damage, and incidents in the wild. In: Proc. European Conf. Computer Vision (ECCV), pp 331–350, https://doi.org/10.1007/978-3-030-58529-7_20

Weber E, Papadopoulos DP, Lapedriza A, et al (2023) Incidents1M: A large-scale dataset of images with natural disasters, damage, and incidents. IEEE Transactions on Pattern Analysis and Machine Intelligence 45(4):4768–4781. https://doi.org/10.1109/TPAMI.2022.3191996

Werner M, Báča T, Štibinger P, et al (2024) Autonomous localization of multiple ionizing radiation sources using miniature single-layer Compton cameras onboard a group of micro aerial vehicles. In: Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS), pp 5710–5717, https://doi.org/10.1109/IROS58592.2024.10802808

Whelan T, Kaess M, Johannsson H, et al (2015) Real-time large-scale dense RGB-D SLAM with volumetric fusion. The International Journal of Robotics Research 34(4–5):598–626. https://doi.org/10.1177/0278364914551008

Whitman J, Zevallos N, Travers M, et al (2018) Snake robot urban search after the 2017 Mexico City earthquake. In: Proc. IEEE Int. Symp. Safety, Security, and Rescue Robotics (SSRR), pp 1–6, https://doi.org/10.1109/SSRR.2018.8468633

Xiao C, Woeppel AB, Clepper GM, et al (2023) Tactile and chemical sensing with haptic feedback for a telepresence explosive ordnance disposal robot. IEEE Transactions on Robotics 39(5):3368–3381. https://doi.org/10.1109/TRO.2023.3278455

Xu Z, Cheng S, Wang L, et al (2026) BCWildfire: A long-term multi-factor dataset and deep learning benchmark for boreal wildfire risk prediction. In: Proc. AAAI Conf. Artificial Intelligence, pp 39486–39494, https://doi.org/10.1609/aaai.v40i46.41299

Yadav I, Sebok M, Tanner HG (2023) Receding horizon navigation and target tracking for aerial detection of transient radioactivity. The International Journal of Robotics Research 42(3):66–82. https://doi.org/10.1177/02783649231169803

Yamanaka R, Okada Y, Ohno K, et al (2025) Vision-LLM-based autonomous navigation and inspection in man-made caves using a wheeled robot equipped with multi-modal sensors. In: Proc. IEEE Int. Symp. Safety, Security, and Rescue Robotics (SSRR), pp 90–95, https://doi.org/10.1109/SSRR68451.2025.11391275

Yanco HA, Drury JL (2007) Rescuing interfaces: A multi-year study of human-robot interaction at the AAAI robot rescue competition. Autonomous Robots 22(4):333–352. https://doi.org/10.1007/s10514-006-9016-5

Yanco HA, Norton A, Ober W, et al (2015) Analysis of human-robot interaction at the DARPA robotics challenge trials. Journal of Field Robotics 32(3):420–444. https://doi.org/10.1002/rob.21568

Yang GZ, Nelson BJ, Murphy RR, et al (2020) Combating COVID-19 - the role of robotics in managing public health and infectious diseases. Science Robotics 5(40):eabb5589. https://doi.org/10.1126/scirobotics.abb5589

Zeng Z, Dang X, Li Y, et al (2024) A mmWave radar SLAM method in subterranean tunnel for low visibility and degradation. IEEE Robotics and Automation Letters 9(5):4281–4288. https://doi.org/10.1109/LRA.2024.3374191

Zhang J, Singh S (2014) LOAM: Lidar odometry and mapping in real-time. In: Robotics: Science and Systems (RSS) X, https://doi.org/10.15607/RSS.2014.X.007

Zhao L, Hu J, Bi J, et al (2024a) Streamlining forest wildfire surveillance: AI-enhanced UAVs utilizing the FLAME aerial video dataset for lightweight and efficient monitoring. In: Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS), pp 8063–8068, https://doi.org/10.1109/IROS58592.2024.10801629

Zhao S, Zhang H, Wang P, et al (2021) Super odometry: IMU-centric LiDAR-visual-inertial estimator for challenging environments. In: Proc. IEEE/RSJ Int. Conf. Intelligent Robots

and Systems (IROS), pp 8729–8736, https://doi.org/10.1109/IROS51168.2021.9635862

Zhao S, Gao Y, Wu T, et al (2024b) SubT-MRS dataset: Pushing SLAM towards all-weather environments. In: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), pp 22647–22657, https://doi.org/10.1109/CVPR52733.2024.02137

Zheng Z, Tian S, Ma A, et al (2023) Scalable multi-temporal remote sensing change data generation via simulating stochastic change process. In: Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), pp 21761–21770, https://doi.org/10.1109/ICCV51070.2023.01994

Zheng Z, Zhong Y, Ma A, et al (2024a) Single-temporal supervised learning for universal remote sensing change detection. International Journal of Computer Vision 132(12):5582–5602. https://doi.org/10.1007/s11263-024-02141-4

Zheng Z, Zhong Y, Zhang L, et al (2024b) Segment any change. In: Advances in Neural Information Processing Systems (NeurIPS), pp 81204–81224, https://doi.org/10.52202/079017-2581

Zheng Z, Ermon S, Kim D, et al (2025) Changen2: Multi-temporal remote sensing generative change foundation model. IEEE Transactions on Pattern Analysis and Machine Intelligence 47(2):725–741. https://doi.org/10.1109/TPAMI.2024.3475824

Zhou B, Gao F, Wang L, et al (2019) Robust and efficient quadrotor trajectory generation for fast autonomous flight. IEEE Robotics and Automation Letters 4(4):3529–3536. https://doi.org/10.1109/LRA.2019.2927938

Zhou N, Zhu Z, Qin H, et al (2026) A boundary adaptive neural network coverage path planning algorithm for maritime search and rescue with AUVs based on target drift prediction. Journal of Field Robotics 43(2):1008–1032. https://doi.org/10.1002/rob.70053

Zhou Q, Chen S, Wang Y, et al (2024) HAZARD challenge: Embodied decision making in dynamically changing environments. In: Proc. Int. Conf. Learning Representations (ICLR), pp 50097–50113

Zhou X, Wen X, Wang Z, et al (2022) Swarm of micro flying robots in the wild. Science Robotics 7(66):eabm5954. https://doi.org/10.1126/scirobotics.abm5954

Zhu Y, Chen J, Zhang X, et al (2025) DEXTER-LLM: Dynamic and explainable coordination of multi-robot systems in unknown environments via large language models. In: Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS), pp 10182–10189, https://doi.org/10.1109/IROS60139.2025.11247301

Zitkovich B, Yu T, Xu S, et al (2023) RT-2: Vision-language-action models transfer web knowledge to robotic control. In: Proc. Conf. Robot Learning (CoRL), pp 2165–2183