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.

When does online vision-language inference help? A survey of socially aware robot navigation

Sungsan Park
Kyung Hee University
Young-Sik Shin
Kyungpook National University
Sanghyun Kim
kim87@khu.ac.kr

Kyung Hee University

Research Article

Keywords:

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11390604/v1

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

Additional Declarations: No competing interests reported.

When does online vision-language inference help? A survey of socially aware robot navigation

Sungsan Park[1], Young-Sik Shin[2*], Sanghyun Kim[1*]

[1]Department of Mechanical Engineering, Kyung Hee University, Yongin, 17104, Republic of Korea.
[2]School of Mechanical Engineering, Kyungpook National University, Daegu, 41566, Republic of Korea.

*Corresponding author(s). E-mail(s): yshin86@knu.ac.kr; kim87@khu.ac.kr;
Contributing authors: san2642@khu.ac.kr;

Abstract

Vision-language models (VLMs) and large language models are increasingly built into the navigation stacks of robots that share space with people, where the models read cues that occupancy maps omit, such as a raised hand or a conversing group. This survey organizes 41 cases reported from 2024 to 2026, 34 on socially aware navigation and 7 adjacent mechanism extensions, by where the model enters the stack: scene interpretation, cost modulation, subgoal or path proposal, hierarchical coordination, direct action output, and offline transfer or generation. For each of these six patterns we describe representative systems and their reported results, and we review the datasets, simulators, benchmarks and metrics behind them. We then discuss what the reported comparisons establish about running the model online, separating system utility, component contribution and online necessity. In these comparisons, cue-aware systems improved the outcomes that hinge on their cue, although whole-system differences limit attribution to the VLM; in one offline study, a model selector helped on hard selections and hurt on routine ones; in one deployment, delayed advice stayed useful when fused with current plans; model-supervised students navigated without runtime teacher queries, but online, cached and distilled alternatives have not been compared with student computation included; and better answers have not yet been linked to better behavior. We also define an invocation ratio (model requests per control decision), propose a minimum reporting set for deployment comparisons, and outline open challenges in real-time reasoning, spatial grounding and safety.

Keywords: Social robot navigation, Vision-language models, Foundation models, Human–robot interaction, Inference latency, Embodied AI

1 Introduction

Socially aware navigation requires robots sharing space with people to do more than avoid collisions: they must respect conventions governing passage, personal space and group interaction (Kruse et al. 2013; Rios-Martinez et al. 2015). A contact-free path may still interrupt a conversation or enter a doorway already claimed by a pedestrian.

Geometric approaches to crowd navigation draw on repulsive potentials in the social force model (Helbing and Molnár 1995), reciprocal velocity constraints (van den Berg et al. 2011) and personal-space costs (Kirby et al. 2009). A gesture or group relation shapes such planners only once it is encoded in their state or cost. Learning-based methods relaxed these hand-specified forms, yet many influential crowd-navigation policies still use current and predicted kinematic states as their principal social representation (Chen et al. 2019; Liu et al. 2023b).

Vision-language models (VLMs) offer an interface to visual concepts and commonsense associations that are not explicit in a geometric state, and instruction-tuned variants can be queried in natural language without task-specific retraining. Recent systems give them different roles. VLM-Social-Nav uses a VLM to supply a socially appropriate velocity target for a local planner's deviation cost (Song et al. 2025b). CoNVOI pairs a Contrastive Language–Image Pre-training (CLIP) context classifier with a large VLM that selects a reference path through marked free-space regions (Sathyamoorthy et al. 2024). GSON detects social groups for a downstream planner (Luo et al. 2025), and the Social-LLaVA paper, which releases the Social robot Navigation via Explainable Interactions (SNEI) question–answer dataset, maps the language outputs of a fine-tuned VLM to robot actions in a proof-of-concept demonstration (Payandeh et al. 2025).

The gains these systems report must be read against the information each system receives. In VLM-Social-Nav, a dynamic window approach (DWA) baseline and the VLM system both succeed in every plain frontal-approach trial, but only the VLM system succeeds when the person raises a hand to stop it (Section 3.3). The stopping-gesture and yielding scenarios of BehAV, which turns instructions into behavioral costs, and the role- or region-conditioned rules of Language-Instructed Social Navigation (LISN) likewise hinge on semantic cues (Weerakoon et al. 2025; Chen et al. 2026a). Direct scene-understanding tests show a different side of these models: on SocialNav-SUB, which evaluates spatial, spatiotemporal and social visual question answering (VQA) on real navigation scenes, the best evaluated VLM falls below the trajectory-rule and human-oracle baselines in aggregate (Munje et al. 2025). A navigation gain therefore shows that a system helps, but not which module produced the gain or whether the model has to run during deployment.

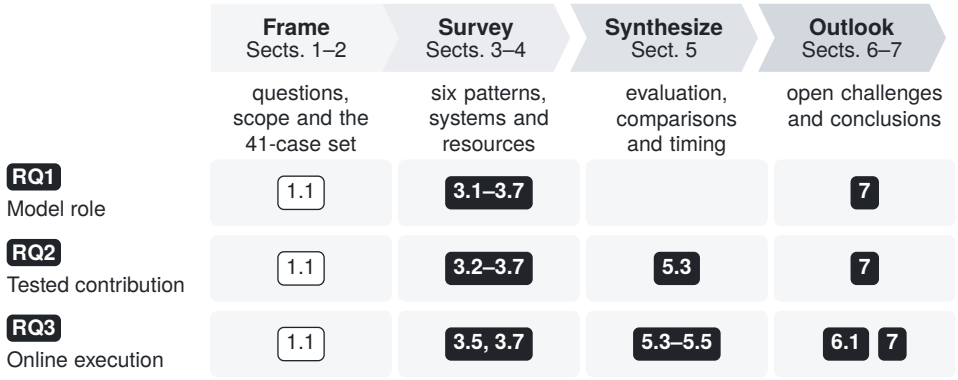


Fig. 1 Organization of the survey. The ribbon gives the four stages: the questions, background and scope (Sections 1–2); the survey of integration patterns, representative systems and research resources (Sections 3–4); the evaluation and evidence synthesis (Section 5); and the open challenges and conclusion (Sections 6–7). Each row marks where a research question is posed (outlined, Section 1.1) and the sections and subsections that answer it (filled)

1.1 Aim and research questions

This survey organizes how recent systems place vision-language and large language models in socially aware navigation and then asks *under which conditions online VLM inference adds value, and how far the reported experiments isolate its contribution.* Online inference here means executing a reviewed VLM, large language model (LLM) or vision-language-action (VLA) component during deployment, whether on board, on a local server or through a cloud service, and whether invoked continuously or conditionally; it may produce text, scores or actions, with or without an explicit verbal reasoning trace. Three research questions guide the survey: **RQ1:** What information or decision does the model supply? **RQ2:** What changes when that input or component is removed or replaced? **RQ3:** Does the comparison establish whether a benefit comes from online execution or from training-time knowledge transfer?

The survey of integration patterns (Section 3) answers RQ1, and the evidence synthesis of Section 5 takes up RQ2 and RQ3. There, removing a component tests one implementation's dependence on it, whereas necessity must also be judged against named alternatives under a stated task, performance requirement and resource budget. Figure 1 maps the three questions onto the sections that address them.

Timing bears directly on RQ3. The Slow Brain system bases its architecture on a 5–20 Hz control-loop range, and a recent review notes that local replanning commonly runs at around or above 10 Hz (Peng et al. 2026b; Cai et al. 2026). Multi-second VLM queries motivate conditional invocation, answer reuse, asynchronous coordination or offline transfer (Sections 3 and 5.4).

1.2 Contributions

This survey makes four contributions.

1. **A pattern-based map of recent systems.** We organize 41 cases from 2024 to 2026 into six overlapping integration patterns according to where the model enters the navigation stack, under a stated assignment convention and with Kruse et al.'s functional classification as a precedent (Sections 2.4 and 3).
2. **A case-level record of reported comparisons.** For each case we record the model's role, what its comparisons changed and measured, and what they support, keeping measured outcomes apart from their interpretation (Section 3).
3. **A survey of research resources and evaluation.** We relate the datasets, simulators and benchmarks behind these results to the evaluation principles of social navigation and to what answer-level scores can and cannot show (Sections 4 and 5).
4. **A synthesis of evidence on online inference.** Separating system utility, component contribution and online necessity, we state what the reported comparisons establish and which comparisons are still needed, define the invocation ratio (model requests per control decision) and propose a minimum reporting set for deployment comparisons (Sections 3.1 and 5.3–5.5).

Section 2 gives the background and the scope of the survey; Section 3 surveys the integration patterns and their representative systems; Section 4 reviews datasets, simulators and benchmarks; Section 5 moves from evaluation to an evidence synthesis, the invocation ratio and a minimum reporting set; Section 6 lists open challenges; and Section 7 concludes.

2 Background and scope

2.1 Requirements of socially aware navigation

Some social requirements can be read partly from trajectories, while others depend on relations between people and on context. The eight principles of Francis et al. (2025) organize our evaluation discussion (Table 1; Section 5): seven are objectives pursued while navigating, and the eighth sets their relative importance by context. Our evidence grouping (the rightmost column of Table 1) distinguishes trajectory proxies from the human responses and contextual judgments that full evaluation requires (Section 5.2).

Hall's proxemics links interpersonal distance to social meaning, with zones that Hall explicitly limits to the observed sample and context (Hall 1966). A questionnaire study of 8,943 participants in 42 countries found differences in stated preferences for conversational distance, a measure distinct from robot-avoidance distance (Sorokowska et al. 2017); Section 6.6 returns to this variation. Human–robot interaction research has studied physical and psychological distancing experimentally (Mumm and Mutlu 2011) and proposed a cross-cultural study methodology (Joosse et al. 2014). Goffman describes civil inattention and the tacit rules by which strangers negotiate passage (Goffman 1963, 1971). Walking groups adopt density-dependent spatial arrangements that independent personal-distance values alone do not represent (Moussaïd et al. 2010). A planner that represents only individual occupancy can treat O-space, the shared area maintained by participants in focused interaction (Rios-Martinez

Table 1 The eight evaluation principles of Francis et al. (2025), with the operational reading used in this survey. The rightmost column, added by this survey, assesses the evidence needed to evaluate each principle: *mixed* means that trajectory-based proxies capture only part of it, and *semantic or contextual* means that evaluation requires interpreting agents or context

Principle	Operational reading	Evidence needed
Safety	Protects people, robots and the environment from harm	Mixed
Comfort	Creates neither annoyance nor stress; distance, no cutting off, naturalness of motion	Mixed
Legibility	Behaves so that goals can be understood	Mixed
Politeness	Shows respect and consideration through motion and communication	Semantic or contextual
Social competency	Complies with relevant social norms	Semantic or contextual
Agent understanding	Infers and accommodates nearby agents' goals	Semantic or contextual
Proactivity	Takes the initiative to prevent or resolve interaction problems	Mixed
Contextual appropriateness	Behaves properly in context; weights the other seven	Semantic or contextual

et al. 2015), as geometrically free, which motivates the group-detection systems of Section 3.2.

2.2 Classical and learning-based social navigation

Geometric methods are both the baselines and the execution layer of many VLM systems. Reactive methods include the dynamic window approach (Fox et al. 1997) and elastic bands (Quinlan and Khatib 1993; Rösmann et al. 2017); social-force models (Helbing and Molnár 1995; Moussaïd et al. 2011) and reciprocal velocity obstacles (van den Berg et al. 2011) represent mutual avoidance; and social costs encode proxemics, visibility and soft social constraints (Hall 1966; Sisbot et al. 2007; Kirby et al. 2009). Cooperative and predictive methods model how people respond to the robot. Joint robot–pedestrian models address the freezing that independent forecasts, which omit people's responses to the robot, can cause even when the forecasts are accurate (Trautman and Krause 2010; Trautman et al. 2015); SICNav embeds an optimal reciprocal collision avoidance (ORCA) pedestrian model in bilevel model predictive control (MPC) (Samavi et al. 2025); and legibility and passing-structure methods communicate intent through motion (Dragan et al. 2013; Mavrogiannis et al. 2018). Trajectory predictors from Social LSTM to Trajectron++ (Alahi et al. 2016; Gupta et al. 2018; Salzmann et al. 2020) provide forecasts that surveys compare (Rudenko et al. 2020b; Kothari et al. 2022). Learned crowd-navigation policies relax hand-designed response rules (Chen et al. 2019; Everett et al. 2021; Liu et al. 2023b) and learn from rewards, from inverse reinforcement learning (Kretzschmar et al. 2016) or from demonstrations such as those of the Socially CompliAnt Navigation Dataset (SCAND) (Karnan et al.

2022). Across these families, social cues such as gestures act only through the planner's state, prediction or cost (Section 1), and a semantic module can supply them while leaving geometric control in place.

These methods can remain strong baselines, depending on the dataset, metric and evaluation setting. On SCAND demonstrations, Raj et al. (2024) report that `move_base`, which lacks an explicit social-cost term, aligns with the demonstrations in about 80% of scenarios at a Hausdorff-distance threshold of 1 m, the highest rate among four geometric planners; the authors also route predicted difficult cases to a learned policy. In simulation, a simple stop-and-go planner can compete with learned policies on time and collisions in the evaluated crowds but fails above a certain crowd density in the social-force environment (Mavrogiannis et al. 2023). In another simulator and scenario mix, several learned policies outperform geometric baselines on success and collisions at doorways and in open spaces, while a DWA baseline leads on corridor success (Alyassi et al. 2025). Comfort ratings and perceived human-likeness can also favor different settings (Zhou et al. 2022), so trajectory similarity, perceived human-likeness and experienced comfort measure different things. A semantic component's added value should therefore be measured against a strong baseline in scenario-specific comparisons.

2.3 Vision-language models and language-guided navigation

The reviewed models build on transformer encoders, image–text alignment and visual instruction tuning (Vaswani et al. 2017; Dosovitskiy et al. 2021; Radford et al. 2021; Li et al. 2023; Liu et al. 2023a). Open-weight families (Bai et al. 2023; Wang et al. 2024c; Chen et al. 2024b), alternatives to proprietary models such as GPT-4 and Gemini (OpenAI 2023; Gemini Team 2023), allow adaptation without guaranteeing onboard feasibility: low-rank adaptation (LoRA) tunes adapters while retaining the backbone at inference (Hu et al. 2022), preference optimization changes trained behavior (Rafailov et al. 2023), and distillation transfers supervision to a student, VLM or otherwise (Hinton et al. 2015). Segmentation and open-vocabulary detection supply regions without generating language (Kirillov et al. 2023; Liu et al. 2025a), and spatial benchmarks such as BLINK and EmbSpatial-Bench expose task-specific spatial limitations (Fu et al. 2025; Du et al. 2024); these limitations argue for separating geometric grounding from semantic judgment.

Adjacent navigation literatures offer architectural precedents, and their tasks vary in how far they involve people. Standard vision-and-language navigation (VLN) benchmarks evaluate instruction following in static scanned environments without moving pedestrians (Anderson et al. 2018b; Krantz et al. 2020), and VLN architectures range from cross-modal transformers and graph reasoning to language-model planning (Gu et al. 2022; Chen et al. 2022; Zhou et al. 2024). General-purpose visual navigation policies (Shah et al. 2023c,d; Sridhar et al. 2024) serve as comparison baselines in Sections 3.3 and 3.7; language-conditioned landmark planning and frontier exploration use language-model priors over where to go (Shah et al. 2023b,a); and SayCan, RT-2 and OpenVLA (Ichter et al. 2023; Zitkovich et al. 2023; Kim et al. 2025) precede the direct action policies of Section 3.6. The central cases of this survey (Section 2.4)

are distinguished by task rather than model family: in them, people's behavior, roles, relations or expressed cues shape shared-space motion.

2.4 Related surveys and scope of this survey

Prior surveys give this survey its social grounding and evaluation vocabulary, and the most recent add foundation-model and VLM taxonomies (Table 2). A 2026 preprint reviewing 25 years of human-aware navigation also lists the increasing adoption of foundation models among its emerging trends (Kashmar et al. 2026). What remains missing is a survey that organizes recent systems by where the model enters the navigation stack and asks what their reported comparisons show about online inference.

Kruse et al. (2013) classify methods by objective and by the navigation module modified, a functional-position precedent for our architectural patterns. Firoozi et al. (2025) already identify real-time inference as a core challenge for foundation-model robotics, and Pan et al. (2026) include social-norm compliance, yet motion conventions near strangers sit outside the center of Pan et al.'s taxonomy.

Two recent reviews overlap more closely. Alyassi et al. (2025) review and benchmark learning-based navigation, discuss VLM roles and distilled representations, and relate global planning, semantic encoding and social context to inference cost. They cite SocialNav-SUB without a dedicated comparison of social-scene VQA benchmarks, time only non-VLM planners on specified hardware and classify whole systems, leaving teacher absence at deployment implicit. In the closest review, Cai et al. (2026) organize VLM social navigation around high-level reasoning, low-level planning and control, and their intermediate mechanisms. Their roadmap covers five paradigms (semantic reasoning, evaluators, grounding, representations and action interfaces), followed by efficient deployment, and annotates each paradigm with its level of supporting evidence.

Cai et al.'s review shares our attention to temporal mismatch, asynchronous reuse, spatial grounding, distillation and lightweight adaptation. The questions differ: Cai et al. map how VLM reasoning connects to deployable navigation, whereas we examine what the reported experiments justify about retaining online inference, replacing it or transferring its benefits before deployment. Our analytical unit is the reported comparison: which input, module or training signal changes, which conditions are retained or unreported, and what the endpoint supports. Sections 3 and 5.3 apply this unit, linking each contrast to the benefit it shows and the deployment choice it leaves unresolved. This contrast-level reading complements Cai et al.'s roadmap and evidence annotations.

Our central scope is navigation in which people are treated as social agents and a vision-language or large language model contributes during inference or training. The corpus was assembled from searches of Google Scholar, IEEE Xplore, arXiv and OpenAlex, supplemented by backward citation checks; arXiv and OpenAlex were queried for records dated from January 2023 to September 2026 whose title or abstract combines people, navigation and language-model terms. A study enters as a central case when three conditions hold: a person's behavior, role, relation or expressed cue could change a robot's or embodied agent's navigation decision; a VLM, LLM or VLA

Table 2 Prior surveys and the position of the present work. The organizing-axis, coverage and scope-difference columns are this survey's assessment of each source. "VLM coverage" is *central* when vision-language methods receive dedicated structural treatment, *partial* for a section or subsection and *none* for no substantive treatment. Möller et al.'s embodied question answering (EQA) coverage concerns earlier vision-language approaches. VLM denotes vision-language model; VQA denotes visual question answering

Survey	Year	Organizing axis	VLM coverage	Scope / analysis difference
Kruse et al. (2013)	2013	Comfort, naturalness, sociability	None	Classical navigation focus
Rios-Martinez et al. (2015)	2015	Proxemics and social spaces	None	Proxemic emphasis
Möller et al. (2021)	2021	Perception and interaction	Partial (EQA)	No foundation-model taxonomy
Gao and Huang (2022)	2022	Evaluation practice	None	Evaluation-focused
Mavrogiannis et al. (2023)	2023	Planning / behavior / evaluation	None	Challenge-based organization
Mirsky et al. (2024)	2024	Conflict-avoidance components	None	Narrow subproblem
Singamaneni et al. (2024)	2024	Four facets incl. situation awareness	None	No foundation-model taxonomy
Francis et al. (2025)	2025	Eight evaluation principles	None	Evaluation-focused
Alyassi et al. (2025)	2025	Learning-based methods, benchmarked	Partial	No VQA comparison; no deployment taxonomy
Firoozi et al. (2025)	2025	Foundation models across robotics	Central	Robotics-wide scope
Cai et al. (2026)	2026	Five paradigms and supporting evidence	Central	Deployment roadmap
Pan et al. (2026)	2026	Perception / planning / control / interaction / coordination	Central	Not centered on stranger-motion norms
This survey	2026	Integration patterns and experiment-level evidence	Central	Online use; replacement tests

contributes during training or deployment; and a quantitative navigation or answer-level evaluation allows the model's role, output interface and endpoint to be coded. Datasets, benchmarks, simulators and scenario generators enter as evaluation resources rather than cases (Section 4); GROVE, a scenario generator retained as a mechanism extension, is the one exception. MUSON (Liu et al. 2025d) and the Social-LLaVA paper, whose authors present its fine-tuned model as a proof of concept for the SNEI dataset (Payandeh et al. 2025), are treated as resources, and their model results inform Section 3.6.3. Beyond Text and TrustNavGPT (Sun et al. 2024a,b) are coded together as one case, as are the two reports of Liu et al. (2025b,c).

The resulting 41 cases comprise 34 central cases and seven mechanism extensions, in three scope strata marked N, D and X in Tables 3–4. Of the central cases,

25 report a quantitative robot-navigation endpoint in simulation or on hardware (N), and nine evaluate answers or offline selections without one (D): AutoSpatial, SocialNav-MoE, E-SocialNav, group competitive learning (GCL), SOPD-SocialNav, MAction-SocialNav, Probing Prompt Design, MOSU and Ruo et al. Central membership marks social scope, not validated robot deployment. The seven extensions (X) illustrate adjacent mechanisms whose evidence transfers only indirectly to social robot deployment: Dual-process and SafeVLA (non-social mechanisms), CrowdVLA (a pedestrian policy), GROVE (language-specified scenario generation), HumAIN (a non-language teacher, an explicit exception to the language-model criterion), SocialNav (semantic traversability and region compliance) and CATNAV, which includes a pedestrian-crossing experiment but primarily evaluates traversability, instruction adherence and caching. Supporting literature of any publication year is cited for context.

Because the corpus was assembled through title and abstract terms, studies indexed elsewhere or described in other vocabulary may be missing; the extensions were selected purposively. Counts describe this case set alone, and the primary-pattern counts depend on the assignment convention of Section 3.1. Preprints and version differences limit publication-level certainty; shared datasets, backbones and implementations reduce independence; and because every case was selected for its use of a reviewed model or, for HumAIN, a non-language teacher, the VLM versus non-VLM comparisons in this set are not representative. These limits do not cancel benefits reported within individual studies, but they preclude a general estimate of VLM advantage or dispensability.

3 Model integration in social navigation

This section surveys the cases by where the model enters the navigation stack. Six patterns distinguish what a model supplies, how its output is coordinated with control and whether it remains in use after training (Section 3.1). Each pattern subsection (Sections 3.2–3.7) opens with what its cases show together and then describes representative systems and the comparisons that remove, replace or vary a component. Tables 3–4 give each case's details, and Section 5.3 synthesizes the comparisons across roles.

3.1 A taxonomy of integration patterns

The patterns sit at different levels (Figure 2). Scene interpreters (I), cost modulators (II), subgoal or path proposers (III) and direct action policies (V) are defined by the output consumed downstream; hierarchical coordination (IV) by how components are coordinated, often across timescales; and offline transfer or generation (VI) by the reviewed model's lifecycle, since the model is absent from the deployed loop. Because the patterns can overlap within one system, each case receives one primary assignment for counting, under an ordered convention. VI applies whenever the reviewed model is absent at deployment or simulator execution. Among online cases, IV applies when inter-component coordination or online adaptation is the principal architectural contribution, and otherwise the downstream output decides I, II, III or V. The convention

is an explicit analytical judgment, and its boundary cases are noted where they arise. VLM-Social-Nav both modulates a cost term and gates its queries on detector output; it receives II because its principal contribution is the social cost term, with gating secondary, whereas delayed-advice fusion is Slow Brain's principal contribution, which makes it IV despite its candidate-scoring interface. CoNVOI classifies context and then selects a reference path; it receives III because the downstream controller consumes the selected path. Seven answer-level generators receive V because their output is an action, even though they have no closed-loop navigation endpoint.

Each case is also coded by scope stratum (Section 2.4), backbone, adaptation, spatial scaffold, evaluation setting, comparator, reported result and timing, and C01–C41 connect the design entries of Table 3 to the results of Table 4. A *spatial scaffold* is an explicit mechanism that supplies the reviewed VLM or LLM (or, for the HumAIN extension, its non-language teacher) with geometric structure it would otherwise have to infer; Section 3.8 gives examples and the coding rule.

How often a model is queried is recorded separately, because one role can operate at different rates and the pattern labels fix neither the query rate nor the model's per-request latency. Over an explicitly reported observation window of duration $T > 0$, count issued requests n_{VLM} to a specified set of reviewed VLM, LLM or VLA components, and count control decisions n_{control}, with observed command rate $f_c = n_{\mathrm{control}}/T$. The *invocation ratio*

$$b = \frac{n_{\mathrm{VLM}}}{n_{\mathrm{control}}} \tag{1}$$

is the mean number of requests issued per control decision, including requests still pending at the window's end. No reviewed study reports this proposed descriptor. It requires $n_{\mathrm{control}} > 0$, so b is undefined for question-answering and standalone-generation systems, whether or not they query a model at runtime. Because raising the command rate lowers b at an unchanged request rate, comparisons must retain the request rate and per-request cost alongside b. Section 5.4 relates b to request capacity and timing.

3.2 Models as scene interpreters

A scene interpreter consumes observations and emits a symbolic or linguistic description of the social situation, such as the context type, who is grouped with whom or what a person appears to be doing, which downstream modules turn into action. All four cases report navigation endpoints, and in each any benefit depends on how the downstream module uses the interpreter's output: GSON's planner detours around groups, SAATT Nav's rule-based planner adjusts speed according to predicted intent, HCSG's trained policy fuses descriptions of people with geometry, and LLM-HNM's optimizer and policy act on users' diagnosed conflicts. The last two are coded I although a learned or optimizing module consumes the description. The four cases' navigation-level comparisons change the model along with other components or remove it with small effects. Replacing an interpreter therefore requires preserving its interface and validating both output quality and downstream behavior.

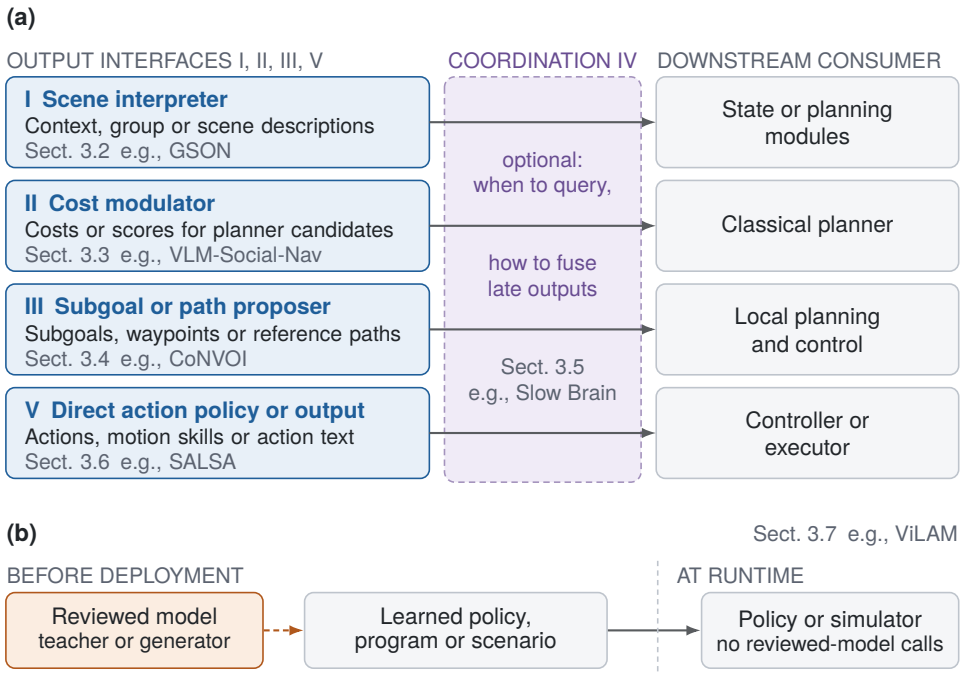


Fig. 2 Where and when models contribute to navigation. (a) At runtime, four output interfaces (I, II, III and V) connect model roles to representative downstream consumers; the rows are not successive stages. Hierarchical coordination (IV), the band crossed by every online output, can accompany any of these interfaces. Model inputs are omitted. (b) Offline transfer or generation (VI) prepares, before deployment (dashed arrow), a policy, program or scenario that later runs without reviewed-model calls. Each pattern names one central navigation case as an example; Table 3 lists all 41 cases. The six patterns can overlap

GSON (Luo et al. 2025) asks a large multimodal model only which pedestrians belong to the same social group, showing it an image annotated with each tracked person's box, mask and identifier. A mid-level planner then replans the stretch of the global path that crosses an ellipse fitted to each group. On 100 annotated scenarios, GPT-4o yields 77.8% exact group matches, against 10.8–14.8% for coordinate-only baselines using density-based spatial clustering of applications with noise (DBSCAN). In four real-world group scenarios run three times each, GSON has the lowest plotted geometric disturbance durations for individuals and groups and the highest mean minimum group distance, at somewhat higher jerk and runtime than several baselines. A timed elastic band planner given the same group estimates performs worse, which the authors attribute to short-horizon local minima when members are widely spaced.

The authors report degraded responsiveness in dense crowds with a 32B open-weight model taking 1.5–2 s per query, propose distillation into a 3B model (on compression, see Section 6.1) and note incomplete groupings under egocentric occlusion. The binary group label also makes the planner avoid every group equally, although crossing a queue may be more acceptable than interrupting a conversation. A wrong label can propagate into a planner that lacks confidence or consistency checks.

In SAATT Nav (Zhang et al. 2026d), an unnamed LLM labels each visible pedestrian near a powered wheelchair as yielding, rushing or keeping a constant velocity. The LLM generates up to eight intent hypotheses per call, a non-language rollout scorer selects one, and a rule-based planner adjusts speed according to the selected intent and, while the language module is active, detours around groups of stationary pedestrians detected by a geometric rule. The model is called every 30 simulation steps of 0.1 s or when a pedestrian comes within 3.0 m, with calls at least 15 steps apart. In three simulated scenarios with 30 seeds per method (360 runs in all), SAATT Nav ranks first on collisions and minimum pedestrian distance in two scenarios and last on travel time in the same two (results are reported as rank orders). Its ablation disables the LLM along with the intent-based speed modulation and the stationary-group detour, so it evaluates the language model and the rules it triggers together.

HCSG (Xu et al. 2026) adds reasoning about people to waypoint-based VLN. At each waypoint the robot pauses while new waypoints are predicted; if a detector finds a person, Qwen3-VL-2B-Instruct describes the person's action and apparent intention from a cropped image, and the CLIP-encoded description is fused with geometric features in a topological map from which a trained graph transformer selects the next waypoint. On the unseen split of HA-VLNCE, the continuous-environment benchmark of the human-aware VLN line (Dong et al. 2026), success is 0.24 against 0.21 for BEVBert and the share of episodes with a human collision 0.36 against 0.55, while the total collision rate is 5.02 against 4.71; in 50 real-robot trials HCSG succeeds 41 times against 19 for BEVBert. Adding the VLM stream lowers the share of episodes with a human collision on the unseen split from 0.40 to 0.36 and raises success from 0.23 to 0.24, so most of the difference from BEVBert remains without the VLM.

LLM-HNM (Li 2025) mediates the conflicting navigation preferences of three users. GPT-4o parses each user's statement into weights for speed, proxemics (interpersonal distance) and privacy, and in the full mediator a language-model classifier also labels the conflict type, power balance and basic needs. These labels select a fairness rule that a non-language optimizer applies before a preference-conditioned policy trained with proximal policy optimization (PPO) executes the compromise in a grid world with fixed human agents. In a within-subjects study with 18 participants in six triads, LLM-HNM averages about 0.5 social-zone violations per run, against about 1.7 for a Pareto-optimal mediator and 4.5 for simple averaging, with slightly longer paths and travel times. Since the compared mediators differ in several components, the result characterizes the full mediation pipeline.

3.3 Models as cost modulators

A cost modulator scores or re-weights candidates that a classical planner has already generated, or produces a cost layer that a classical planner consumes. With feasibility and collision checks left downstream, system safety and its comparison across studies depend on the implemented checks and assumptions. Six cases have this primary pattern, four with navigation endpoints and two evaluated offline or at the answer level. Their clearest advantages over baselines appear where the deciding cue is semantic, such as a raised hand or a pedestrian to yield to, and G2-Nav's ablation supports the model-derived social term within one pipeline.

VLM-Social-Nav (Song et al. 2025b) adds a social term to the weighted goal and obstacle costs of a dynamic-window planner, and only this term uses VLM output: it penalizes the weighted deviation of each candidate linear/angular velocity pair from a target pair. Given the current image, the robot's current action and a short list of etiquette rules, GPT-4V answers with a direction from {left, straight, right} and a speed change from {slow down, speed up, constant, stop}, which is parsed into the target pair. The answer is a point in the robot's own action space, expressed in words and limited to twelve combinations.

In four indoor scenarios, with 21 trials per method per scenario, VLM-Social-Nav is compared with a DWA baseline and behavior cloning trained on SCAND. In the stopping-gesture scenario, both baselines reach zero success and VLM-Social-Nav 100%, a result associated with its gesture-sensitive semantic input. Averaged over the four scenarios, success is 27.38 percentage points higher and the collision rate 19.05 points lower than with DWA. In the plain frontal approach and the narrow doorway, VLM-Social-Nav and DWA both reach 100% success, and VLM-Social-Nav collides less often in each. A user study with seven participants per scenario gives the method the highest mean score in every scenario. Outdoors, BehAV (Weerakoon et al. 2025) turns free-form instructions into spatial behavioral costs through LLM-assigned weights and CLIPSeg segmentation, and in its stopping-gesture and yielding scenarios the GA-Nav, ViNT and NoMaD baselines record zero success, as both baselines do in VLM-Social-Nav's gesture condition.

G2-Nav (Liao et al. 2026) tests the contribution of a model-derived cost term. Its VLM returns a bounded social score per relevant tracked object, which enters the costmap as a scaled elliptical Gaussian beside goal, occupancy and traversability terms. Newly detected objects carry a default score until the VLM responds, which decouples VLM response latency from the immediate control loop; sustained queries still require sufficient inference capacity. On recorded SCAND trials, replacing the VLM's per-object scores with the default score raises L2 error from 0.55 to 1.12 m and maximum average orientation error from 15.64 to 21.08 degrees.

Obstacles to Etiquette (Fang et al. 2026) fine-tunes a Qwen2.5-VL-7B path selector on GPT-4o-generated choices and rationales from SCAND and rematches each returned choice to the current candidates before reactive control, so it evaluates planner-generated candidates (pattern II), although its selected-path reference overlaps with pattern III. A standalone test of its candidate generator shows that lengthening the prediction horizon from 0 to 2.4 s reduces candidate-path group intersection from 39.17% to 8.89%, which highlights the role of upstream candidate construction; the VLM's closed-loop contribution is not reported separately.

Two further cases stop short of a navigation endpoint. MOSU (Liang et al. 2025) draws numbered candidate trajectories on the camera image and asks Gemini to rank them under rules that include keeping away from pedestrian groups and slowing when people approach; the ranking enters a weighted score, a speed mode caps the local planner, and the quantitative evaluation is offline on the Global Navigation Dataset (GND) against six baselines, including PIVOT. Ruo et al. (2025) let an LLM set a navigation mode, maximum speed and social distance from a user's stated age, mobility and comfort with robots, settings that a control-barrier-function controller

then enforces; the returned settings are judged for correctness over 100 text requests, and the navigation runs are demonstrations. Table 4 lists the headline results of these and the other systems.

The cost interface bounds what the model can request: VLM-Social-Nav's direction/speed vocabulary cannot directly request behaviors such as joining a queue or communicating intent, whereas a spatial costmap can encode richer route preferences, and behaviors outside the planner's action and goal space require another interface. Retaining a planner keeps geometric planning apart from semantic scoring but does not by itself impose hard safety constraints: BehAV uses unconstrained MPC with obstacle costs from light detection and ranging (LiDAR), and G2-Nav uses costmap-based path optimization (Section 6.3). The baseline comparisons in this group involve complete systems that differ in inputs, observations or training, and the safety measures of VLM-Social-Nav and G2-Nav come from supervised trials: preventive manual interventions count as collisions in VLM-Social-Nav, whereas G2-Nav reports supervisor pauses as separate intervention counts and excludes them from its safety-violation and goal times.

3.4 Models as subgoal and path proposers

A subgoal or path proposer emits reference paths, waypoints or navigation subgoals, often at a lower rate than the control loop, and local planning and control convert these into motion. In both pattern III cases, CoNVOI and CATNAV, classical geometry constructs a few marked candidates, and the model's selection becomes the path that a controller then follows. Their comparisons show two ways to reduce what the model must do: filtering candidates against occupancy reduced CoNVOI's unacceptable reference paths and inference times, and caching scene-level costs lowered CATNAV's query rate.

In CoNVOI (Sathyamoorthy et al. 2024), a local CLIP model classifies the context zero-shot and selects a predefined behavior phrase such as "stay on the pavement." A second, large VLM (GPT-4V or Gemini) then selects, in line with that phrase, among obstacle-free image regions numbered by correlating the red–green–blue (RGB) image with a local occupancy map, and the selected numbers form a reference path. Near the end of a path in an unchanged context, an extrapolated waypoint, accepted only after obstacle and paved-surface checks, avoids some large-VLM queries.

The input ablation tests occupancy filtering directly: both variants keep the numbering, and the ablated one also permits markers on obstacles. For GPT-4V in the first indoor scenario, removing the filter raises reference-path error $0.187 \rightarrow 1.576$ and inference time, from query to returned path including network latency, $3.48 \rightarrow 4.86$ s. With both VLMs in all four scenarios, the filter lowers the share of unacceptable paths and the inference time, and it lowers path error except for Gemini in the third, outdoor scenario.

CATNAV (Potnis et al. 2026), an extension that primarily targets quadruped traversability and instruction adherence with one pedestrian-crossing task (Section 2.4), applies the same division of labor. Gemini 3 Flash supplies class-level traversal costs that CLIPSeg and depth projection convert into a costmap, a visuosemantic cache reuses risk estimates for similar CLIP-embedded scenes, and a further

model query selects among four geometrically generated paths overlaid on the RGB image. The selected path goes to a tracking controller, which places CATNAV in pattern III. Raising the cache's novelty threshold from 0.1, a setting the authors describe as effectively disabling caching, to 0.55 lowers the query rate from 0.042 to 0.006 queries per second and scene queries from 125 to 9; total calls fall from 420 to 166, including path-selector calls from 295 to 157, so the comparison supports reduced online use while every configuration retains a runtime language model.

A bare waypoint conveys an intended route, while explicit speed, timing and yielding instructions require additional information or downstream inference. Waypoints and costs can both carry errors forward: a persistently tracked wrong waypoint can direct the robot toward an unsuitable region, and a heavily weighted erroneous cost can dominate candidate selection. How far such errors propagate depends on planning horizon, replanning cadence, cost weights and safety constraints. CoNVOI's filter ablation and CATNAV's cache comparison each vary one component within one system; neither measures this propagation.

3.5 Hierarchical coordination of models and planners

In hierarchical coordination, distinct reasoning, planning or adaptation components coordinate invocation, information flow or action fusion. The nine pattern IV cases do so, often across different update timescales, in four ways: conditional invocation (HUMA, Dual-process, PAD), gating of a low-level policy by a high-level model (Walk With Me), fusion of delayed or competing outputs with current plans or actions (Slow Brain, TIC-VLA, LISN and, through adaptive action fusion, SALM) and online adaptation of a lightweight model (OLiVia-Nav). Invocation may be state-dependent or fixed-cadence, with or without a gate, and the pattern can accompany any output interface; Tables 3–4 give case details. Most timed VLM outputs take seconds, although the timed operations differ, and SALM, Walk With Me and PAD report no timing. Timing and its handling changed outcomes: success fell under HUMA's *Realistic* timing condition relative to a frozen scene, and Slow Brain's advice lowered takeovers when matched to or fused with current candidates but raised them when stale trajectories were executed directly. The ratio b summarizes calls and leaves the invocation rule unspecified.

HUMA (Ahmadi et al. 2026), whose authors call semantic reasoning unnecessary overhead in uncrowded environments and critical near humans, pairs a reinforcement-learning policy for ordinary navigation with a LoRA-adapted Qwen3-VL-2B that intervenes only when Personal Space Compliance over recent proximity readings falls below a tolerance. Simulated success on Social-HM3D, a social-navigation benchmark built on Habitat-Matterport 3D (Gong et al. 2025), is 62.07% with the scene frozen during VLM inference and 56.90% under the paper's *Realistic* timing. A trigger ablation spans 390–3,250 VLM calls, with the highest observed success at an intermediate tested threshold. Without an episode or time denominator, these counts yield neither b nor a general optimum.

The trigger can also be structural. Dual-process (Lee et al. 2026), a non-social extension, targets VLM steering decisions that repeat across many consecutive frames in straight hallways. After keyframe comparison detects a structural change, a

lightweight System One matches scene descriptors against a condition-to-action table the VLM generated at initialization, and System Two is invoked when matching fails to yield a unique action. Across five real buildings, the system completes 67.5% of episodes, against 5.83% for continuous VLM-only control. System One resolves more than half the decisions, taking 0.7–0.9 s each, against 17–19 s per System Two call on the same onboard hardware. The deliberately asymmetric gate escalates every evaluated case that needs the VLM and about a third of unneeded ones, and a less sensitive threshold misses a turn. Similar structural triggers might cut redundant queries in social encounters, but this case does not test that.

In Walk With Me (Zhang et al. 2026a), the gate decides whether another component may act: a high-level VLM chooses a destination, labels each step routine or complex and answers proceed or stop-and-wait, and a low-level navigation policy plans local motion only after a proceed answer, which at crossings and in crowds must also exceed a confidence threshold. Twelve of 20 real trials of last-mile delivery and guidance for blind users succeed. On the two delivery scenarios, where the full system reaches 70% success, swapping the high-level model gives 3 to 6 successes in 10 trials and swapping the low-level policy gives 2 to 6, so the outcome depends on both components.

PAD (Bejarano Sepulveda et al. 2026) fuses simultaneous localization and mapping (SLAM), speech recognition and a slower, unnamed VLM stream into a situation map, and a meta-controller hands control to an LLM dialogue module when a person keeps blocking a corridor, or otherwise to a planner. With eight participants and one trial per condition on a TIAGo++ robot, the SLAM-only condition, which drops every model component, succeeds in 2 of 8 trials and both dialogue strategies in 8 of 8, while the more assertive strategy lengthens trials to 162.21 s from 99.95 s for the conservative one.

Among fusion designs, Slow Brain (Peng et al. 2026b) targets a *trajectory scoring gap*: a learned sidewalk planner proposes good candidates in real time but often fails to score the best one highest, sending the robot onto grass or toward pedestrians. A VLM selects an index from the proposal set while the planner keeps generating. Offline, a Gemini 3 Flash selector lowers average displacement error (ADE) on the hard subset (1.16 versus 1.64 m) and raises it on routine scenes (0.60 versus 0.502 m). The deployed system issues Gemini 2.5 Flash-Lite requests at 1 Hz with typical cloud round trips of 1.5–3.0 s, while candidates and waypoints update at 5 Hz. A training-free fusion layer motion-compensates the stale selected path and compares its remaining horizon with current candidates, discounting its influence with age. On five real-world routes with repeated supervised runs, takeovers per 100 m are 3.49 for local-only control, 4.40 for streaming without fusion, 1.74 and 2.18 when the stale choice is matched to the closest current candidate (with sequential and streamed requests, respectively), and 0.87 for probability fusion with streaming. Because operators restore a safe pose and resume the route after each intervention, the rates measure takeovers in supervised runs, a different endpoint from unaided route completion. In this urban-sidewalk task, which includes pedestrian encounters but extends beyond social navigation, fusion was useful, although it does not remove the limits set by request capacity and output age (Section 5.4).

TIC-VLA (Huang et al. 2026b) addresses a related timing problem with learned fusion, conditioning a faster action expert on delayed semantic states from the key–value cache of an InternVL3-1B backbone, along with current visual features, robot motion and latency metadata. Enabling reasoning-token generation in the same fine-tuned model raises simulated success from 40.00% to 55.29%, while the collision rate rises from 25.88% to 28.24% and the VLM forward rate falls from 4 to 0.5 Hz, with the action policy at 10 Hz. The authors attribute the lower collision rate without reasoning to reduced activity and more frequent failure. Both variants keep the VLM. On hardware, the authors report action/reasoning latencies including 85.73/3,430.73 ms on an RTX 4060.

In LISN, Chen et al. (2026a) separate fast and slow loops for language-instructed navigation: the VLM selects tool calls and controller parameters, and a fast LiDAR-based controller updates a tool-fed costmap between VLM calls. The five simulated task-level success rates average 91.3%; the authors reimplement one baseline and migrate the other, running both with GPT-4o. For the Follow Doctor task, LISN succeeds in 100% of trials and both baselines in 0%. Using example trajectories, the authors trace this gap to delayed VLM responses that make the baselines lose the moving target, whereas LISN's fast map updates keep track of it; the compared systems also differ in perception tools, tracking and controllers.

SALM (Wang et al. 2024d) fuses at the action level. GPT-4o issues high-level instructions and velocity proposals and also sets feedback-based weights for competing language-model and reinforcement-learning actions. Positions and velocities enter as text, which serves as a runtime spatial scaffold (Section 3.8). Over 50 simulated trials per task, mean success is 86% for SALM, 75% without the language navigation and feedback branches and 63% with fixed 0.5/0.5 fusion weights. The fixed-weight contrast favors adaptive weighting; the branch ablation removes both branches together, and all variants keep the high-level language interface.

OLiVia-Nav (Narasimhan et al. 2025) sits at the boundary with offline transfer (VI), where its initial construction belongs: captions from a large VLM align lightweight image and text encoders whose social-context embeddings condition trajectory generation and selection. Its lifelong-learning updater places it in pattern IV and responds to the offline-adaptation limitation discussed in Section 3.7: the large VLM captions images collected during navigation, and both encoders are updated in batches, which keeps inference lightweight but adds teacher and optimization cost. On a held-out test set from the MuSoHu human-navigation dataset (Nguyen et al. 2023) and before any update, removing the social-context image embedding raises the mean squared error (MSE)/Hausdorff losses from 0.2981/1.8087 to 0.4412/2.3772, removing the text embedding also raises them, and one update lowers the full model's losses, leaving open whether continued teacher queries are needed.

For gated members, the trigger trades missed assistance against reintroduced latency, and the reviewed triggers, generally geometric or perceptual predicates such as proximity, scene change or detector output, may miss socially salient but geometrically mild situations; no reviewed evaluation tests this risk. Fixed-cadence and online-adaptive members face stale-output fusion and update stability, respectively, as their central concerns.

3.6 Models as direct action policies and outputs

A direct action policy maps observations directly to actions, motion skills or action sentences, as in deployed or simulated VLA policies, systems that execute generated action plans through tool calls, and answer-level next-action generators. Pattern V cases supply two kinds of evidence. Policies evaluated in simulation or deployment (HSAC-LLM, TrustNavGPT, SALSA, SafeVLA, CrowdVLA and SocialNav) show that targeted components, such as spoken interaction, vocal cues, safety constraints and counterfactual supervision, improve reported outcomes, mostly on adjacent tasks, simplified simulators or offline data, whereas the policy of Liu et al. (2025c) outperforms CrowdNav++, a reinforcement-learning crowd-navigation policy (Liu et al. 2023b), in success and collisions only when that baseline is held to the language model's slower decision rate. The answer-level generators of Section 3.6.3 make up the second kind of evidence. A direct policy that replaces a classical planner needs its safety mechanisms and their assumptions specified and evaluated, since a safety guarantee requires more than training data.

HSAC-LLM (Wen et al. 2025) lets the robot negotiate passage with pedestrians by speech. A hybrid soft actor-critic policy chooses velocities and an interaction code, GPT-4 turns the code into an utterance to the pedestrian, and an unnamed language network encodes the pedestrian's reply for the policy. In a two-dimensional simulator whose scripted pedestrians respond to the codes, success over 500 tests per scene is 0.91, 0.93 and 0.91 in three scenes, against 0.85, 0.77 and 0.85 for a twin delayed deep deterministic policy gradient (TD3) baseline. A one-way variant trained without the LLM module has lower success in all three scenes, most in the first (0.86 against 0.91). Real-robot runs are presented qualitatively.

TrustNavGPT (Sun et al. 2024b) and its companion report Beyond Text (Sun et al. 2024a) let an LLM decide whether to follow a person's spoken directions or ask someone else, using vocal cues of uncertainty. The system selects among five candidate plans generated by few-shot prompting and converts action/location pairs into tool calls. On 500 recorded instructions, preferred-option selection rises from 22.16% for text alone to 49.30% with chain-of-thought reasoning and 70.46% with audio cues added; adding all three vocal cues without reasoning gives 36.93%. Beyond Text reports gains from reasoning and vocal cues for two other selection models, starting from text-only baselines that differ widely (57.88% for text-davinci-003 and 43.11% for Gemini 1.5 Pro). Target finding reaches 80% in ten RoboTHOR scenes, against 25% for LM-Nav.

Liu et al. (2025c) fine-tune a language model to issue robot motion commands from fused LiDAR and camera input. Pedestrians reach the model as bird's-eye-view tensors with numeric positions and velocities, or as tracklets labeled with categories such as child, elderly person or wheelchair user. The first report trains an input encoder and then applies LoRA fine-tuning; the companion report (Liu et al. 2025b) describes chain-of-thought distillation from a larger teacher, with DeepSeek-R1 named as an example, into a 7B student and gives the closed-loop comparison. In simulation with one decision per second, success is 95% with 1% collisions for the student, 94% and 0% for DeepSeek-R1 and 70% and 4% for CrowdNav++; at its native 0.25 s interval CrowdNav++ reaches 100% and 0%, although its average time to success is 14.19 s against 7.95 s for the student.

3.6.1 Aligning actions with internal representations

Wang et al. (2026a) pair linear probes with behavioral counterfactual tests in an OmniVLA navigation policy (Hirose et al. 2026), supporting a distinction between decodable information and its behavioral use. Person–object separation peaks around layer nine, impending danger is decodable at roughly 74% across depth, and both signals remain decodable at the action-head input. Yet with static objects inpainted at pedestrian locations, the base policy's counterfactual accuracy (CF-Acc) is 52.8% and its differential clearance 0.002 m. The authors report greater avoidance of a more visually salient pedestrian but no systematic person–object differentiation across the tested behavioral measures.

Their method, SALSA, adds intermediate-feature access to the action head, counterfactual supervision and three-tier future-risk targets, using human demonstrations and LiDAR-derived training targets without manual social labels. On 4,902 held-out SCAND frames, CF-Acc rises to 93.0% and near-collisions (predicted waypoints within 0.6 m of the ground-projected LiDAR scan) fall from 2,460 to 334. In small monocular-RGB Unitree trials, the aligned policy reaches the goal in five of five pedestrian-scenario runs, against one of five for the base policy. In a source ablation on a variant that lacks SALSA's weight-regularization and layer-freezing protections, removing the mid-layer projection during safety alignment lowers CF-Acc from 74% to 41%, but LoRA training also changes the representations, so how much of SALSA's CF-Acc gain comes from reaching existing features rather than altering them remains open (Section 5.3).

3.6.2 Safety, pedestrian and traversability policies

Three extensions test policy mechanisms outside the central social scope. SafeVLA (Zhang et al. 2025) casts safety alignment as a constrained Markov decision process solved by Lagrangian safe reinforcement learning and reports, averaged over three tasks, an 83.58% reduction in cumulative safety-violation cost and a 3.85% gain in task success over FLaRe, a task-focused reinforcement-learning baseline. None of its five safety-critical components (corners, blind spots, fragile collections, critical points and dangerous equipment) involves people. Two analyses bear on this headline result. SafeVLA's integrated safety approach (ISA) attains lower cost and higher success than the FLaRe reward-shaping variant with the same imitation-learning foundation, supporting the adaptive constrained recipe under these settings. Training ISA in simplified one-room scenes without safety-critical components raises cumulative cost from 1.854 to 5.01 and lowers success from 0.865 to 0.645, although scene complexity and hazard exposure change together. Apart from a dual-arm Safety-PickUp demonstration, the evidence is simulated, and transfer to social constraints remains to be shown.

Hwang et al. (2026) invert the roles in CrowdVLA, applying a LoRA-fine-tuned vision-language-action model to simulated *pedestrians* in reconstructed ETH (Pellegrini et al. 2009) and UCY (Lerner et al. 2007) scenes. The results concern pedestrian control, but two ablations are relevant to robot action design. The policy chooses among 64 *motion skills*, 20-frame trajectory segments clustered from real pedestrian data. Removing skills raises teacher-forced next-segment average displacement error

from 0.1288 to 0.5876 on Zara01 and from 0.2088 to 0.5978 on ETH-Hotel, with per-decision resets to ground truth that exclude rollout drift. CrowdVLA also learns from question–answer pairs generated by counterfactual skill rollouts that record collisions, norm violations and inefficient detours. Without these question–answer pairs, goal success falls and the social-region violation rate rises in both of the paper's synthetic scenarios, Hallway and Intersection (from 0.0170 to 0.1369 in Intersection).

SocialNav (Chen et al. 2026b), a semantic-traversability extension, conditions a waypoint-generating flow policy on VLM features, giving it primary pattern V despite its hierarchical architecture. An ablation that removes the social reward while keeping the runtime architecture reduces region-compliance measures in the tested training design. The reported deployment rate of over 5 Hz refers to the full system.

3.6.3 Answer-level action generation

Seven central cases generate next-action answers at runtime and evaluate them without a closed-loop navigation endpoint, so b is undefined for their evaluations and their quantitative results stay at the answer level. Most build on SNEI, which comprises over 40,000 human-annotated visual question–answer pairs on 2,000 SCAND interactions. The paper that released it, treated here as an evaluation resource (Section 2.4), fine-tunes a proof-of-concept model, Social-LLaVA, that leads GPT-4V and Gemini 1.5 Pro on intention similarity (0.26 versus 0.21 and 0.22) but trails both on relative-position and overall similarity (Payandeh et al. 2025); its authors acknowledge ambiguity in social interpretation and ground-truth labels, so the scores can depend on annotator choices. Supervision, compression and prompting all shift these answer scores, each in ways that depend on the dataset and setup.

Supervision. Kong et al. (2025) fine-tune AutoSpatial, built on a LLaVA-1.6-OneVision-7B backbone, on structured spatial VQA with a prescribed vocabulary of image zones, metric distance bands and compass directions, combining rule-based labels for individual pedestrians with 72 manually annotated scenes. On an automatically labeled test split of CODA, the dataset from which AutoSpatial builds its automatic spatial VQA, the spatial score is 0.376 for Vanilla-LLaVA, 0.404 with manual annotations alone, 0.680 with automatic labels alone and 0.710 with both, while reasoning-only training reaches 0.518–0.520, below the full configuration; the authors also report VLM-judge score gains of up to 20.50% on the action component over training on manual annotations alone. Reasoning supervision does not help uniformly: Liu et al. (2025d) build MUSON from 10,110 egocentric MuSoHu samples under the same five-step chain-of-thought schema as SNEI, and chain-of-thought supervision of a small vision-language model raises action accuracy from 0.5499 to 0.6899 on MUSON but lowers it from 0.6235 to 0.5741 on SNEI. Their diagnosis, unreliable alignment between SNEI's reasoning and actions, motivates consistency checks between reasoning and action annotations.

Compression. Four systems compress the action model yet retain a specialized VLM at inference and therefore remain in pattern V. SocialNav-MoE (Kawabata et al. 2025) trains a 5.74B mixture-of-experts VLM and, timed only on the action-generation turn, reaches 1.709 frames per second (FPS), 8.1× and 19.6× the throughput of GPT-4o and Claude, although remotely queried and local models need not share hardware

or service conditions. E-SocialNav (Xiao et al. 2026) builds on a Phi-2-2.7B language model with a SigLIP encoder, tuning the projector by supervised fine-tuning and then a LoRA adapter by direct preference optimization, and reports 2.354 FPS and action accuracy of 0.550, against 0.483 for the fine-tuned Social-LLaVA, 0.450 for GPT-4o and 0.417 for Claude. The two papers report the same 265/60 train/test image counts and the same baseline throughputs for the two proprietary models, yet their GPT-4o Sentence Mover's Similarity scores diverge (0.376 and 0.651), so we interpret both as source-specific results. In GCL (Zhang et al. 2026b), a 3B Learner trained jointly with a 4B Guide reaches, at its best temperature setting, an SNEI Action-F1 (semantic similarity to reference action text) of 0.968, above its own supervised fine-tuning baseline (0.692), the fine-tuned 4B model (0.816) and a separately fine-tuned 8B baseline (0.755). SOPD-SocialNav (Zhang et al. 2026c) distills an 8B teacher into a 3B student that remains online, and reports 0.984 model FPS separately from 2.74 s end-to-end latency including communication. SNEI-derived cases also use different splits, 8:2 for SOPD-SocialNav and 265/60 images for SocialNav-MoE, which may give them different training and test membership.

Prompting. Adding a meta-cognitive prompt (MCP) to MAction-SocialNav, which generates a ranked set of acceptable actions (Wang et al. 2026b), raises all-predictions-in-ground-truth accuracy (APG) from 0.532 to 0.595. In Probing Prompt Design (Xiao and Yamasaki 2026), a reasoning-focused prompt with an added human-comparison framing raises GPT-4o action accuracy from 0.550 to 0.633 on SNEI and from 0.531 to 0.563 on a MUSON subset, relative to omitting the system prompt, while fine-tuned TinyLLaVA variants show model- and dataset-dependent effects. Table 4 keeps the evaluation setting beside each of these results.

3.7 Offline knowledge transfer and scenario generation

In offline transfer or generation, the reviewed teacher or generator annotates data, generates scenarios, supplies attention maps or provides language rationales before deployment, and it is absent from the deployed policy or simulator execution loop, where a smaller learned policy, a symbolic program or a generated scenario remains. A deployed control system then has $b = 0$ for those components, although other models in its stack may still be queried, and a specialized VLM or LLM still queried at runtime falls outside pattern VI, even with an offline teacher. The cases indicate that VLM-derived supervision, together with demonstration-derived data, can train students that navigate without teacher queries in the tested scenarios, that LLM-generated text can train a deployed module that runs without the LLM, and that a language model can generate executable pedestrian scenarios for simulation; in one case, demonstrations ranked by a language model seem to have reduced navigation success. During learning, the teacher's outputs shape the student's supervision, and, at runtime, computation remains for the student, a generated behavior tree and classical simulation components. Standalone scenario generation lacks a control-decision denominator, so its b is undefined.

3.7.1 Attention and feature distillation

In ViLAM, Elnoor et al. (2025) fine-tune a LoRA-adapted copy of the pretrained vision-action model VANP (Nazeri et al. 2024), aligning its attention with that of the frozen original and with spatial supervision built from VLM responses: GPT-4o rates three marked frontiers for crowding, and the response-derived scores form a social map. The student's maps then supply a traversability cost to a local planner. In four Husky scenarios, success is 100/90/90/80% against 80/60/60/70% for the best baseline in each scenario. Scenario 4 adds a Fréchet distance of 1.034 for ViLAM versus CoNVOI's 1.452. Because the GPT-4o annotator is used only offline, the deployed system has $b = 0$. The design trades online adaptivity for tractability. ViLAM is trained offline on outdoor SCAND trajectories only, so its social priors face a limited outdoor-to-indoor test in two indoor scenarios and, as the authors note, may need further tuning in crowded or unseen settings. Narasimhan et al. (2025) target the same limit in OLiVia-Nav by moving updates online (Section 3.5); sustained generalization across these shifts remains to be shown for both.

HumAIN (Song et al. 2026) is the non-language exception, a teacher–student design that does not test removing language from a VLM. Its transformer teacher fuses past images, skeletal keypoints extracted offline, robot state and the goal, and a lightweight student given RGB history and a goal learns trajectories against ground-truth targets while aligning its latent features with the frozen teacher's, so the deployed student skips explicit skeletal estimation. The reported 29.8% average relative reduction across four baselines and three offline out-of-distribution trajectory-error metrics reflects both teacher-feature transfer and goal modality, since HumAIN uses numerical goal coordinates whereas three of the four baselines (ViNT, NoMaD and GNM) use goal images. Removing skeletal input from the teacher changes results modestly, with the full model better on the displacement metrics and the pose-free variant marginally better on orientation error.

3.7.2 Language-derived supervision

In Narrate2Nav, Payandeh et al. (2026) use language-informed supervision to shape a visual representation. Qwen2-VL-72B-Instruct supplies textual annotations that, with visual observations and recorded trajectories, train an approximately 330M teacher that includes a CLIP text encoder. A Barlow Twins objective aligns an approximately 29M ResNet-50 student, which receives RGB history and a two-dimensional goal, with a teacher representation computed from future frames. Geometry stays in the pipeline, from waypoints overlaid on annotation images to the student's robot-frame goal, and only the annotation stage calls a language model, so the design tests offline language supervision on top of retained geometry.

The reported 52.94% offline gain is a relative reduction in average displacement error from 0.34 for NoMaD to 0.16, and the 41.67% real-world improvement compares aggregate success of 8.5/10 with 6/10 for GNM. Across four retained robot scenarios, with Navigation in Crowds excluded because all models performed poorly, Narrate2Nav ranges from a leading 10/10 in Intersection and Frontal Approach to a trailing 4/10 in Narrow Passageway.

Removing the text encoder while keeping distillation and the redundancy-reduction objective yields a reported 14.8% aggregate performance drop; average orientation, average displacement and final displacement errors are 0.04/0.16/0.24 with text and 0.06/0.19/0.24 without. The text-free variant still improves over the compared baselines. These results support a contribution from language supervision.

Read together, the HumAIN and Narrate2Nav ablations show that the gains cannot be attributed to the privileged pose or language signal separately from the transfer architecture that carries it.

CommNav (Sacco et al. 2026) uses a language model to produce training text. Qwen3-8B converts simulator records into sentences; the records state whether, when, where and in which direction a bystander saw the person the robot is looking for, and the sentences train a BERT-based module that estimates the person's location for a reinforcement-learning policy in Habitat 3.0c, the authors' communication-enabled extension of Habitat 3.0 (Puig et al. 2024). The robot's own stack contains no LLM, and Qwen3-8B also stands in for the bystanders in the language condition; the evaluation is simulated. Over 400 episodes, episode success is 0.24 with structured cue input, 0.20 with LLM-generated sentences and 0.23 with colloquial sentences from a human study (100 episodes), against 0.14 for a policy that can interact with people but has no communication module (0.16 when that policy is trained for as many steps as the policy fine-tuned with the structured module). The structured and LLM-generated inputs reach the same finding success (0.78); the structured module has the higher episode success, and the language-trained module finds the person in fewer steps (546 against 572). Because the structured module is pretrained on 2.4M records and the language module on 7,000 generated sentences, with different encoders, the comparison does not isolate the form of the input.

Kathuria (2025) uses GPT-4o-mini to rank demonstrations for inverse reinforcement learning in a simulated narrow crossing where the oncoming person has priority. Each demonstration is drawn as a top-down image showing the robot's and the person's past paths, the goal and the demonstrated path, and few-shot prompting with hand-scored examples yields a rating from one to five. The ratings enter a ranking loss in the reward learner, and the deployed policy runs without the model. Against a human expert's ratings of 180 demonstrations, exact agreement is 69.68% and agreement within one rating point 82.98%, with both label sets skewed toward the top rating. When the policy is trained with these ratings, its success falls to 76.66%, compared with about 92% for the same learner without ratings, and the author attributes the drop to mislabeled good demonstrations. The two figures come from separate experiments, so the comparison is indicative; within that limit, the model-generated labels lowered navigation success.

3.7.3 Language-driven scenario generation

Outside robot control, GROVE (Nguyen et al. 2026) turns a free-text prompt into a running pedestrian simulation: an LLM composes retrieved behavior-node specifications into a behavior tree, which the authors call a *contract* between the generation and simulation sides that grounds language-derived decisions in executable logic; a Theta* planner runs *on* that tree to inject waypoints between navigation goals; and

a social force model handles local avoidance. The authors added the global planner because omitting it produced frequent wall collisions and implausible straight-line paths; the planner secures only map-level feasibility, which does not extend to the simulator or to social behavior as a whole. They adopted retrieval because supplying the full node specification in-context increased hallucinated node types. In GPT-5-based judging of three scenarios per preset (64 screenshots in total), GROVE has the highest average score across the Emergency, Normal and Queuing presets and consistently leads the two reimplemented baselines on prompt alignment. Once the behavior trees are generated and geometrically post-processed, the trees, Theta* waypoints (velocity fields in the Emergency preset) and social-force components run without further language-model calls.

3.8 Design attributes across patterns

Across the six patterns, training regime and deployment role vary independently: preference training can leave a model online, and teacher distillation can either keep a student VLM online, as in SOPD-SocialNav, or remove teacher calls from execution, as in ViLAM. By arXiv identifier year, or publication year where no identifier exists, 8 cases fall in 2024, 13 in 2025 and 20 in 2026. Section 2.4 and Table 4 separate 34 central cases from seven extensions and, within the former, 25 navigation evaluations from nine answer-level or offline evaluations, and Figure 3 maps these strata against the primary patterns and timing reports. The map shows where the evidence is concentrated: 12 of the 25 navigation evaluations fall in patterns II and IV, pattern I contributes four, pattern V four against seven answer-level ones, pattern VI four and pattern III one.

Secondary attributes refine each design beyond its primary pattern. The *adaptation regime* in Table 3 distinguishes zero-shot or few-shot prompting, supervised fine-tuning, preference or reinforcement optimization, distillation or self-supervision, and online adaptation. The *backbone* entry records model family, size and specialized or distilled variant where reported, and it is only a partial guide to latency, which also depends on hardware, prompt length and decoding. The *evaluation* entry of Table 4 distinguishes real-robot deployment, simulation, offline and question-answering evaluation, and human-subject study, with a precedent in Mirsky et al. (2024), and gives the reported scale.

Spatial scaffolds, as defined in Section 3.1, recur during training and at runtime: CoNVOI numbers navigable regions against an occupancy map (Sathyamoorthy et al. 2024), ViLAM marks three training-image frontiers for crowding supervision (Elnoor et al. 2025), SALM passes positions and velocities as text, and the non-social Dual-process extension supplies segmentation-derived spatial descriptors in prompts (Lee et al. 2026). AutoSpatial uses its prescribed vocabulary in data construction and in the system prompt, and supplies bounding boxes and color-to-pedestrian correspondence only during training, which leaves open how much spatial competence the weights hold apart from the prompted vocabulary. Detector-based query gating alone, as in Song et al. (2025b), falls outside the definition: it controls when the VLM is queried but supplies no scene geometry. Scaffold coding in Table 3 covers training and inference, including offline annotation or generation prompts, and excludes geometry used only as a target, evaluation reference, model output or downstream controller input.

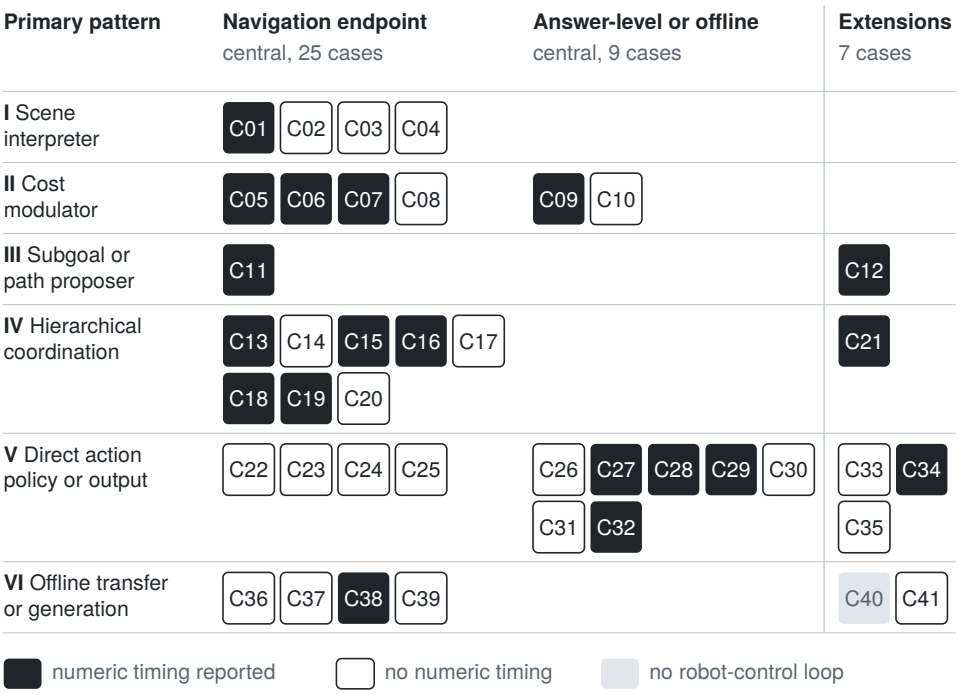


Fig. 3 Evidence map of the 41 cases by primary pattern, scope stratum and timing reporting. Each square is one case, labeled with its ID in Tables 3–4. Columns separate three groups: central cases with a quantitative navigation endpoint, central cases with answer-level or offline endpoints only, and extensions. Filled squares have a numeric timing value in Table 4; the gray square has no robot-control loop

Phase qualifiers matter: SALSA's entry, "none (R)," concerns its runtime policy and excludes the offline inpainting model that builds its counterfactual images from category/context text and a spatial mask produced by Segment Anything Model 2 (SAM2). These designs make explicit geometry a plausible aid to semantic reasoning, though their heterogeneous tasks do not isolate a common causal effect (Sections 5.3 and 6.2).

Table 3: Design characteristics of the 41 cases, grouped by primary architectural pattern. Assignments support counting; patterns can overlap. N, D and X mark, respectively, central cases with navigation endpoints, central cases with answer-level or offline evaluation only and scope extensions. Scaffold denotes geometric model input under the rule of Section 3.8: T, training; R, runtime; G, offline generation; none, absent from the described interface; n/i, not identified; parentheses restrict a code to a phase or component. n/r, not reported. Evaluation and results appear in Table 4

Case	Year	Backbone	Adaptation	Scaffold
(I) Scene interpreter				
C01: GSON [N] (Luo et al. 2025)	2024	GPT-4o / Qwen2.5-VL-32B	Zero-shot	R
C02: LLM-HNM [N] (Li 2025)	2025	GPT-4o; PPO policy	Few-shot prompting; RL	none
C03: HCSG [N] (Xu et al. 2026)	2026	Qwen3-VL-2B + CLIP; trained waypoint policy	Zero-shot VLM; trained policy	none
C04: SAATT Nav [N] (Zhang et al. 2026d)	2026	LLM (unnamed)	Prompting	R
(II) Cost modulator				
C05: BehAV [N] (Weerakoon et al. 2025)	2024	GPT-4 + GPT-4o + CLIPSeg	Zero-shot	n/i
C06: VLM-Social-Nav [N] (Song et al. 2025b)	2024	GPT-4V	Zero-shot	n/i
C07: G2-Nav [N] (Liao et al. 2026)	2026	Qwen3.5-2B / GPT-5.4-Nano	Zero-shot	R
C08: Obstacles to Etiquette [N] (Fang et al. 2026)	2026	Qwen2.5-VL-7B (GPT-4o teacher)	Distillation + SFT	T, R
C09: MOSU [D][a] (Liang et al. 2025)	2025	Gemini (version n/r)	Zero-shot	R
C10: Ruo et al. [D][a] (Ruo et al. 2025)	2025	llama3.1:8b / command-r7b	Zero-shot function calling	R
(III) Subgoal/path proposer				
C11: CoNVOI [N] (Sathyamoorthy et al. 2024)	2024	CLIP + GPT-4V / Gemini	Zero-shot	R
C12: CATNAV [X][b] (Potnis et al. 2026)	2026	Gemini 3 Flash + CLIP + CLIPSeg	Zero-shot + caching	R

Continued on next page

Table 3 continued

Case	Year	Backbone	Adaptation	Scaffold
(IV) Hierarchical coordination				
C13: OLiVia-Nav [N] (Narasimhan et al. 2025)	2024	SC-CLIP (GPT-4o teacher)	Distillation + online adaptation	n/i
C14: SALM [N] (Wang et al. 2024d)	2024	GPT-4o; RL policy	Prompting; RL	R
C15: LISN [N] (Chen et al. 2026a)	2025	GPT-4o + Grounded-SAM 2 + RoboPoint	Zero-shot + tools	n/i
C16: HUMA [N] (Ahmadi et al. 2026)	2026	Qwen3-VL-2B; RL policy	LoRA fine-tuning	R
C17: PAD [N] (Bejarano Sepulveda et al. 2026)	2026	VLM + LLMs (unnamed)	n/r	n/i
C18: Slow Brain [N] (Peng et al. 2026b)	2026	Gemini 2.5 Flash-Lite	Zero-shot	R
C19: TIC-VLA [N] (Huang et al. 2026b)	2026	InternVL3-1B + action expert	SFT + imitation + RL	T (teacher)
C20: Walk With Me [N] (Zhang et al. 2026a)	2026	VLM (six tested); low-level policy (five tested)	Prompting	none
C21: Dual-process [X][c] (Lee et al. 2026)	2026	Gemma-3-4B	Zero-shot	R
(V) Direct action policy/output				
C22: HSAC-LLM [N] (Wen et al. 2025)	2024	GPT-4; hybrid SAC policy	Prompting; RL	none
C23: TrustNavGPT [N][d] (Sun et al. 2024b)	2024	LLM + Whisper (GPT-4 + GPT-3.5 in Beyond Text)	Few-shot	R
C24: Liu et al. [N][e] (Liu et al. 2025c)	2025	7B LLM (unnamed, distilled)	LoRA + chain-of-thought distillation	T, R
C25: SALSA [N] (Wang et al. 2026a)	2026	OmniVLA	LoRA post-training	none (R)
C26: AutoSpatial [D][a] (Kong et al. 2025)	2025	LLaVA-1.6-OneVision-7B	Instruction tuning	T
C27: MAction-SocialNav [D][a] (Wang et al. 2026b)	2025	NVILA-Lite-2B	LoRA + projector tuning	n/i

Continued on next page

Table 3 continued

Case	Year	Backbone	Adaptation	Scaffold
C28: SocialNav-MoE [D][a] (Kawabata et al. 2025)	2025	5.74B MoE (Phi-2 + SigLIP)	SFT + RFT + MoEFT	n/i
C29: E-SocialNav [D][a] (Xiao et al. 2026)	2026	Phi-2-2.7B + SigLIP	SFT + DPO	n/i
C30: GCL [D][a] (Zhang et al. 2026b)	2026	Qwen2.5-VL-3B learner + Qwen3-VL-4B guide	Supervised GCL	n/i
C31: Probing Prompt Design [D][a] (Xiao and Yamasaki 2026)	2026	GPT-4o / TinyLLaVA (SigLIP)	Prompting / SFT (small models)	n/i
C32: SOPD-SocialNav [D][a] (Zhang et al. 2026c)	2026	Qwen2.5-VL-3B (8B teacher)	Selective distillation	n/i
C33: SafeVLA [X][c] (Zhang et al. 2025)	2025	SPOC-DINOv2 VLA	Constrained RL	none
C34: SocialNav [X][f] (Chen et al. 2026b)	2025	Qwen2.5-VL-3B + flow-matching action expert	Imitation + RL	T, R
C35: CrowdVLA [X][g] (Hwang et al. 2026)	2026	Qwen3-VL-2B	LoRA	T

Continued on next page

Table 3 continued

Case	Year	Backbone	Adaptation	Scaffold
(VI) Offline transfer or generation				
C36: Kathuria [N] (Kathuria 2025)	2025	GPT-4o-mini ranker; inverse-RL reward learner	Few-shot ranking; inverse RL	G
C37: Narrate2Nav [N] (Payandeh et al. 2026)	2025	Qwen2-VL-72B annotator; ∼330M teacher; ∼29M student	Self-supervised distillation	T
C38: ViLAM [N] (Elnoor et al. 2025)	2025	GPT-4o; VANP student	Distillation	T
C39: CommNav [N] (Sacco et al. 2026)	2026	Qwen3-8B generator; BERT + DDPPO policy	Offline generation; RL	G
C40: GROVE [X][h] (Nguyen et al. 2026)	2026	Gemini 2.5 Flash / Qwen3-0.6B (selector) + Gemini 3 Pro / Qwen3-4B (tree generator) + RAG	Zero-shot generation	G
C41: HumAIN [X][i] (Song et al. 2026)	2026	Multimodal Transformer teacher; vision student	Distillation	T

Table 3 notes. In Backbone and Adaptation, + joins models or stages used together, a semicolon separates them from other learned modules such as policies or students, and / separates tested alternatives. Year denotes the arXiv identifier year, not a revision or venue year; LLM-HNM, Ruo et al., PAD, Liu et al. and Kathuria, which have no arXiv identifier, use the publication year. [a] Evaluated without quantitative closed-loop navigation endpoints (answers, parameter outputs or offline selections); MOSU states a real-world evaluation but reports no robot results, and SOPD-SocialNav adds qualitative robot examples. [b] Traversability, path-selection and caching extension. [c] Outside the central social scope: non-social tasks, retained as mechanism boundaries. [d] Coded with its companion report, Beyond Text (Sun et al. 2024a), which presents the same audio-cue experiment with further ablations. [e] Coded with its companion report (Liu et al. 2025b), which gives the closed-loop comparison with CrowdNav++. [f] Semantic-traversability and region-compliance extension. [g] Models pedestrians rather than the robot. [h] Generates pedestrian scenarios rather than a robot policy. [i] Non-language multimodal teacher.

Abbreviations. RL, reinforcement learning; SFT, supervised fine-tuning; RFT, reinforcement fine-tuning; MoE, mixture of experts; MoEFT, mixture-of-experts fine-tuning; DPO, direct preference optimization; GCL, group competitive learning; RAG, retrieval-augmented generation; PPO, proximal policy optimization; DDPPO, decentralized distributed proximal policy optimization; SAC, soft actor-critic; LoRA, low-rank adaptation.

Table 4: Evaluation and reported results for the cases of Table 3, grouped in the same way; [N], [D] and [X] mark the same scope strata. Evaluation gives the setting and its reported scale, and Compared with names the reported comparators, including those of the headline result. Headline results keep the source's metrics and values; timing keeps source units and measurement boundaries, so values are not directly comparable across rows. n/r denotes not reported (for timing, no numeric value for the listed configuration); n/a denotes no applicable robot-control loop

Case	Evaluation	Compared with	Headline result	Timing
(I) Scene interpreter				
C01: GSON [N]	Real, 4 scenarios × 3 runs; offline grouping, 100 scenarios	TEB (with and without group estimates), SACSON, GMPC, RLCA; DBSCAN (grouping)	Exact grouping 77.8% vs. 10.8–14.8% (DBSCAN); real: lowest plotted disturbance durations of 6 methods	1.5–2 s/query (32B model, dense crowds)
C02: LLM-HNM [N]	Sim (grid world) + user study; 18 participants	Pareto-optimal and averaging mediators	Social-zone violations/run ≈0.5 vs. ≈1.7 (Pareto) and ≈4.5 (averaging); slightly longer paths	n/r
C03: HCSG [N]	Sim (HA-VLNCE unseen); Real, 5 scenarios × 10 trials	8 VLN baselines incl. BEVBert; no-VLM ablation	Unseen SR 0.24 vs. 0.21 (BEVBert) and 0.23 (no VLM); real success 41/50 vs. 19/50 (BEVBert)	n/r
C04: SAATT Nav [N]	Sim (2D); 3 scenarios × 30 seeds per method	A*, SFM; LLM-disabled ablation	Rank orders only: first of 4 on collisions and minimum pedestrian distance in 2 of 3 scenarios, last on travel time there	n/r (periodic and proximity-triggered calls)
(II) Cost modulator				
C05: BehAV [N]	Real (outdoor); 5 scenarios × ≥10 trials	GA-Nav, CoNVOI, ViNT, NoMaD	BFA 76.14–89.65% vs. 38.67–81.43% (baselines), highest in each scenario	Query ∼6 s (3.59 s offline); system 9.5 Hz
C06: VLM-Social-Nav [N]	Real (indoor); 4 scenarios × 21 trials; user study	DWA; behavior cloning	Vs. DWA, 4-scenario mean: success +27.38 pp, collisions −19.05 pp (incl. preventive interventions)	2–3 s/query

Continued on next page

Table 4 continued

Case	Evaluation	Compared with	Headline result	Timing
C07: G2-Nav [N]	Real, 6 trials per method; offline SCAND, 83 trials	5 baselines incl. SF, PAP; default-score ablation	Real: goal time 278 vs. 322–518 s (baselines); time within 0.3 m 6.9 vs. 4.4 s (SF, lowest)	1.5 s/query (recorded, 2B); 4 s/query (real, online)
C08: Obstacles to Etiquette [N]	Real (Spot); 4 scenarios × 3 trials	5 baselines incl. VLM-Social-Nav, GSON	Social-zone interruption 0.00 s and max. ratio 0.00% in all 4 scenarios; lowest baseline (scenario a) 0.71 s, 2.86%	n/r (off-board H100)
C09: MOSU [D]	Offline (GND); no robot results reported	6 baselines incl. PIVOT, VL-TGS	Traversability 77% vs. 70% (PIVOT); distance to target 73% vs. 70% (VL-TGS), described as comparable	2.30 s inference (boundary n/r)
C10: Ruo et al. [D]	Offline answers, 100 requests; Gazebo demos	None	Parameter accuracy 87/76/85% (mode/speed/distance)	n/r
(III) Subgoal/path proposer				
C11: CoNVOI [N]	Real, 4 scenarios (trials n/r); offline ablation	4 baselines; teleoperation reference; filter ablation	Lowest Fréchet distance to teleoperation in all 4 scenarios; path error 0.187 vs. 1.576 without filter (GPT-4V, scenario 1)	3.28–5.26 s/query (incl. network)
C12: CATNAV [X]	Real (Go1); ∼2 h of driving per cache setting	Threshold 0.1 (caching effectively off)	Query rate 0.006 vs. 0.042 Q/s; total calls 166 vs. 420 (path selector 157 vs. 295)	3.717 vs. 17.35 s mean latency (boundary n/r)
(IV) Hierarchical coordination				
C13: OLiVia-Nav [N]	Real (Jackal); 4 scenarios × 3 trials; offline ablation	VLM-Social-Nav, MultiSoc; embedding ablations	Lowest MSE, Hausdorff loss and PSV in all 4 scenarios; PSV 1.2/0.4/2.1/3.3 vs. 1.9/2.6/3.2/4.5 s (best baseline)	5 Hz mean method response
C14: SALM [N]	Sim; 2 tasks × 50 test cases	4 baselines incl. SRNN; branch and fixed-weight ablations	Mean SR 86% vs. 75% (without language branches; also SRNN) and 63% (fixed 0.5/0.5 fusion)	n/r

Continued on next page

Table 4 continued

Case	Evaluation	Compared with	Headline result	Timing
C15: LISN [N]	Sim (LISN-Bench); 5 tasks	VLM-Nav, VLM-Social-Nav (reimplemented)	Mean SR 91.3% (5 tasks); Follow Doctor 100% vs. 0% (both baselines)	7,094.16 ms slow loop; 5.27 ms fast loop
C16: HUMA [N]	Sim (Social-HM3D, Social-MP3D); real demo	5 baselines incl. Falcon, NavThinker (published values); timing variants	Freeze timing SR 62.07% vs. 59.46% (NavThinker, Social-HM3D) and 70.35% vs. 51.67% (Falcon, Social-MP3D); Realistic 56.90% (Social-HM3D)	2,425 ms per call
C17: PAD [N]	Real (TIAGo++); 8 participants, 1 trial per condition	SLAM-only (no VLM/LLM); two dialogue strategies	Success 8/8 per dialogue strategy vs. 2/8 (SLAM-only); duration 99.95 vs. 162.21 s (conservative vs. assertive)	n/r
C18: Slow Brain [N]	Real, 5 routes (supervised); offline logs	Local-only; streaming without fusion; matching; planner's top choice (offline)	Takeovers/100 m 0.87 (probability fusion) vs. 3.49 (local-only), 4.40 (streaming without fusion); offline hard-subset ADE 1.16 (Gemini 3 Flash) vs. 1.64 m (planner's top choice)	1.5–3.0 s round trip (typical); 1 Hz requests
C19: TIC-VLA [N]	Sim (DynaNav), 85 test cases; Real (Go2), 4 tasks × 5 trials	Reasoning off; VLN/VLA baselines	Sim, reasoning on vs. off: SR 55.29 vs. 40.00%, CR 28.24 vs. 25.88%; VLM kept in both	85.73 ms action; 3,430.73 ms reasoning (RTX 4060)
C20: Walk With Me [N]	Real (outdoor); 4 scenarios × 5 trials; swaps, 10 trials each	None; 6 high-level VLMs, 5 low-level policies	Success 12/20 (60%); delivery swaps 3–6/10 (VLM) and 2–6/10 (low-level policy) vs. 70% (full system)	n/r
C21: Dual-process [X]	Real (5 buildings, non-social); 120 episodes	Continuous VLM-only control	Episode success 67.5% vs. 5.83% (VLM only); System One resolves more than half of decisions	0.7–0.9 s System One; 17–19 s System Two

Continued on next page

Table 4 continued

Case	Evaluation	Compared with	Headline result	Timing
(V) Direct action policy/output				
C22: HSAC-LLM [N]	Sim (2D, 1 pedestrian); 3 scenes × 500 tests	A2C, PPO, TD3; one-way variant	Success 0.91/0.93/0.91 vs. 0.85/0.77/0.85 (TD3) and 0.86/0.91/0.87 (one-way variant without LLM module)	n/r
C23: TrustNavGPT [N]	Offline, 500 audio clips; Sim (RoboTHOR), 10 scenes	Text only, CoT; LM-Nav	Preferred option 70.46% vs. 22.16% (text only), 49.30% (CoT); target finding 80% vs. 25% (LM-Nav)	n/r
C24: Liu et al. [N]	Sim (episodes n/r); replayed field data	CrowdNav++ at 1 s and 0.25 s	Success/collisions 95/1% vs. 70/4% (CrowdNav++, 1 s) and 100/0% (CrowdNav++, 0.25 s)	n/r (1 s decision interval)
C25: SALSA [N]	Offline (SCAND), 4,902 frames; small Unitree trials	Base OmniVLA; stage ablations	Near-collision frames 334 vs. 2,460; CF-Acc 93.0% vs. 52.8% (base OmniVLA)	n/r
C26: AutoSpatial [D]	VQA (CODA auto-labeled test; SNEI)	Untuned; manual only, auto only	CODA 0.710 vs. 0.376 (untuned), 0.404 (manual only) and 0.680 (auto only)	n/r
C27: MAction-SocialNav [D]	VQA, 79 test samples	No MCP; GPT-4o, Claude	APG 0.595 vs. 0.532 (no MCP) and 0.000/0.025 (GPT-4o/Claude)	1.524 FPS (boundary n/r)
C28: SocialNav-MoE [D]	VQA (SNEI), 60 test images	GPT-4o, Claude	SMS 0.551 vs. 0.376/0.417 (GPT-4o/Claude); 8.1×/19.6× their throughput	1.709 FPS (action turn only)
C29: E-SocialNav [D]	VQA (SNEI-derived), 60 test samples	Social-LLaVA; GPT-4o, Claude	AA 0.550 vs. 0.483 (Social-LLaVA), 0.450 (GPT-4o) and 0.417 (Claude)	2.354 FPS (phase n/r)
C30: GCL [D]	VQA (SNEI, MUSON)	SFT baselines (3B, 4B, 8B)	SNEI Action-F1 0.968 (3B learner, best setting) vs. 0.692/0.816/0.755 (3B/4B/8B SFT)	n/r
C31: Probing Prompt Design [D]	VQA, 60 SNEI + 160 MUSON samples	No system prompt	GPT-4o AA, best prompt vs. none: SNEI 0.633 vs. 0.550; MUSON 0.563 vs. 0.531	n/r

Continued on next page

Table 4 continued

Case	Evaluation	Compared with	Headline result	Timing
C32: SOPD-SocialNav [D]	VQA (SNEI, MUSON); robot examples	SFT; distillation baselines	SNEI accuracy 0.889 vs. 0.863 (best distillation baseline) and 0.692 (SFT); student VLM stays online	0.984 model FPS; 2.74 s end-to-end
C33: SafeVLA [X]	Sim (Safety-CHORES), 3 tasks; dual-arm demo	FLaRe, FLaRe-RS, SPOC	3-task average vs. FLaRe: cumulative cost −83.58%, success +3.85% (as reported)	n/r
C34: SocialNav [X]	Sim, 9 scenes × 20 test cases; Real (Go2), 3 sites × 20 trials	4 baselines incl. CityWalker; social-reward ablation	DCR 82.5 vs. 61.4 (no social reward) and 36.1 (CityWalker); real SR 85.0 vs. 62.5 (CityWalker)	>5 Hz deployment (call window n/r)
C35: CrowdVLA [X]	Sim (pedestrians), 4 scenes	GBM, CCP; skill and question-answer ablations	Intersection SRVR 0.0170 vs. 0.0290–0.1475 (baselines) and 0.1369 (no question-answer pairs)	n/r (20-frame skills)
(VI) Offline transfer or generation				
C36: Kathuria [N]	Sim (narrow crossing), runs n/r; 180 rated demonstrations	Learner without ratings; human ratings	Success 76.66% vs. about 92% (learner without ratings, separate experiment); exact agreement 69.68% with human ratings	n/r
C37: Narrate2Nav [N]	Offline (unseen set); Real, 4 scenarios × 10 trials	GNM, ViNT, NoMaD; text ablation	ADE 0.16 vs. 0.34 (NoMaD); real success 8.5/10 vs. 6/10 (GNM) and 5.75/10 (ViNT, NoMaD)	n/r
C38: ViLAM [N]	Real (Husky); 4 scenarios × 10 trials	DWA, CoNVOI, VANP	Success 100/90/90/80% vs. 80/60/60/70% (best baseline per scenario); scenario-4 Fréchet 1.034 vs. 1.452 (CoNVOI)	∼20 Hz (boundary n/r)
C39: CommNav [N]	Sim (Habitat 3.0c); 400 episodes (human text: 100)	No module; LLM-text and human-text modules	Episode success 0.24 (structured) vs. 0.20 (LLM text), 0.23 (human text) and 0.14 without module (0.16 at matched steps)	n/r

Continued on next page

Table 4 continued

Case	Evaluation	Compared with	Headline result	Timing
C40: GROVE [X]	Sim; 3 presets × 3 scenarios, GPT-5 judging	Text-Crowd, TRACE (reimplemented)	Highest average score in all 3 presets: 5.33/6.57/6.62 vs. 4.43/5.50/4.61 (best baseline)	n/a (robot loop); generation n/r
C41: HumAIN [X]	Offline (OOD set, ∼2,000 samples); robot demos	4 baselines incl. GNM, ViNT; teacher ablations	ADE/FDE/AOE 0.578/1.064/0.159 vs. 0.753/1.254/0.203 (best baseline per metric)	n/r

Table 4 notes. Real, real-robot trials; Sim, simulation; Offline, recorded or logged data; VQA, visual question answering; published values, baseline results taken from the original papers. FPS is the source's frames-per-second metric, not a robot-command rate. pp, percentage points; SR, success rate; CR, collision rate; BFA, behavior-following accuracy; PSV, personal-space violation duration; MSE, mean squared error; DCR, distance compliance rate; SRVR, social-region violation rate; CF-Acc, counterfactual accuracy; APG, all-predictions-in-ground-truth accuracy; MCP, meta-cognitive prompt; SMS, Sentence Mover's Similarity; AA, action accuracy; ADE and FDE, average and final displacement error; AOE, average orientation error; Q/s, queries per second; OOD, out-of-distribution; CoT, chain-of-thought prompting; SFT, supervised fine-tuning; SF and SFM, social force model; TEB, timed elastic band; DWA, dynamic window approach; A2C, advantage actor-critic; PPO, proximal policy optimization; SLAM, simultaneous localization and mapping; TD3, twin delayed deep deterministic policy gradient; GND, Global Navigation Dataset; Social-MP3D and Social-HM3D, social-navigation benchmarks built on Matterport3D and Habitat-Matterport 3D. TIC-VLA's hardware timing and simulation ablation describe separate tests (Section 3.5).

4 Datasets, simulators and benchmarks

The case comparisons rest on how training data represent social encounters, how simulated pedestrians respond and what benchmarks score. Each resource fixes what a comparison can test. Table 5 lists selected resources, and Section 5 builds on them to assess the contribution evidence. The searches of Section 2.4 also returned 14 evaluation resources. They comprise SNEI, MUSON and SocialNav-SUB; CrowdVLA, which the search returned as a resource but which, as a pedestrian policy rather than a simulator, is coded as an extension (C35); eight simulation platforms, benchmarks and scenario generators for human-centric navigation, several of which use language models to generate scenes, pedestrian behavior or instructions (Shcherbyna et al. 2025; Marpally et al. 2024; Huang et al. 2026a; Dai et al. 2026; Li et al. 2024; Peng et al. 2026a; Wang et al. 2024b; Sun et al. 2025); and two benchmarks that evaluate VLMs on human-perception or embodied tasks (Shi et al. 2025; Cheng et al. 2025).

4.1 Real-world datasets

Real-world datasets determine whose behavior a model imitates and which cues it can observe. SCAND and MuSoHu recur in the reviewed corpus, though each underlies only some systems. SCAND (Karnan et al. 2022) records teleoperated navigation from the robots' own sensors, capturing behavior humans considered appropriate *for a robot*. It supplies the training data for VLM-Social-Nav's behavior-cloning baseline (Section 3.3) and the annotation corpus behind SNEI, and its scenes underlie

SocialNav-SUB, so its collection conditions bear on the interpretation of all three. MuSoHu (Nguyen et al. 2023) records *human* navigation with casual, neutral or rush context labels. Its trajectories demonstrate pedestrian motion only, and their transfer to a wheeled robot and a quadruped is a small proof of concept.

Other corpora mainly serve perception, prediction or scene rating. THÖR (Rudenko et al. 2020a) and THÖR-MAGNI (Schreiter et al. 2025) track indoor human motion; JRDB (Martín-Martín et al. 2023) provides egocentric person tracking and its JRDB-Act extension (Ehsanpour et al. 2022) adds social-group labels relevant to scene interpreters (pattern I); and SocNav1 (Manso et al. 2020) rates the disturbance of robot placements in synthetic static scenes. TBD (Wang et al. 2024a) contributes person-pushed recordings with metric trajectory labels for social-navigation research. Prediction work (Section 2.2) draws on heterogeneous trajectory sources, from ETH (Pellegrini et al. 2009), UCY (Lerner et al. 2007) and Stanford Drone (Robicquet et al. 2016) to public-building tracking corpora (Brščić et al. 2013) including L-CAS (Yan et al. 2017).

SCAND and MuSoHu both come from North American settings, so validation under other regional norms requires evidence beyond their reuse. Later corpora add language-reasoning annotations produced through different processes. SNEI attaches more than 40,000 manually annotated visual question–answer pairs to 2,000 SCAND interactions (Payandeh et al. 2025), whereas nine experts verify or revise MUSON's 10,110 model-assisted annotations before a final consistency check (Liu et al. 2025d). MUSON's authors describe SNEI's action-label distribution as long-tailed and report inconsistencies between its reasoning and action annotations.

4.2 Simulators

Simulators allow closed-loop behavior to be tested, and two studies show that measured crowd dynamics and learned-policy trade-offs depend on how pedestrians are generated. Replay in SocNavBench (Biswas et al. 2022) makes comparisons repeatable, while the replayed pedestrians ignore the robot's motion. SEAN 2.0 (Tsoi et al. 2022), which extends SEAN (Tsoi et al. 2020), and HuNavSim (Pérez-Higueras et al. 2023) instead generate pedestrian behavior through behavior graphs and configurable Robot Operating System 2 (ROS 2) models, respectively. SocialGym 2.0 (Sprague et al. 2023) extends SocialGym (Holtz and Biswas 2022) to multi-robot policy learning; the Habitat line brings photorealistic simulation, with humanoid avatars added in Habitat 3.0 (Savva et al. 2019; Szot et al. 2021; Puig et al. 2024); and the Falcon study introduces Social-HM3D and Social-MP3D (Gong et al. 2025). The Menge crowd simulator, the static-obstacle BARN benchmark and iGibson 2.0 supply related infrastructure (Curtis et al. 2016; Perille et al. 2020; Li et al. 2022).

NavIsaacLab (Xia et al. 2026) quantifies one limitation of scripted pedestrians. Its articulated pedestrians, based on the Skinned Multi-Person Linear (SMPL) body model, follow trajectory-diffusion paths under a joint-level controller trained with adversarial motion priors. With that controller fixed, replacing the diffusion-generated paths with social-force ones raises the rate at which pedestrians collide with each other from 7.81 to 17.17%, lowers their navigation success from 88.67 to 74.24% and raises their fall rate from 1.12 to 4.99%. The controller misses some reference locomotion clips

Table 5 Selected datasets, simulators and benchmarks used in the surveyed literature. RGB, red–green–blue; LiDAR, light detection and ranging; IMU, inertial measurement unit; VQA, visual question answering; MCQ, multiple-choice question; BEV, bird's-eye view; GPS, Global Positioning System; ROS 2, Robot Operating System 2

Resource	Year	Content	Modality	Primary use here
SCAND[a]	2022	8.7 h; 138 trajectories; 25 mi	RGB, LiDAR, odometry, IMU, joystick commands	Training; source of SNEI/SocialNav-SUB
MuSoHu[b]	2023	∼20 h; 300 trajectories; 100 km	360° RGB, stereo/depth cameras, LiDAR, IMU, audio	Training; context labels
THÖR[c]	2020	>600 tracked indoor human trajectories	Motion capture	Prediction; evaluation
JRDB / JRDB-Act[d]	2023/ 2022	64 min; 54 sequences; Act: actions/groups	RGB/stereo, depth, LiDAR, audio, GPS, encoders, motion	Perception/groups
SNEI[e]	2025	>40K VQA pairs on 2K interactions	RGB + text	Fine-tuning; VQA evaluation
MUSON[f]	2025	10,110 samples; 8 scene types; 6 actions	RGB + text	Fine-tuning/evaluation
SocialNav-SUB[g]	2025	4,968 MCQs; 60 scenes; 153 participants	RGB + BEV + text	Diagnostic evaluation
SocNavBench[h]	2022	Recorded pedestrian trajectory replay	Simulation	Evaluation
SEAN 2.0[i]	2022	Social-situation generation	Simulation	Evaluation
HuNavSim[j]	2023	ROS 2 human behavior models	Simulation	Benchmarking
SocialGym 2.0[k]	2023	Multi-agent social navigation	Simulation	Benchmarking
Habitat 3.0[l]	2024	Photorealistic human–robot cohabitation	Simulation	Training/evaluation
NavIsaacLab[m]	2026	Physics + learned pedestrian policies	Simulation	Benchmarking

Table 5 sources. [a](Karnan et al. 2022); [b](Nguyen et al. 2023); [c](Rudenko et al. 2020a); [d](Martín-Martín et al. 2023; Ehsanpour et al. 2022); [e](Payandeh et al. 2025); [f](Liu et al. 2025d); [g](Munje et al. 2025); [h](Biswas et al. 2022); [i](Tsoi et al. 2022); [j](Pérez-Higueras et al. 2023); [k](Sprague et al. 2023); [l](Puig et al. 2024); [m](Xia et al. 2026).

under the paper's motion-tracking criterion, which bounds claims of physical fidelity. NavIsaacLab's diffusion planner and body controller build on Rempe et al. (2023), who report one to three seconds to sample one character depending on guidance, a simulation and training cost distinct from robot-control throughput.

Policy evaluation shows the same dependence. Mavrogiannis et al. (2023) initialize four learned-policy families with ORCA or social-force demonstrations and refine them by reinforcement learning, and the training pedestrian model shifts the reported time–collision trade-offs. The authors describe social-force-trained variants as generally faster and more collision-prone and report better transfer from social-force training to ORCA testing than the reverse. GROVE's scenarios (Nguyen et al. 2026), specified by an LLM as behavior trees and executed with a global planner and social-force avoidance, carry generator assumptions of their own. All these findings hold within their simulators and policies, and their bearing on general social realism and real pedestrians is untested.

HA-VLN 2.0 (Dong et al. 2026) scores human-aware behavior among moving people separately from goal reaching. Its collision checks at step endpoints make collision measurements depend on the action step, so cross-agent comparisons need the action discretization and collision-checking procedure to be reported. We treat HA-VLN 2.0, including its LLM-assisted instructions and agents, as evaluation evidence rather than an integration case.

4.3 Question-answering benchmarks

Question-answering benchmarks test answer-level reasoning directly and supply the diagnostic evidence introduced in Section 1. Answer quality shifts with each model's inputs and supervision.

SocialNav-SUB (Munje et al. 2025) scores 4,968 multiple-choice spatial, spatiotemporal and social questions on 60 SCAND scenarios against the answer distributions of 153 participants. A rule baseline using pedestrian positions and trajectories reaches an overall probability of agreement (PA) with human responses of 0.64, against 0.62 for the best evaluated VLM, and leads every evaluated VLM in each category, social questions included (0.71 versus 0.66). Because it reads structured positions, its lead does not show that it could replace a VLM that must perceive the scene. Category-wise best VLM scores match or exceed the mean human PA, and all remain below the human-oracle baseline. A bird's-eye view helps GPT-4o in every category but has small, mixed effects for Gemini and LLaVA, and supplied ground-truth spatial and spatiotemporal answers improve Gemini's social answers, consistent with a spatial bottleneck in that model. A preliminary open-loop waypoint-selection test finds that human-oracle context improves alignment for three models.

The annotation contrast between SNEI and MUSON reappears in training. In the MUSON study, chain-of-thought training raises a small model's action accuracy on MUSON but lowers it on SNEI (Liu et al. 2025d), as Section 3.6.3 details; the authors attribute this to annotation consistency but test that explanation only on these two datasets and this model. Two further diagnostics qualify answer gains. In a study of whether VLMs can assess proxemic risk from egocentric robot images, Rudas and Kuzmenko (2026) assign labels manually without measured distance thresholds,

and the models' classifications and spatial grounding can disagree. Prompt effects in Probing Prompt Design (Section 3.6.3) depend on model and dataset, with some prompts scoring below runs without a system prompt.

General spatial benchmarks provide complementary diagnostics of spatial answers and of tested adaptation recipes. GPT-4o reaches 59.0% test-set accuracy on BLINK's 14 tasks against a 95.7% human reference. Text-only GPT-4-0125-preview given GPT-4V's task-agnostic dense captions instead of the images scores 36.0% on BLINK, compared with 51.1% for GPT-4V with the images and 38.1% for random choice; the gap is specific to that caption pipeline (Fu et al. 2025). SpatialVLM's two billion synthetic question–answer pairs improve spatial VQA (Chen et al. 2024a). In a matched ablation on EmbSpatial-Bench, adding language-backbone LoRA to connector-only tuning raises MiniGPT-v2's scores under both generation and likelihood scoring, and likelihood-based scores exceed generated-answer scores in both configurations (Du et al. 2024). In RoboSpatial's LLaVA-NeXT ablation, training with RoboSpatial and auxiliary grounding data raises total validation accuracy from 30.3% to 60.5%, and on RoboSpatial-Home, a benchmark of novel indoor scenes, fine-tuned models score 8.2–21.3% on point-output free-space questions (Song et al. 2025a). RoboSpatial's source scenes exclude humans and animals, and closed-loop social navigation lies outside every benchmark and diagnostic in this subsection.

5 Evaluation and evidence synthesis

Judging an online model's contribution requires an appropriate endpoint and an informative comparison. We first review the metrics in use and map them onto evaluation principles, then discuss what the reported comparisons establish for system utility, component contribution and online necessity, and finally relate the invocation ratio to request capacity and examine timing evidence, from which a minimum reporting set follows.

5.1 Evaluation metrics

Reported measurements form three groups: task metrics, social metrics and measures of answer and explanation quality prominent in VLM work. Each answers a different question.

Task metrics capture goal completion and motion efficiency: success rate, success weighted by path length (SPL) (Anderson et al. 2018a), path length and time to goal. A robot can reach its goal efficiently while forcing pedestrians to give way, so social appropriateness needs measures of its own.

Social metrics characterize how the robot navigates among people: personal-space violations, minimum and average distance to pedestrians, time to collision, path irregularity and freezing rate; the last two also reflect motion quality and failure to progress, so their social reading depends on the scenario. For legged robots, locomotion noise adds an acoustic channel: Park et al. (2026) report, alongside personal-space compliance, the A-weighted sound-level increment at the listener and the fraction of a run during which it exceeds 5 dBA. Subjective instruments such as Godspeed (Bartneck et al. 2009), the Robotic Social Attributes Scale (RoSAS) (Carpinella et al. 2017) and

the NASA Task Load Index (NASA-TLX) (Hart and Staveland 1988) target distinct constructs, from perceived safety to workload, and favorable ratings do not establish collision safety or norm compliance. Because human-subject studies in the screened literature differ in instruments and protocols, their responses cannot be pooled. Their results can also concern a specific intervention: the 30-person comparison of Sotomi et al. (2025), with and without explanations, evaluates the explanations, and its navigation measures cover the whole VLM-based system without isolating the VLM's contribution to the robot's motion. Validation protocols for social-navigation policies have been proposed (Pirk et al. 2022), and Francis et al. (2025) recommend reporting a battery of metrics with distributional information and caution against claims of statistical significance without the corresponding tests.

Answer and explanation quality is assessed through visual question answering accuracy, agreement with a human-answer distribution in SocialNav-SUB (Munje et al. 2025), semantic similarity to reference answers in SocialNav-MoE (Kawabata et al. 2025), and human-judge or model-judge ratings of free-form explanations. These scores evaluate observable outputs; they do not test the faithfulness of internal reasoning or establish the link to executed behavior (Section 6.5). MUSON's collision-rate label, a manual judgment of whether a predicted action would cause a collision, likewise remains annotation-level, distinct from a controller's collision rate (Liu et al. 2025d).

5.2 Metrics and evaluation principles

Table 6 pairs each of the eight overlapping principles of Francis et al. (2025) with a reported or suggested assessment and keeps each measure tied to the construct it captures. The principles' relative importance depends on context: in interviews with eight robotics experts, collision avoidance was almost universally the top priority, while other priorities varied by domain and role (Tsoi et al. 2024). Collision and distance measures cover safety and a geometric proxy for comfort, whereas user-study items, scenario success, time to first action and VQA accuracy measure other constructs. Time to first action measures latency and bears on proactivity only alongside a scenario-based assessment of initiative. The last two columns contrast what VQA can probe with the behavior or human response that an answer score leaves unmeasured. Higher-level norms such as queue order or letting passengers leave an elevator, which Gao and Huang (2022) distinguish from motion conventions such as passing side, need evidence beyond smooth motion. Comfort likewise needs human-response evidence beyond distance and smoothness (Francis et al. 2025). In the terms of Dragan et al. (2013), legibility concerns how readily observers infer the goal, and predictability concerns expected motion given a known goal.

5.3 Evidence on the contribution of online VLM inference

RQ1 is answered by the role and interface coding of Section 3, whereas RQ2 and RQ3 call for different comparisons (Figure 4). For RQ2, whole-system gains have to be distinguished from the results of within-system removal or replacement tests of a specified input, representation, branch or training signal; such tests hold observations, candidate paths and downstream planning fixed where possible while varying the input

Table 6 Reported measures and suggested assessments for the evaluation principles of Francis et al. (2025), with the related knowledge VQA can probe and the behavior or human response that VQA evidence leaves unmeasured. VQA denotes visual question answering; MCQ denotes multiple-choice question

Principle	Example assessment	VQA can probe	Beyond VQA evidence
Safety	Collision rate; minimum distance	Hazard judgment	Realized safety
Comfort	Personal-space violation; human rating	Predicted comfort	Experienced comfort
Legibility	Observer goal/intent inference accuracy, confidence or timing	Intent prediction	Observer response
Politeness	Yielding success; user-study item	Consideration judgment	Respectful behavior
Social competency	Scenario success; user rating	Norm knowledge	Enacted skill
Agent understanding	VQA/MCQ answer accuracy	Intent/relation inference	Behavioral adaptation
Proactivity	Scenario-based assessment of initiative; time to first action reported separately	Judgment of initiative	Enacted proactivity
Contextual appropriateness	Scenario-specific action accuracy	Judged action	Executed behavior

or module of interest. For RQ3, online execution has to be set against training-time transfer and against implementable alternatives at a stated performance target, with runtime, annotation, engineering and adaptation costs reported; online necessity is judged only against such named alternatives. Oracle social labels can isolate the value of information, while a deployable means of obtaining it must be tested on its own. Identical RGB inputs can also hide unequal pretraining resources, so comparisons should state matched conditions and unknown costs. For studies claiming only system utility, missing tests mark open design questions but do not weaken the claim. Because the comparisons differ in tasks, baselines, hardware and outcome definitions, we do not pool effect sizes.

Applied to the cases, these questions yield five findings. Table 7 lists selected comparisons behind them, and Table 8 turns the findings into conditional design judgments. Most findings rest on one or two studies, and each of the first four holds only for the implementation tested; none establishes that matched alternatives would fail to supply the same function. The navigation stratum supplies the direct robot-navigation evidence for cue access, delayed-advice fusion and teacher-free deployment; the evidence on candidate selection (Slow Brain's hard/routine split) and on the text contribution to knowledge transfer (Narrate2Nav's text ablation) uses offline trajectory metrics, and answer-level cases add evidence about prompting, supervision and output quality.

Cue access. In the cases that test a specific cue (GSON for social groups; VLM-Social-Nav and BehAV for stop gestures), the cue-aware system outperformed its

Table 7 Selected comparisons behind the findings of Section 5.3. Table and figure numbers in the first column refer to the cited paper; case IDs link to Tables 3–4. ADE: average displacement error

Case / source	Changed element / endpoint	Supported inference	Unresolved comparison
C06: VLM-Social-Nav (Song et al. 2025b); Table I	Whole systems; four robot-navigation scenarios.	Cue-aware pipeline utility.	Alternative with the same semantic cues.
C07: G2-Nav (Liao et al. 2026); Table 1	VLM per-object scores vs. default score; recorded-trajectory metrics.	Scoring contribution.	Equivalent alternative scorer.
C11: CoNVOI (Sathyamoorthy et al. 2024); Table I	Occupancy filtering removed; numbered markers retained; reference paths.	Fewer unacceptable paths and shorter inference with filtering.	Selector with the same filtered inputs.
C18: Slow Brain (Peng et al. 2026b); Fig. 2	Offline selector changed; shared candidates; ADE.	Better hard-subset, worse routine selection.	Selectors with matched cues and budgets.
C18: Slow Brain; Table 3	Robot local-only, streaming-only, probability fusion; takeovers/100 m.	Local-only: 3.49; fusion lowers this to 0.87 and streaming raises it to 4.40.	Matched delay-handling alternatives.
C19: TIC-VLA (Huang et al. 2026b); Table 4	Reasoning tokens off/on; success, collisions and VLM forward rate.	Higher success and collision rates; lower VLM forward rate (simulation).	VLM replacement beyond token disabling.
C32: SOPD-SocialNav (Zhang et al. 2026c); Tables I–II	Distillation/token-filter variants; answer datasets.	Training contribution; student VLM stays online.	Quantitative closed-loop and VLM-free comparisons.
C37: Narrate2Nav (Payandeh et al. 2026); Table 3	Text supervision retained/removed; student trajectory errors.	Metric-dependent supervision benefit.	Deployable online language policy vs. student; matched past/current inputs.
C38: ViLAM (Elnoor et al. 2025); Table I	Whole systems; four robot-navigation scenarios.	Student navigation without runtime teacher queries.	Online/distilled alternatives; matched cues and budgets.

Table 8 Findings, evidence and bounded design implications for online inference. Case IDs link to Tables 3–4. These judgments are conditional and do not rank systems

Finding	Observed support	Design implication	Needed comparison
Cue access	C01: grouping; C05, C06: gestures (Luo et al. 2025; Weerakoon et al. 2025; Song et al. 2025b).	Cue-aware input can help when occupancy omits the deciding cue.	Cue-matched alternative; stated training/runtime budgets.
Candidate selection	C18: better hard-subset, worse routine selection (Peng et al. 2026b).	Conditional model selection may help on hard-to-rank candidates.	Selector matched on cues, candidates and budget; hard/routine strata.
Delayed advice	C18: versus local-only, streaming raises takeovers and fusion lowers them; C16: timing variants (Peng et al. 2026b; Ahmadi et al. 2026).	Fusing delayed advice with current plans can preserve its value.	Matched delays; completion counts; observation-to-command timing.
Knowledge transfer	C38: teacher-free deployment; C37: text ablation; C39: structured cues vs. LLM text, unmatched data (Elnoor et al. 2025; Payandeh et al. 2026; Sacco et al. 2026).	Test transfer with student computation included.	Online, cached and distilled alternatives; stated data and budgets.
Representation use	C25: behavior tests; C30: action-text endpoints (Wang et al. 2026a; Zhang et al. 2026b).	Validate the consuming module and target outcome.	Paired diagnostic and closed-loop tests; relevant human response.

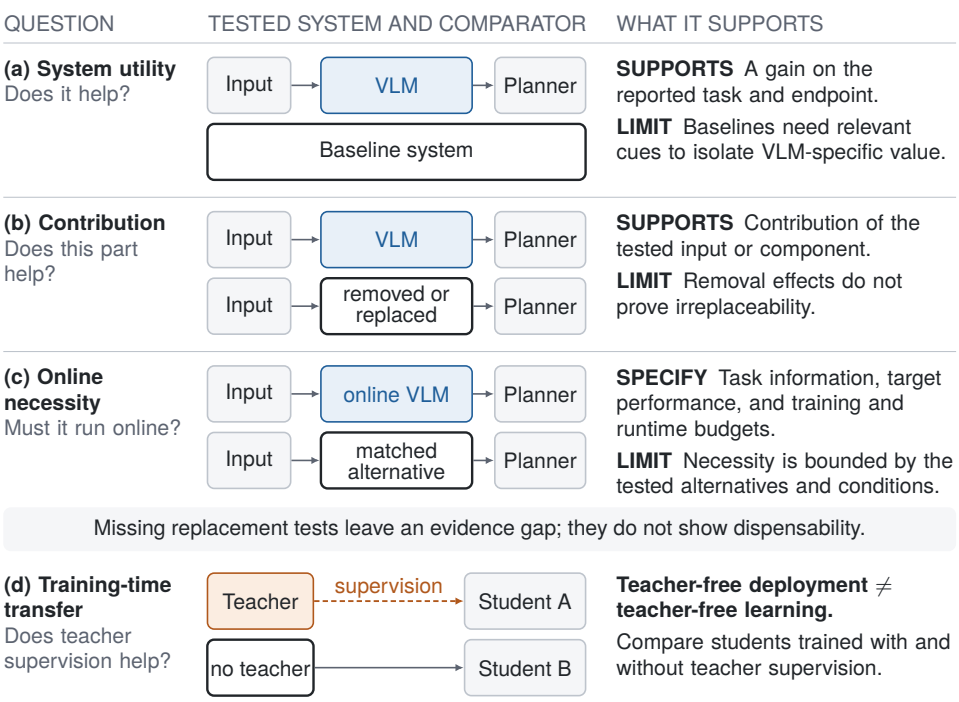


Fig. 4 Comparison questions for assessing system utility, component contribution and online necessity (Sections 1.1 and 5.3). In each row, the upper chain is the tested system and the lower chain its comparator; the outlined white block marks what the comparison changes. Panels (a)–(c) ask distinct questions, and a positive result in one does not settle the others. Panel (d) shows the comparison that training-time transfer needs: students trained with and without teacher supervision (dashed arrow), both of which run without the teacher

baselines on the outcomes that depend on that cue. These are whole-system comparisons, so they show that occupancy alone can miss task-relevant information rather than how much the VLM itself contributes (Luo et al. 2025; Song et al. 2025b; Weerakoon et al. 2025). More focused input tests support narrower claims: excluding occupied numbered regions reduces CoNVOI's unacceptable reference paths in every tested condition, with numbering kept in both variants (Sathyamoorthy et al. 2024), replacing G2-Nav's per-object VLM scores with a default score raises its error against recorded trajectories (Liao et al. 2026), and supplied spatial answers improve Gemini's subsequent social answers in the SocialNav-SUB diagnostic benchmark (Munje et al. 2025). HUMA's comparison of depth-based and RGB-based input changes a composite representation in which depth, contours and goal encoding vary together (Ahmadi et al. 2026). Removal tests show a small effect in HCSG (Xu et al. 2026) and, in SAATT Nav (Zhang et al. 2026d), remove the model together with the rules it triggers (Section 3.2).

Candidate selection. Evidence isolating model-based selection among feasible candidates comes from one offline study: Slow Brain's Gemini 3 Flash selector improves

ADE on the hard subset and worsens it on routine scenes (Peng et al. 2026b), a contrast that motivates evaluating conditional invocation (Section 3.5). Slow Brain and Obstacles to Etiquette also make the candidate-set limit explicit: a selector can only choose among the paths the proposer supplies (Peng et al. 2026b; Fang et al. 2026).

Delayed advice. One supervised deployment shows that delayed advice can stay useful when fused with current plans. In Slow Brain's Gemini 2.5 Flash-Lite deployment, probability fusion lowered takeovers from 3.49 per 100 m under local-only control to 0.87, whereas streaming without fusion raised them to 4.40 (Peng et al. 2026b). In HUMA's simulation, delay itself had a cost: success fell from 62.07% with the scene frozen during inference (*Freeze*) to 56.90% under the paper's *Realistic* timing condition, one of three timing conditions it reports (Ahmadi et al. 2026). TIC-VLA's within-model reasoning comparison shows why success, collisions and activity should be read together to judge the timing trade-off (Section 3.5). SALM's fixed-weight comparison addresses action fusion (Section 3.5). An ablation outside the case set supports non-language lookahead, though social metrics do not improve uniformly: in NavThinker, a non-VLM system, enabling its world-model lookahead without the trajectory-based social reward changes Social-HM3D success from 43.36% to 55.87%, human collisions from 47.33% to 39.19% and personal-space compliance from 90.81% to 90.58% (Hu et al. 2026). Named replacements, such as another adaptive weighting mechanism in place of SALM's LLM or non-language lookahead matched against online VLM reasoning, are still untested.

Knowledge transfer. Students trained with model supervision ran teacher-free: ViLAM and Narrate2Nav issued no runtime queries to the language models used for supervision (Elnoor et al. 2025; Payandeh et al. 2026), although no study yet compares online, cached and distilled alternatives with student computation included. Narrate2Nav's text ablation worsens average orientation error and ADE while leaving final displacement error unchanged, supporting a metric-dependent training contribution (Section 3.7). OLiVia-Nav (Narasimhan et al. 2025) adds evidence for a teacher-assisted update alongside offline transfer (Section 3.5): removing its social embeddings worsens held-out trajectory losses, and one update lowers the MSE and Hausdorff losses. SOPD-SocialNav's distillation compresses the teacher into a student VLM that still runs at deployment (Zhang et al. 2026c). In CommNav (Sacco et al. 2026), a module trained on structured cue records replaces one trained on LLM-generated sentences, although the two modules differ in encoder and training data; the structured module reaches higher episode success (0.24 versus 0.20) at equal finding success, while the language-trained module finds the person in fewer steps (Section 3.7).

Representation use. Better answers or representations need validation in the module that consumes them and on its target outcome, through paired diagnostic and closed-loop tests. SALSA improves the base VLA's offline person–object differentiation and near-collision measures, and its authors add small robot trials (Wang et al. 2026a). Because its ablations concern routing, supervision and alignment, the complete gain belongs to the combined intervention, of which feature access is one part. EmbSpatial-Bench (Du et al. 2024) shows a different mechanism, readout sensitivity (Section 4.3): the same tuned MiniGPT-v2 performs better with likelihood-based readout than with generated answers. GCL and the other answer-level generators (Section 3.6.3)

report action-text endpoints only (Zhang et al. 2026b), and for action-text systems no reviewed study yet links answer quality to executed navigation.

What the extensions add. The extensions clarify mechanisms in adjacent tasks. HumAIN's skeletal-input ablation gives modest, metric-dependent changes in students distilled from non-language teachers (Song et al. 2026), CrowdVLA tests skills and counterfactual supervision in a pedestrian policy (Hwang et al. 2026), and GROVE's global planner supports executable generated scenarios (Nguyen et al. 2026). Direct evidence on VLM use and teacher-free deployment in social robot navigation still comes from the central navigation cases, so these mechanisms need testing in that setting.

Across the findings of Table 8, whose last column lists the comparison each finding needs, future comparisons should name the replacement, provide relevant task information, state training and runtime budgets, and assess both task and social outcomes.

5.4 Invocation ratio and request capacity

The invocation ratio b, defined in Section 3.1 by Eq. (1), counts model requests per control decision at command rate f_c. For one serial server with mean wall-clock occupancy τ per request, the average-capacity screen is

$$b \cdot \tau \cdot f_c \lesssim 1, \tag{2}$$

because the issued-request rate is $\lambda_{\mathrm{VLM}} = b f_c$; the left-hand side is the server utilization $\rho = \lambda_{\mathrm{VLM}} \tau$, so the screen is the utilization condition of a single-server queue. With m parallel service slots, the screen becomes $\lambda_{\mathrm{VLM}} \tau \lesssim m$. Both screens bound mean capacity only; queuing delay, response deadlines, output age (the time from an observation to the use of the resulting output), batching, pipelining and variable service times lie outside them. Figure 5 contrasts a synchronous request, during which control waits, with an asynchronous request whose output is used at the first control decision after it arrives.

Prompt and response length affect occupancy: VLM-Social-Nav constrains its answer to direction and speed tokens (Song et al. 2025b). Asynchronous control can continue while requests are pending, as when Slow Brain (Peng et al. 2026b) fuses delayed advice with current plans (Section 3.5), although its reported network round trip differs from server occupancy τ. A quadruped system refreshes a fine-tuned VLM's compact behavior token at about 0.25 Hz while a social costmap and a 50 Hz locomotion policy act on the most recent cached token, so when a query is delayed, the costmap and policy keep using the previous social context (Park et al. 2026).

The capacity problem admits four recurring responses, which systems can combine (pattern numbers refer to Section 3.1):

1. Remove runtime teacher or generator queries by moving model use before deployment (pattern VI), as in ViLAM, whose GPT-4o annotator runs only offline, a concrete $b = 0$ example (Section 3.7).
2. Reduce b by invoking the model conditionally (pattern IV and, in practice, parts of patterns I and II).

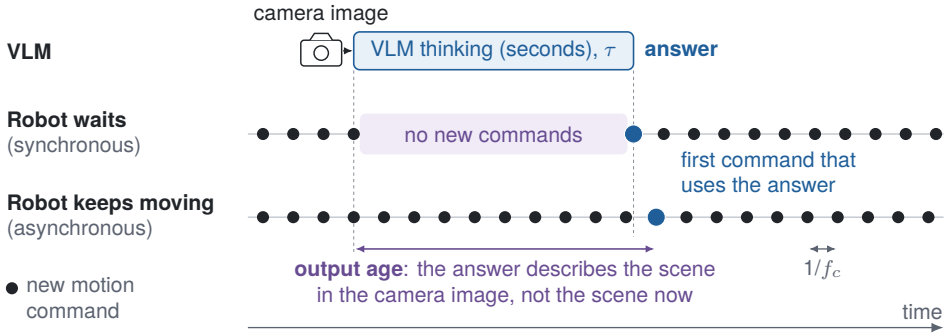


Fig. 5 What the robot experiences while a VLM answers. A request sends a camera image to the VLM, which takes seconds to answer; dots are motion commands, spaced by the command interval $1\ f_c$. A synchronous controller waits, so the robot receives no new motion commands until the answer arrives. An asynchronous controller keeps sending commands, but the answer it eventually uses describes the scene in the camera image rather than the current scene; the time from capturing that image to using the answer is the output age. Section 5.5 proposes reporting these timing quantities together with the invocation ratio b. Schematic, not to scale

3. Seek to reduce τ by shrinking, quantizing or distilling the model.
4. Decouple timescales: a slow VLM supplies a slowly varying quantity while a fast classical component closes the loop (patterns III and IV, and pattern II with a cost layer refreshed at a low rate).

5.5 Timing reports and a minimum reporting set

Reported timing values describe distinct operations under varying measurement boundaries, and the design choices that address study-specific delays set no field-wide rate threshold. This heterogeneity calls for a common reporting set.

Numeric timing appears in 16 of the 34 central cases (Figure 3): 11 of the 25 with a quantitative navigation endpoint and five of the nine answer-level or offline cases. Adding the three extensions with numeric entries (Dual-process, SocialNav and CATNAV) gives 19 of the 41 cases, a count of any numeric timing in Table 4. GROVE lacks a robot-control loop (n/a).

The entries range from answer throughput and individual queries to system ticks and planner updates, on hardware from embedded modules to hosted services, with varying windows and summary statistics. Firoozi et al. (2025) offer a useful adjacent format that lists model, size, hardware and inference/control speeds together. The GPT-4o and Claude throughputs, identical in SocialNav-MoE and E-SocialNav (Section 3.6.3), lack matched timing boundaries and independent replication (Kawabata et al. 2025; Xiao et al. 2026).

Timings for the same navigation baseline vary across studies. Song et al. (2025b) report two to three seconds per GPT-4V query for VLM-Social-Nav after prompt optimization. Chen et al. (2026a) reimplement it and report mean slow-branch latencies over ten runs, all with GPT-4o and timed from image and instruction input to the VLM-agent response: 1,751.80 ms for VLM-Social-Nav, 7,094.16 ms for LISN

and 9,072.10 ms for VLM-Nav. Their fast branch averages 5.27 ms over 100 runs for LiDAR, costmap, social-force planning and control output. These measurements compare components and implementations; they do not compare interface roles or apportion the perception–latency trade-off among individual tools. Narasimhan et al. (2025) time their VLM-Social-Nav reimplementation at 0.6 Hz against a method-level mean response speed of 5 Hz for OLiVia-Nav and attribute VLM-Social-Nav's longer personal-space violations to the resulting action delays, without an encoder-only timing measurement. These baseline values are of comparable order of magnitude but differ in implementation and measurement scope, so one pooled throughput estimate would mask them.

Rates and ratios also need explicit denominators. Peng et al. (2026b) motivate asynchronous operation by contrasting one-to-three-second queries with a 5–20 Hz control-loop range. Their offline whole-model comparison reports median end-to-end latency from 1.7 s for Gemini 2.5 Flash-Lite to 22 s for GPT-5 Nano, with the best ADE from Gemini 3 Flash at 8.1 s; because entire models differ, the comparison confounds model size with reasoning cost. In deployment, requests are issued at 1 Hz while candidates and waypoints update at 5 Hz, and the ratio $1/5 = 0.2$ counts requests per planner update, not a measured command-level b. Issued requests, completed responses and accepted advice need not be equal in number.

Each study handles latency in its own way. Sathyamoorthy et al. (2024) cap the robot at 0.3 m/s explicitly so that navigation stays uninterrupted during model inference. Ahmadi et al. (2026) add a *Pause* condition (55.17% success) to the *Freeze* and *Realistic* results of Section 5.3; Freeze stops the scene during VLM inference, while the Pause and Realistic update schedules are unspecified. Weerakoon et al. (2025) report that a competing VLM baseline handles static crosswalk cues but misses dynamic cues from stop gestures and pedestrians, and attribute the misses to its low inference rate, although latency and scene understanding are not varied independently. Their own system, with a small segmentation model in the local stack, runs at a reported overall 9.5 Hz, whereas its internet-based GPT-4o queries take about 6 s during navigation (3.59 s in an offline landmark test).

Gate-use counts are reported in incompatible units. HUMA reports threshold-dependent total call counts (390–3,250); in Dual-process, more than half of decisions use the fast path (Lee et al. 2026). CATNAV additionally separates scene-query and path-selector counts from query rate, and its cached configuration still makes online model calls (Section 3.4).

We propose a minimum reporting set for deployment comparisons (Figure 5):

1. **Model and execution setting:** model identity, disclosed size and quantization, deployment hardware and network use.
2. **Latency and control:** observation-to-command latency distribution and command rate, with measurement boundaries and whether timing was measured during operation or offline.
3. **Invocation ratio:** model requests and control decisions counted over the same observation window, giving b wherever its denominator is defined, together with the request rate and per-request model latency.
4. **Asynchronous operation:** completed-response rate and output age at use.

Together, these items distinguish request capacity, control timing and stale advice. Invocation frequency alone leaves b undetermined, and component timing can inform a deployment trade-off only when its measurement boundaries are stated.

6 Open challenges and future directions

6.1 Real-time reasoning

Two runtime questions remain open: whether task-specific social reasoning degrades gracefully under compression, and whether learned invocation can improve on hand-specified gates.

The compression comparisons reviewed here involve unmatched models. In a non-social navigation comparison on a Jetson Orin NX, TinyLLaVA 1.1B performs poorly, with accuracy around 20% across the tested maximum token limits of 100, 150 and 300, whereas Gemma-3-4B, from another model family, performs substantially better (Lee et al. 2026). A fine-tuned 5.74B model scores higher on text similarity than tested frontier models on a held-out split of its task-specific training corpus (Kawabata et al. 2025). Isolating compression effects requires matched comparisons within one model family, including social situations outside the task-specific training distribution.

Hand-specified gates in Section 3.5 (Ahmadi et al. 2026; Lee et al. 2026) yield setting-specific invocation counts, reported in different units (Section 5.5), and do not identify a general optimum. The learned gate of Raj et al. (2024) predicts thresholded deviation from a demonstration, a target that excludes component cost and general collision risk. A concrete study would compare learned and hand-specified VLM gates under common observation and resource constraints.

6.2 Spatial grounding

The SocialNav-SUB rule baseline's advantage (Section 4.3) raises a representation question: which input (numbered regions, detected entities, segmentation-derived text or learned spatial tokens) best supports geometric accuracy and social utility within latency constraints while allowing reliable uncertainty estimates. That advantage is, however, specific to the benchmark's questions and metric and does not show that every social judgment reduces to geometry (Munje et al. 2025).

Geometry enters at different stages. CoNVOI supplies numbered regions at inference, AutoSpatial uses a prescribed spatial vocabulary during data construction and prompting (Sections 3.4, 3.6.3 and 3.8), and SpatialVLM and HumAIN supply geometric supervision during training (Chen et al. 2024a; Song et al. 2026). These interventions differ in the geometry they supply, not only in when they supply it. Input-representation tests therefore need one trained model under matched observations and tasks and should change one input element at a time; HUMA's depth/RGB test (Section 5.3) changes several at once. Training-supervision tests need matched backbones, initialization and training budgets, then a common evaluation. Both should score representation probes apart from executed behavior, as SALSA's action-head results show (Wang et al. 2026a).

Marker-dependent systems also need a test of whether the model reads the scaffold correctly, alongside end-task evaluation. In a manually analyzed random sample of 140 GPT-4V errors (ten per task), Fu et al. (2025) attribute 20% to incorrect circle locations and report sensitivity to marker color and size. SocialNav-SUB's bird's-eye-view ablation (Munje et al. 2025) shows that a scaffold's benefit also varies by model (Section 4.3). Treating a scaffold as a uniform remedy requires showing that the target model uses it.

6.3 Safety, hallucination and adversarial robustness

Following the specification and robustness framing of Amodei et al. (2016), the open problem is verifying whole-stack safety, social norms included. Several systems in online patterns II–IV keep a conventional planner or controller downstream, but any formal guarantees it provides can be claimed for the whole perception–reasoning–control stack only when their transfer to that stack is demonstrated. SafeVLA's constrained learning addresses environment- and object-level hazards in an adjacent VLA setting (Zhang et al. 2025) and has not been tested on queuing, group-intrusion or deference norms. Methods for uncertainty-aware deferral (Ren et al. 2023), for evaluating and mitigating hallucination (Ji et al. 2023) and for adversarial testing of language-controlled robots (Robey et al. 2025) come from outside this navigation corpus.

Within the corpus, SafeVLA also reports out-of-distribution and perturbation tests, which differ from attacks under an explicit adversary model and from hallucination-specific diagnosis. TrustNavGPT's specified textual attack, which replaces uncertain transcript phrases with more definite wording while vocal information remains available, lowers reported preferred-option selection from 70.46% to 46.90% (Sun et al. 2024b). TrustNavGPT's requests for human help when visual evidence is insufficient, the supervisor's pauses in G2-Nav's evaluation protocol (Liao et al. 2026) and the takeovers in Slow Brain's runs (Peng et al. 2026b) are not quantitatively validated handovers of control triggered by a calibrated uncertainty threshold. Studies should report attack-specific, hallucination and decision-deferral evaluations separately from each other and from any generic robustness score.

6.4 Sim-to-real transfer and data scarcity

It is unclear how well policies trained on reused, generated or distilled data handle unseen real pedestrian behavior. Several systems reuse SCAND-derived annotations or MuSoHu. GROVE's scenario generation (Nguyen et al. 2026) and NavIsaacLab's data-driven simulation (Xia et al. 2026) expand controllable evaluation conditions and make generator assumptions explicit; their real-world transfer is an empirical question. Distilled policies should be evaluated for out-of-distribution detection, uncertainty calibration and runtime latency, since teacher coverage limits and generated-label errors can persist after transfer.

6.5 Evaluation standardization

A paired evaluation should run the same models and interfaces on scene-understanding and navigation tasks under matched observations, without assuming that every offline or non-robot system can run unchanged on one platform. A preregistered analysis could then ask whether diagnostic accuracy predicts task success, safety or human-rated appropriateness. Human-response measures need clearer reporting and greater reuse of instruments with documented reliability and validity evidence (Bartneck et al. 2009). Since human-response scores cannot be pooled across studies (Section 5.1), each study's result must stand on its own, and its sample size should be justified accordingly.

6.6 Personalization and cultural variation

Preferred interpersonal distances vary across the countries sampled by Sorokowska et al. (2017), whereas SCAND was collected in Austin, Texas (Karnan et al. 2022), and MuSoHu on the George Mason University campus and in the Washington, DC, metropolitan area (Nguyen et al. 2023), local settings that bound claims of geographic coverage. Song et al. (2025b) note that a prompt can specify right- or left-side passing according to culture, although their reported navigation experiments do not compare such prompts across cultures. User-aware VLMs condition responses on visual and contextual user cues in social human–robot interaction, evidence adjacent to navigation (Rahimi et al. 2025). Culturally stratified navigation evaluation should compare user groups and test language instructions against reconfigured geometric or learned alternatives. No reviewed study tests longitudinal preference adaptation.

7 Conclusion

This survey organizes 34 central cases (25 with quantitative navigation evaluations and nine with answer-level or offline evaluations) and seven mechanism extensions by where the model enters the navigation stack, and reads their comparisons for system utility, component contribution and online necessity. Within the scope defined in Section 2.4, the comparisons show more clearly where online inference can help than whether it is required: they support conditional utility and make no general case for keeping or removing online inference.

RQ1: Model role. Models supply scene interpretations, costs, subgoals or reference paths, or actions; coordinate navigation modules; or provide offline supervision and scenarios. Each role locates the contribution to be tested, while invocation timing remains a separate design choice.

RQ2: Tested contribution. In the reported encounters, cue-aware systems outperformed their baselines when the deciding cue was missing from occupancy, and in one offline selection study a model selector helped on hard-to-rank candidates, though only among those the proposer supplied. The clearest cue-access results come from whole-system comparisons; removal tests show dependence within one implementation, and named replacements are rare. The evidence therefore establishes system utility and component dependence more often than necessity. Answer-level gains have

yet to be linked to executed navigation, and representation-level gains have reached only small robot trials.

RQ3: Online execution. In one supervised deployment, fusion with current plans kept delayed advice useful, and training-time transfer can retain benefits without querying the teacher at deployment. The reviewed comparisons do not settle whether a benefit requires online execution, because no study has set online, cached and distilled alternatives against one another with student computation included. Benefits are also not guaranteed: HUMA's success dropped to 56.90% under its *Realistic* timing condition from 62.07% with the scene frozen during inference, and, in one offline-transfer case, demonstrations ranked by a language model appear to have lowered navigation success (Sections 3.5, 3.7 and 5.3). Numeric timing entries exist for 16 of the 34 central cases, but different timing boundaries, measurement windows and asynchronous operation prevent a common deployment-cost estimate; the minimum reporting set of Section 5.5 addresses this gap. Table 8 connects these results to conditional design judgments.

Closing these gaps requires head-to-head comparisons, run with explicit observation access, training resources and execution deadlines and with annotation and adaptation costs counted. Evaluation should measure social behavior and human response alongside task success.

Acknowledgments

We acknowledge the use of Claude (Anthropic, https://claude.ai) for assistance with writing and English proofreading. All AI-generated content was reviewed and validated by the authors.

Statements and Declarations

Funding

This research was supported by the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIT) (Nos. RS-2024-00461583, RS-2024-00411007 and RS-2026-25618697) and by the Korea Basic Science Institute (National Research Facilities and Equipment Center) grant funded by the Ministry of Science and ICT (No. RS-2025-00564593).

Competing interests

The authors declare no competing interests.

Author contributions

Sungsan Park conducted the literature review and drafted the manuscript. Young-Sik Shin and Sanghyun Kim supervised the work and revised the manuscript.

Data availability

No datasets were generated or analyzed during the current study.

References

Ahmadi A, Rahimi H, Jacquet Crétides A, et al (2026) Think when it matters: Conditional VLM reasoning for social navigation with RL policies. https://doi.org/10.48550/arXiv.2607.10991, arXiv:2607.10991v1

Alahi A, Goel K, Ramanathan V, et al (2016) Social LSTM: Human trajectory prediction in crowded spaces. In: IEEE Conf Comput Vis Pattern Recognit, pp 961–971, https://doi.org/10.1109/CVPR.2016.110

Alyassi R, Cadena C, Riener R, et al (2025) Social robot navigation: A review and benchmarking of learning-based methods. Front Robot AI 12:1658643. https://doi.org/10.3389/frobt.2025.1658643

Amodei D, Olah C, Steinhardt J, et al (2016) Concrete problems in AI safety. https://doi.org/10.48550/arXiv.1606.06565, arXiv:1606.06565

Anderson P, Chang A, Chaplot DS, et al (2018a) On evaluation of embodied navigation agents. https://doi.org/10.48550/arXiv.1807.06757, arXiv:1807.06757

Anderson P, Wu Q, Teney D, et al (2018b) Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In: IEEE/CVF Conf Comput Vis Pattern Recognit, pp 3674–3683, https://doi.org/10.1109/CVPR.2018.00387

Bai J, Bai S, Yang S, et al (2023) Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. https://doi.org/10.48550/arXiv.2308.12966, arXiv:2308.12966v3

Bartneck C, Kulić D, Croft E, et al (2009) Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. Int J Soc Robot 1(1):71–81. https://doi.org/10.1007/s12369-008-0001-3

Bejarano Sepulveda EJ, Bo V, Sanfeliu A, et al (2026) Perception–Awareness–Decision: Socially-aware robot navigation and interaction. In: Companion ACM/IEEE Int Conf Hum-Robot Interact, pp 258–262, https://doi.org/10.1145/3776734.3794394

van den Berg J, Guy SJ, Lin M, et al (2011) Reciprocal n-body collision avoidance. In: Robotics Research. Springer, Berlin, p 3–19, https://doi.org/10.1007/978-3-642-19457-3_1

Biswas A, Wang A, Silvera G, et al (2022) SocNavBench: A grounded simulation testing framework for evaluating social navigation. ACM Trans Hum-Robot Interact

11(3):1–24. https://doi.org/10.1145/3476413

Brščić D, Kanda T, Ikeda T, et al (2013) Person tracking in large public spaces using 3-D range sensors. IEEE Trans Hum-Mach Syst 43(6):522–534. https://doi.org/10.1109/THMS.2013.2283945

Cai R, Yamasaki T, Xiao L (2026) Vision-language models for deployable social robot navigation: Bridging semantic reasoning and low-level control. https://doi.org/10.48550/arXiv.2606.28760, arXiv:2606.28760

Carpinella CM, Wyman AB, Perez MA, et al (2017) The Robotic Social Attributes Scale (RoSAS): Development and validation. In: ACM/IEEE Int Conf Hum-Robot Interact, pp 254–262, https://doi.org/10.1145/2909824.3020208

Chen B, Xu Z, Kirmani S, et al (2024a) SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. In: IEEE/CVF Conf Comput Vis Pattern Recognit, pp 14455–14465, https://doi.org/10.1109/CVPR52733.2024.01370

Chen C, Liu Y, Kreiss S, et al (2019) Crowd-robot interaction: Crowd-aware robot navigation with attention-based deep reinforcement learning. In: IEEE Int Conf Robot Autom, pp 6015–6022, https://doi.org/10.1109/ICRA.2019.8794134

Chen J, Li Y, Jiang P, et al (2026a) LISN: Language-instructed social navigation with VLM-based controller modulating. In: IEEE Int Conf Robot Autom, pp 1855–1862, https://doi.org/10.1109/icra57385.2026.11695757, arXiv:2512.09920

Chen S, Guhur PL, Tapaswi M, et al (2022) Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In: IEEE/CVF Conf Comput Vis Pattern Recognit, pp 16516–16526, https://doi.org/10.1109/CVPR52688.2022.01604

Chen Z, Wu J, Wang W, et al (2024b) InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: IEEE/CVF Conf Comput Vis Pattern Recognit, pp 24185–24198, https://doi.org/10.1109/CVPR52733.2024.02283

Chen Z, Guo Y, Chu Z, et al (2026b) SocialNav: Training human-inspired foundation model for socially-aware embodied navigation. In: IEEE/CVF Conf Comput Vis Pattern Recognit, pp 28796–28806, arXiv:2511.21135

Cheng Z, Tu Y, Li R, et al (2025) EmbodiedEval: Evaluate multimodal LLMs as embodied agents. https://doi.org/10.48550/arXiv.2501.11858, arXiv:2501.11858

Curtis S, Best A, Manocha D (2016) Menge: A modular framework for simulating crowd movement. Collect Dyn 1:A1. https://doi.org/10.17815/CD.2016.1

Dai H, Wang X, Wang L, et al (2026) DPed-VLN: A benchmark for socially compliant vision-and-language navigation in dynamic pedestrian environments. https://doi.org/10.48550/arXiv.2609.21504, arXiv:2609.21504

Dong Y, Wu F, He Q, et al (2026) HA-VLN 2.0: An open benchmark and leaderboard for human-aware navigation in discrete and continuous environments with dynamic multi-human interactions. https://doi.org/10.48550/arXiv.2503.14229, arXiv:2503.14229

Dosovitskiy A, Beyer L, Kolesnikov A, et al (2021) An image is worth 16×16 words: Transformers for image recognition at scale. In: Int Conf Learn Represent, URL https://openreview.net/forum?id=YicbFdNTTy

Dragan AD, Lee KCT, Srinivasa SS (2013) Legibility and predictability of robot motion. In: ACM/IEEE Int Conf Hum-Robot Interact, pp 301–308, https://doi.org/10.1109/HRI.2013.6483603

Du M, Wu B, Li Z, et al (2024) EmbSpatial-Bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In: Proc 62nd Annu Meet Assoc Comput Linguist (Vol 2: Short Papers), pp 346–355, https://doi.org/10.18653/v1/2024.acl-short.33

Ehsanpour M, Saleh F, Savarese S, et al (2022) JRDB-Act: A large-scale dataset for spatio-temporal action, social group and activity detection. In: IEEE/CVF Conf Comput Vis Pattern Recognit, pp 20951–20960, https://doi.org/10.1109/CVPR52688.2022.02031

Elnoor M, Weerakoon K, Seneviratne G, et al (2025) ViLAM: Distilling vision-language reasoning into attention maps for social robot navigation. https://doi.org/10.48550/arXiv.2503.09820, arXiv:2503.09820v2

Everett M, Chen YF, How JP (2021) Collision avoidance in pedestrian-rich environments with deep reinforcement learning. IEEE Access 9:10357–10377. https://doi.org/10.1109/ACCESS.2021.3050338

Fang Z, Xiao A, Hsu D, et al (2026) From obstacles to etiquette: Robot social navigation with VLM-informed path selection. IEEE Robot Autom Lett 11(4):3947–3954. https://doi.org/10.1109/lra.2026.3662586, arXiv:2602.09002v1

Firoozi R, Tucker J, Tian S, et al (2025) Foundation models in robotics: Applications, challenges, and the future. Int J Robot Res 44(5):701–739. https://doi.org/10.1177/02783649241281508

Fox D, Burgard W, Thrun S (1997) The dynamic window approach to collision avoidance. IEEE Robot Autom Mag 4(1):23–33. https://doi.org/10.1109/100.580977

Francis A, Pérez-D'Arpino C, Li C, et al (2025) Principles and guidelines for evaluating social robot navigation algorithms. ACM Trans Hum-Robot Interact 14(2):1–65. https://doi.org/10.1145/3700599

Fu X, Hu Y, Li B, et al (2025) BLINK: Multimodal large language models can see but not perceive. In: Eur Conf Comput Vis, pp 148–166, https://doi.org/10.1007/978-3-031-73337-6_9, arXiv:2404.12390

Gao Y, Huang CM (2022) Evaluation of socially-aware robot navigation. Front Robot AI 8:721317. https://doi.org/10.3389/frobt.2021.721317

Gemini Team (2023) Gemini: A family of highly capable multimodal models. https://doi.org/10.48550/arXiv.2312.11805, arXiv:2312.11805v5

Goffman E (1963) Behavior in Public Places: Notes on the Social Organization of Gatherings. Free Press, New York

Goffman E (1971) Relations in Public: Microstudies of the Public Order. Basic Books, New York

Gong Z, Hu T, Qiu R, et al (2025) From cognition to precognition: A future-aware framework for social navigation. In: IEEE Int Conf Robot Autom, pp 9122–9129, https://doi.org/10.1109/icra55743.2025.11127910, arXiv:2409.13244v2

Gu J, Stefani E, Wu Q, et al (2022) Vision-and-language navigation: A survey of tasks, methods, and future directions. In: Proc 60th Annu Meet Assoc Comput Linguist (Vol 1: Long Papers), pp 7606–7623, https://doi.org/10.18653/v1/2022.acl-long.524

Gupta A, Johnson J, Fei-Fei L, et al (2018) Social GAN: Socially acceptable trajectories with generative adversarial networks. In: IEEE/CVF Conf Comput Vis Pattern Recognit, pp 2255–2264, https://doi.org/10.1109/CVPR.2018.00240

Hall ET (1966) The Hidden Dimension. Doubleday, Garden City, NY

Hart SG, Staveland LE (1988) Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In: Human Mental Workload. North-Holland, Amsterdam, p 139–183, https://doi.org/10.1016/S0166-4115(08)62386-9

Helbing D, Molnár P (1995) Social force model for pedestrian dynamics. Phys Rev E 51(5):4282–4286. https://doi.org/10.1103/PhysRevE.51.4282

Hinton G, Vinyals O, Dean J (2015) Distilling the knowledge in a neural network. https://doi.org/10.48550/arXiv.1503.02531, arXiv:1503.02531

Hirose N, Glossop C, Shah D, et al (2026) OmniVLA: An omni-modal vision-language-action model for robot navigation. In: IEEE Int Conf Robot Autom, pp 12651–12658, https://doi.org/10.1109/icra57385.2026.11697566, arXiv:2509.19480

Holtz J, Biswas J (2022) SocialGym: A framework for benchmarking social robot navigation. In: IEEE/RSJ Int Conf Intell Robots Syst, pp 11246–11252, https://doi.org/10.1109/IROS47612.2022.9982021

Hu EJ, Shen Y, Wallis P, et al (2022) LoRA: Low-rank adaptation of large language models. In: Int Conf Learn Represent, URL https://openreview.net/forum?id=nZeVKeeFYf9

Hu T, Gong Z, Kong L, et al (2026) NavThinker: Action-conditioned world models for coupled prediction and planning in social navigation. https://doi.org/10.48550/arXiv.2603.15359, arXiv:2603.15359v2

Huang W, Yang D, Li J, et al (2026a) SONG: A photorealistic 3D Gaussian simulation platform for benchmarking social navigation. https://doi.org/10.48550/arXiv.2607.25219, arXiv:2607.25219

Huang Z, Zhang Y, Liu J, et al (2026b) TIC-VLA: A think-in-control vision-language-action model for robot navigation in dynamic environments. In: Proc 43rd Int Conf Mach Learn, Proc Mach Learn Res, vol 306. PMLR, pp 48344–48367, URL https://proceedings.mlr.press/v306/huang26cd.html, arXiv:2602.02459

Hwang J, Hong SE, Kim J, et al (2026) CrowdVLA: Embodied vision-language-action agents for context-aware crowd simulation. https://doi.org/10.48550/arXiv.2604.05525, arXiv:2604.05525v1

Ichter B, Brohan A, Chebotar Y, et al (2023) Do as I can, not as I say: Grounding language in robotic affordances. In: Proc 6th Conf Robot Learn, Proc Mach Learn Res, vol 205. PMLR, pp 287–318, URL https://proceedings.mlr.press/v205/ichter23a.html

Ji Z, Lee N, Frieske R, et al (2023) Survey of hallucination in natural language generation. ACM Comput Surv 55(12):1–38. https://doi.org/10.1145/3571730

Joosse M, Lohse M, Evers V (2014) Lost in proxemics: Spatial behavior for cross-cultural HRI. In: HRI Worksh Cult-Aware Robot, URL https://www.researchgate.net/publication/302477938_Lost_in_Proxemics_Spatial_Behavior_for_Cross-Cultural_HRI

Karnan H, Nair A, Xiao X, et al (2022) Socially compliant navigation dataset (SCAND): A large-scale dataset of demonstrations for social navigation. IEEE Robot Autom Lett 7(4):11807–11814. https://doi.org/10.1109/LRA.2022.3184025

Kashmar O, Rollo F, Kashiri N, et al (2026) Human-aware robot navigation: Problem formulation, approaches, challenges. SSRN preprint, https://doi.org/10.2139/ssrn.6486782

Kathuria T (2025) Motion and behavior planning for socially assistive robots. PhD thesis, University of Michigan, https://doi.org/10.7302/25667

Kawabata T, Zhang X, Xiao L (2025) SocialNav-MoE: A mixture-of-experts vision language model for socially compliant navigation with reinforcement fine-tuning. https://doi.org/10.48550/arXiv.2512.14757, arXiv:2512.14757

Kim MJ, Pertsch K, Karamcheti S, et al (2025) OpenVLA: An open-source vision-language-action model. In: Proc 8th Conf Robot Learn, Proc Mach Learn Res, vol 270. PMLR, pp 2679–2713, URL https://proceedings.mlr.press/v270/kim25c.html

Kirby R, Simmons R, Forlizzi J (2009) COMPANION: A constraint-optimizing method for person-acceptable navigation. In: IEEE Int Symp Robot Hum Interact Commun, pp 607–612, https://doi.org/10.1109/ROMAN.2009.5326271

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

Kong Y, Song D, Liang J, et al (2025) AutoSpatial: Visual-language reasoning for social robot navigation through efficient spatial reasoning learning. In: IEEE/RSJ Int Conf Intell Robots Syst, pp 11298–11304, https://doi.org/10.1109/iros60139.2025.11245987, arXiv:2503.07557

Kothari P, Kreiss S, Alahi A (2022) Human trajectory forecasting in crowds: A deep learning perspective. IEEE Trans Intell Transp Syst 23(7):7386–7400. https://doi.org/10.1109/TITS.2021.3069362

Krantz J, Wijmans E, Majumdar A, et al (2020) Beyond the nav-graph: Vision-and-language navigation in continuous environments. In: Eur Conf Comput Vis, pp 104–120, https://doi.org/10.1007/978-3-030-58604-1_7

Kretzschmar H, Spies M, Sprunk C, et al (2016) Socially compliant mobile robot navigation via inverse reinforcement learning. Int J Robot Res 35(11):1289–1307. https://doi.org/10.1177/0278364915619772

Kruse T, Pandey AK, Alami R, et al (2013) Human-aware robot navigation: A survey. Robot Auton Syst 61(12):1726–1743. https://doi.org/10.1016/j.robot.2013.05.007

Lee J, Shin H, Ko J (2026) IROS: A dual-process architecture for real-time VLM-based indoor navigation. https://doi.org/10.48550/arXiv.2601.21506, arXiv:2601.21506v1

Lerner A, Chrysanthou Y, Lischinski D (2007) Crowds by example. Comput Graph Forum 26(3):655–664. https://doi.org/10.1111/j.1467-8659.2007.01089.x

Li C, Xia F, Martín-Martín R, et al (2022) iGibson 2.0: Object-centric simulation for robot learning of everyday household tasks. In: Proc 5th Conf Robot Learn, Proc Mach Learn Res, vol 164. PMLR, pp 455–465, URL https://proceedings.mlr.press/

v164/li22b.html

Li C, Sun X, Chen P, et al (2024) CoNav: A benchmark for human-centered collaborative navigation. https://doi.org/10.48550/arXiv.2406.02425, arXiv:2406.02425

Li J, Li D, Savarese S, et al (2023) BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proc 40th Int Conf Mach Learn, Proc Mach Learn Res, vol 202. PMLR, pp 19730–19742, URL https://proceedings.mlr.press/v202/li23q.html

Li T (2025) Robot mediation and navigation: Navigating multi-user conflicts with large language model and reinforcement learning. Master's thesis, KTH Royal Institute of Technology, URL https://kth.diva-portal.org/smash/get/diva2:2037655/FULLTEXT01.pdf, TRITA-EECS-EX-2025:908

Liang J, Weerakoon K, Song D, et al (2025) MOSU: Autonomous long-range robot navigation with multi-modal scene understanding. https://doi.org/10.48550/arXiv.2507.04686, arXiv:2507.04686

Liao Y, Lan Y, Yang Y, et al (2026) G2-Nav: Grounded and guarded vision-language costmaps for robot social navigation. https://doi.org/10.48550/arXiv.2607.16956, arXiv:2607.16956v2

Liu H, Li C, Wu Q, et al (2023a) Visual instruction tuning. In: Adv Neural Inf Process Syst, vol 36. Curran Associates, Inc., pp 34892–34916, https://doi.org/10.52202/075280-1516

Liu S, Chang P, Huang Z, et al (2023b) Intention aware robot crowd navigation with attention-based interaction graph. In: IEEE Int Conf Robot Autom, pp 12015–12021, https://doi.org/10.1109/ICRA48891.2023.10160660

Liu S, Zeng Z, Ren T, et al (2025a) Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In: Eur Conf Comput Vis, pp 38–55, https://doi.org/10.1007/978-3-031-72970-6_3

Liu X, Farid A, Amano T, et al (2025b) LLM-powered embodied intelligence for socially-aware robot navigation in human-robot interaction. In: Proc 19th Int Symp Spatial Temporal Data, pp 145–148, https://doi.org/10.1145/3748777.3748807

Liu X, Farid A, Ukyoh R, et al (2025c) LLM-driven adaptive autonomous robot navigation via multimodal fusion for diverse environments. In: IEEE Intell Veh Symp, pp 2361–2368, https://doi.org/10.1109/IV64158.2025.11097694

Liu Z, Zhang X, Wang Z, et al (2025d) MUSON: A reasoning-oriented multimodal dataset for socially compliant navigation in urban environments. https://doi.org/10.48550/arXiv.2512.22867, arXiv:2512.22867v2

Luo S, Sun P, Zhu J, et al (2025) GSON: A group-based social navigation framework with large multimodal model. IEEE Robot Autom Lett 10(10):9646–9653. https://doi.org/10.1109/lra.2025.3595038, arXiv:2409.18084v3

Manso LJ, Nuñez P, Calderita LV, et al (2020) SocNav1: A dataset to benchmark and learn social navigation conventions. Data 5(1):7. https://doi.org/10.3390/data5010007

Marpally SR, Goyal P, Soh H (2024) SocRATES: Towards automated scenario-based testing of social navigation algorithms. https://doi.org/10.48550/arXiv.2412.19595, arXiv:2412.19595

Martín-Martín R, Patel M, Rezatofighi H, et al (2023) JRDB: A dataset and benchmark of egocentric robot visual perception of humans in built environments. IEEE Trans Pattern Anal Mach Intell 45(6):6748–6765. https://doi.org/10.1109/TPAMI.2021.3070543

Mavrogiannis C, Baldini F, Wang A, et al (2023) Core challenges of social robot navigation: A survey. ACM Trans Hum-Robot Interact 12(3):1–39. https://doi.org/10.1145/3583741

Mavrogiannis CI, Thomason WB, Knepper RA (2018) Social momentum: A framework for legible navigation in dynamic multi-agent environments. In: ACM/IEEE Int Conf Hum-Robot Interact, pp 361–369, https://doi.org/10.1145/3171221.3171255

Mirsky R, Xiao X, Hart J, et al (2024) Conflict avoidance in social navigation: A survey. ACM Trans Hum-Robot Interact 13(1):1–36. https://doi.org/10.1145/3647983

Möller R, Furnari A, Battiato S, et al (2021) A survey on human-aware robot navigation. Robot Auton Syst 145:103837. https://doi.org/10.1016/j.robot.2021.103837

Moussaïd M, Perozo N, Garnier S, et al (2010) The walking behaviour of pedestrian social groups and its impact on crowd dynamics. PLoS One 5(4):e10047. https://doi.org/10.1371/journal.pone.0010047

Moussaïd M, Helbing D, Theraulaz G (2011) How simple rules determine pedestrian behavior and crowd disasters. Proc Natl Acad Sci USA 108(17):6884–6888. https://doi.org/10.1073/pnas.1016507108

Mumm J, Mutlu B (2011) Human-robot proxemics: Physical and psychological distancing in human-robot interaction. In: ACM/IEEE Int Conf Hum-Robot Interact, pp 331–338, https://doi.org/10.1145/1957656.1957786

Munje MJ, Tang C, Liu S, et al (2025) SocialNav-SUB: Benchmarking VLMs for scene understanding in social robot navigation. In: Proc 9th Conf Robot Learn, Proc Mach Learn Res, vol 305. PMLR, pp 1120–1143, URL https://proceedings.

mlr.press/v305/munje25a.html, arXiv:2509.08757

Narasimhan S, Tan AH, Choi D, et al (2025) OLiVia-Nav: An online lifelong vision language approach for mobile robot social navigation. In: IEEE Int Conf Robot Autom, pp 9130–9137, https://doi.org/10.1109/icra55743.2025.11128004, arXiv:2409.13675v2

Nazeri M, Wang J, Payandeh A, et al (2024) VANP: Learning where to see for navigation with self-supervised vision-action pre-training. In: IEEE/RSJ Int Conf Intell Robots Syst, pp 2741–2746, https://doi.org/10.1109/iros58592.2024.10802451, arXiv:2403.08109

Nguyen DM, Nazeri M, Payandeh A, et al (2023) Toward human-like social robot navigation: A large-scale, multi-modal, social human navigation dataset. In: IEEE/RSJ Int Conf Intell Robots Syst, pp 7442–7447, https://doi.org/10.1109/IROS55552.2023.10342447

Nguyen DT, Shcherbyna V, Do DA, et al (2026) GROVE: Grounded pedestrian simulation via natural language for interactive social robot navigation. https://doi.org/10.48550/arXiv.2606.25504, arXiv:2606.25504

OpenAI (2023) GPT-4 technical report. https://doi.org/10.48550/arXiv.2303.08774, arXiv:2303.08774v6

Pan H, Huang S, Yang J, et al (2026) Robot navigation via foundation language models: A review. ACM Comput Surv 58(11):1–38. https://doi.org/10.1145/3802539

Park S, Shin YS, Kim S (2026) TACET: Context-appropriate acoustic-social navigation for quadrupeds. https://doi.org/10.48550/arXiv.2610.03828, arXiv:2610.03828

Payandeh A, Song D, Nazeri M, et al (2025) Social-LLaVA: Enhancing social robot navigation through human-language reasoning. In: IEEE/RSJ Int Conf Intell Robots Syst, pp 17192–17198, https://doi.org/10.1109/iros60139.2025.11247618, arXiv:2501.09024

Payandeh A, Pokhrel A, Song D, et al (2026) Narrate2Nav: Real-time visual navigation with implicit language reasoning in human-centric environments. In: IEEE Int Conf Robot Autom, pp 21696–21703, https://doi.org/10.1109/icra57385.2026.11696112, arXiv:2506.14233

Pellegrini S, Ess A, Schindler K, et al (2009) You'll never walk alone: Modeling social behavior for multi-target tracking. In: IEEE Int Conf Comput Vis, pp 261–268, https://doi.org/10.1109/ICCV.2009.5459260

Peng Y, Pan Y, He X, et al (2026a) FreeAskWorld: An interactive and closed-loop simulator for human-centric embodied AI. Proc AAAI Conf Artif Intell 40(2):926–934. https://doi.org/10.1609/aaai.v40i2.37061

Peng Z, He H, Li Q, et al (2026b) Slow brain, fast planner: Latency-resilient VLM-augmented urban navigation. https://doi.org/10.48550/arXiv.2606.20458, arXiv:2606.20458v1

Pérez-Higueras N, Otero R, Caballero F, et al (2023) HuNavSim: A ROS 2 human navigation simulator for benchmarking human-aware robot navigation. IEEE Robot Autom Lett 8(11):7130–7137. https://doi.org/10.1109/LRA.2023.3316072

Perille D, Truong A, Xiao X, et al (2020) Benchmarking metric ground navigation. In: IEEE Int Symp Saf Secur Rescue Robot, pp 116–121, https://doi.org/10.1109/SSRR50563.2020.9292572

Pirk S, Lee E, Xiao X, et al (2022) A protocol for validating social navigation policies. https://doi.org/10.48550/arXiv.2204.05443, arXiv:2204.05443

Potnis A, Affonso F, Gummadi S, et al (2026) CATNAV: Cached vision-language traversability for efficient zero-shot robot navigation. https://doi.org/10.48550/arXiv.2603.22800, arXiv:2603.22800

Puig X, Undersander E, Szot A, et al (2024) Habitat 3.0: A co-habitat for humans, avatars and robots. In: Int Conf Learn Represent, URL https://openreview.net/forum?id=4znwzG92CE

Quinlan S, Khatib O (1993) Elastic bands: Connecting path planning and control. In: IEEE Int Conf Robot Autom, pp 802–807, https://doi.org/10.1109/ROBOT.1993.291936

Radford A, Kim JW, Hallacy C, et al (2021) Learning transferable visual models from natural language supervision. In: Proc 38th Int Conf Mach Learn, Proc Mach Learn Res, vol 139. PMLR, pp 8748–8763, URL https://proceedings.mlr.press/v139/radford21a.html

Rafailov R, Sharma A, Mitchell E, et al (2023) Direct preference optimization: Your language model is secretly a reward model. In: Adv Neural Inf Process Syst, vol 36. Curran Associates, Inc., pp 53728–53741, https://doi.org/10.52202/075280-2338

Rahimi H, Bahaj A, Abrini M, et al (2025) USER-VLM 360: Personalized vision language models with user-aware tuning for social human-robot interactions. In: Proc 27th Int Conf Multimodal Interact, pp 326–336, https://doi.org/10.1145/3716553.3750767

Raj AH, Hu Z, Karnan H, et al (2024) Rethinking social robot navigation: Leveraging the best of two worlds. In: IEEE Int Conf Robot Autom, pp 16330–16337, https://doi.org/10.1109/ICRA57147.2024.10611710, arXiv:2309.13466

Rempe D, Luo Z, Peng XB, et al (2023) Trace and Pace: Controllable pedestrian animation via guided trajectory diffusion. In: IEEE/CVF Conf Comput Vis Pattern

Recognit, pp 13756–13766, https://doi.org/10.1109/cvpr52729.2023.01322

Ren AZ, Dixit A, Bodrova A, et al (2023) Robots that ask for help: Uncertainty alignment for large language model planners. In: Proc 7th Conf Robot Learn, Proc Mach Learn Res, vol 229. PMLR, pp 661–682, URL https://proceedings.mlr.press/v229/ren23a.html, arXiv:2307.01928

Rios-Martinez J, Spalanzani A, Laugier C (2015) From proxemics theory to socially-aware navigation: A survey. Int J Soc Robot 7(2):137–153. https://doi.org/10.1007/s12369-014-0251-1

Robey A, Ravichandran Z, Kumar V, et al (2025) Jailbreaking LLM-controlled robots. In: IEEE Int Conf Robot Autom, pp 11948–11956, https://doi.org/10.1109/icra55743.2025.11128119, arXiv:2410.13691

Robicquet A, Sadeghian A, Alahi A, et al (2016) Learning social etiquette: Human trajectory understanding in crowded scenes. In: Eur Conf Comput Vis, pp 549–565, https://doi.org/10.1007/978-3-319-46484-8_33

Rösmann C, Hoffmann F, Bertram T (2017) Integrated online trajectory planning and optimization in distinctive topologies. Robot Auton Syst 88:142–153. https://doi.org/10.1016/j.robot.2016.11.007

Rudas V, Kuzmenko D (2026) Can vision-language models assess proxemic risk from egocentric robot images? https://doi.org/10.48550/arXiv.2608.12515, arXiv:2608.12515

Rudenko A, Kucner TP, Swaminathan CS, et al (2020a) THÖR: Human-robot navigation data collection and accurate motion trajectories dataset. IEEE Robot Autom Lett 5(2):676–682. https://doi.org/10.1109/LRA.2020.2965416

Rudenko A, Palmieri L, Herman M, et al (2020b) Human motion trajectory prediction: A survey. Int J Robot Res 39(8):895–935. https://doi.org/10.1177/0278364920917446

Ruo A, Cacace J, Dalmau-Moreno M, et al (2025) An LLM-based architecture for socially intelligent robot navigation based on social cues. In: IEEE Int Conf Robot Hum Interact Commun, pp 1494–1500, https://doi.org/10.1109/ro-man63969.2025.11217843

Sacco V, Scofano L, Spinelli I, et al (2026) Robots ask the way: Communication-enabled social navigation. https://doi.org/10.48550/arXiv.2607.01044, arXiv:2607.01044

Salzmann T, Ivanovic B, Chakravarty P, et al (2020) Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In: Eur Conf Comput Vis, pp 683–700, https://doi.org/10.1007/978-3-030-58523-5_40

Samavi S, Han JR, Shkurti F, et al (2025) SICNav: Safe and interactive crowd navigation using model predictive control and bilevel optimization. IEEE Trans Robot 41:801–818. https://doi.org/10.1109/TRO.2024.3484634

Sathyamoorthy AJ, Weerakoon K, Elnoor M, et al (2024) CoNVOI: Context-aware navigation using vision language models in outdoor and indoor environments. In: IEEE/RSJ Int Conf Intell Robots Syst, pp 13837–13844, https://doi.org/10.1109/iros58592.2024.10802716, arXiv:2403.15637v1

Savva M, Kadian A, Maksymets O, et al (2019) Habitat: A platform for embodied AI research. In: IEEE/CVF Int Conf Comput Vis, pp 9338–9346, https://doi.org/10.1109/ICCV.2019.00943

Schreiter T, Rodrigues de Almeida T, Zhu Y, et al (2025) THÖR-MAGNI: A large-scale indoor motion capture recording of human movement and robot interaction. Int J Robot Res 44(4):568–591. https://doi.org/10.1177/02783649241274794

Shah D, Equi MR, Osiński B, et al (2023a) Navigation with large language models: Semantic guesswork as a heuristic for planning. In: Proc 7th Conf Robot Learn, Proc Mach Learn Res, vol 229. PMLR, pp 2683–2699, URL https://proceedings.mlr.press/v229/shah23c.html

Shah D, Osiński B, Ichter B, et al (2023b) LM-Nav: Robotic navigation with large pre-trained models of language, vision, and action. In: Proc 6th Conf Robot Learn, Proc Mach Learn Res, vol 205. PMLR, pp 492–504, URL https://proceedings.mlr.press/v205/shah23b.html

Shah D, Sridhar A, Bhorkar A, et al (2023c) GNM: A general navigation model to drive any robot. In: IEEE Int Conf Robot Autom, pp 7226–7233, https://doi.org/10.1109/ICRA48891.2023.10161227

Shah D, Sridhar A, Dashora N, et al (2023d) ViNT: A foundation model for visual navigation. In: Proc 7th Conf Robot Learn, Proc Mach Learn Res, vol 229. PMLR, pp 711–733, URL https://proceedings.mlr.press/v229/shah23a.html

Shcherbyna V, Kästner L, Diaz D, et al (2025) Arena 4.0: A comprehensive ROS2 development and benchmarking platform for human-centric navigation using generative-model-based environment generation. In: IEEE Int Conf Robot Autom, pp 9138–9144, https://doi.org/10.1109/icra55743.2025.11127635

Shi Z, Zhao E, Dennler NS, et al (2025) HRIBench: Benchmarking vision-language models for real-time human perception in human-robot interaction. https://doi.org/10.48550/arXiv.2506.20566, arXiv:2506.20566

Singamaneni PT, Bachiller-Burgos P, Manso LJ, et al (2024) A survey on socially aware robot navigation: Taxonomy and future challenges. Int J Robot Res 43(10):1533–1572. https://doi.org/10.1177/02783649241230562

Sisbot EA, Marin-Urias LF, Alami R, et al (2007) A human aware mobile robot motion planner. IEEE Trans Robot 23(5):874–883. https://doi.org/10.1109/TRO.2007.904911

Song CH, Blukis V, Tremblay J, et al (2025a) RoboSpatial: Teaching spatial understanding to 2D and 3D vision-language models for robotics. In: IEEE/CVF Conf Comput Vis Pattern Recognit, pp 15768–15780, https://doi.org/10.1109/cvpr52734.2025.01470

Song D, Liang J, Payandeh A, et al (2025b) VLM-Social-Nav: Socially aware robot navigation through scoring using vision-language models. IEEE Robot Autom Lett 10(1):508–515. https://doi.org/10.1109/LRA.2024.3511409

Song D, Le N, Chen J, et al (2026) HumAIN: Human-aware implicit social robot navigation. https://doi.org/10.48550/arXiv.2607.07357, arXiv:2607.07357v1

Sorokowska A, Sorokowski P, Hilpert P, et al (2017) Preferred interpersonal distances: A global comparison. J Cross-Cult Psychol 48(4):577–592. https://doi.org/10.1177/0022022117698039

Sotomi O, Kodi D, Arab A (2025) Transparent social navigation for autonomous mobile robots via vision-language models. In: IEEE Int Conf Robot Hum Interact Commun, pp 1646–1651, https://doi.org/10.1109/RO-MAN63969.2025.11217693

Sprague Z, Chandra R, Holtz J, et al (2023) SocialGym 2.0: Simulator for multi-agent social robot navigation in shared human spaces. https://doi.org/10.48550/arXiv.2303.05584, arXiv:2303.05584

Sridhar A, Shah D, Glossop C, et al (2024) NoMaD: Goal masked diffusion policies for navigation and exploration. In: IEEE Int Conf Robot Autom, pp 63–70, https://doi.org/10.1109/ICRA57147.2024.10610665

Sun X, Meng H, Chakraborty S, et al (2024a) Beyond Text: Utilizing vocal cues to improve decision making in LLMs for robot navigation tasks. Trans Mach Learn Res, URL https://openreview.net/forum?id=ojWtq4n7Ag, arXiv:2402.03494

Sun X, Zhang Y, Tang X, et al (2024b) TrustNavGPT: Modeling uncertainty to improve trustworthiness of audio-guided LLM-based robot navigation. In: IEEE/RSJ Int Conf Intell Robots Syst, pp 8794–8801, https://doi.org/10.1109/iros58592.2024.10801932, arXiv:2408.01867v1

Sun Z, Wu K, Fu C, et al (2025) TongSIM: A general platform for simulating intelligent machines. https://doi.org/10.48550/arXiv.2512.20206, arXiv:2512.20206

Szot A, Clegg A, Undersander E, et al (2021) Habitat 2.0: Training home assistants to rearrange their habitat. In: Adv Neural Inf Process Syst, vol 34. Curran Associates, Inc., pp 251–266, URL https://proceedings.neurips.cc/paper/2021/hash/

021bbc7ee20b71134d53e20206bd6feb-Abstract.html

Trautman P, Krause A (2010) Unfreezing the robot: Navigation in dense, interacting crowds. In: IEEE/RSJ Int Conf Intell Robots Syst, pp 797–803, https://doi.org/10.1109/IROS.2010.5654369

Trautman P, Ma J, Murray RM, et al (2015) Robot navigation in dense human crowds: Statistical models and experimental studies of human-robot cooperation. Int J Robot Res 34(3):335–356. https://doi.org/10.1177/0278364914557874

Tsoi N, Hussein M, Espinoza J, et al (2020) SEAN: Social environment for autonomous navigation. In: Proc 8th Int Conf Hum-Agent Interact, pp 281–283, https://doi.org/10.1145/3406499.3418760

Tsoi N, Xiang A, Yu P, et al (2022) SEAN 2.0: Formalizing and generating social situations for robot navigation. IEEE Robot Autom Lett 7(4):11047–11054. https://doi.org/10.1109/LRA.2022.3196783

Tsoi N, Romero J, Vázquez M (2024) How do robot experts measure the success of social robot navigation? In: Companion ACM/IEEE Int Conf Hum-Robot Interact, pp 1063–1066, https://doi.org/10.1145/3610978.3640636

Vaswani A, Shazeer N, Parmar N, et al (2017) Attention is all you need. In: Adv Neural Inf Process Syst, vol 30. Curran Associates, Inc., pp 5998–6008, URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Wang A, Sato D, Corzo Y, et al (2024a) TBD pedestrian data collection: Towards rich, portable, and large-scale natural pedestrian data. In: IEEE Int Conf Robot Autom, pp 637–644, https://doi.org/10.1109/ICRA57147.2024.10610335, arXiv:2309.17187v2

Wang H, Chen J, Huang W, et al (2024b) GRUtopia: Dream general robots in a city at scale. https://doi.org/10.48550/arXiv.2407.10943, arXiv:2407.10943

Wang P, Bai S, Tan S, et al (2024c) Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. https://doi.org/10.48550/arXiv.2409.12191, arXiv:2409.12191

Wang Q, Wu X, Shi G, et al (2026a) Act on what you see: Unlocking safe social navigation in vision-language-action models. https://doi.org/10.48550/arXiv.2606.10495, arXiv:2606.10495v2

Wang W, Obi I, Bera A, et al (2024d) Unifying large language model and deep reinforcement learning for human-in-loop interactive socially-aware navigation. https://doi.org/10.48550/arXiv.2403.15648, arXiv:2403.15648

Wang Z, Zhang X, Liu Z, et al (2026b) MAction-SocialNav: Multi-action socially compliant navigation via reasoning-enhanced prompt tuning. IEEE Robot Autom Lett 11(8):9747–9754. https://doi.org/10.1109/LRA.2026.3706951, arXiv:2512.21722

Weerakoon K, Elnoor M, Seneviratne G, et al (2025) BehAV: Behavioral rule guided autonomy using VLMs for robot navigation in outdoor scenes. In: IEEE Int Conf Robot Autom, pp 7044–7051, https://doi.org/10.1109/icra55743.2025.11127890, arXiv:2409.16484

Wen C, Liu Y, Bethala GCR, et al (2025) Socially-aware robot navigation enhanced by bidirectional natural language conversations using large language models. In: IEEE/RSJ Int Conf Intell Robots Syst, pp 15314–15321, https://doi.org/10.1109/iros60139.2025.11246738, arXiv:2409.04965

Xia B, Bao H, Zhu J, et al (2026) NavIsaacLab: Generating realistic crowd via parallel robot learning for benchmarking human-aware navigation. https://doi.org/10.48550/arXiv.2606.26265, arXiv:2606.26265

Xiao L, Yamasaki T (2026) Probing prompt design for socially compliant robot navigation with vision language models. https://doi.org/10.48550/arXiv.2601.14622, arXiv:2601.14622

Xiao L, Song D, Xiao X, et al (2026) E-SocialNav: Efficient socially compliant navigation with language models. In: IEEE Int Conf Acoust Speech Signal Process, pp 20077–20081, https://doi.org/10.1109/icassp55912.2026.11462076, arXiv:2603.20664

Xu H, Li T, Chen W, et al (2026) HCSG: Human-centric semantic-geometric reasoning for vision-language navigation. https://doi.org/10.48550/arXiv.2605.13321, arXiv:2605.13321

Yan Z, Duckett T, Bellotto N (2017) Online learning for human classification in 3D LiDAR-based tracking. In: IEEE/RSJ Int Conf Intell Robots Syst, pp 864–871, https://doi.org/10.1109/IROS.2017.8202247

Zhang B, Zhang Y, Ji J, et al (2025) SafeVLA: Towards safety alignment of vision-language-action model via constrained learning. In: Adv Neural Inf Process Syst, vol 38. Curran Associates, Inc., pp 170196–170234, https://doi.org/10.52202/085713-5128, arXiv:2503.03480

Zhang L, Hao X, Bu X, et al (2026a) Walk With Me: Long-horizon social navigation for human-centric outdoor assistance. https://doi.org/10.48550/arXiv.2604.26839, arXiv:2604.26839

Zhang X, Konno A, Yamasaki T, et al (2026b) Enhancing lightweight vision language models through group competitive learning for socially compliant navigation. https://doi.org/10.48550/arXiv.2603.11447, arXiv:2603.11447

Zhang X, Wang Z, Xiao L (2026c) SOPD-SocialNav: Selective on-policy distillation for vision-language social navigation. https://doi.org/10.48550/arXiv.2607.19850, arXiv:2607.19850

Zhang Y, Mehra SY, Duerstock BS, et al (2026d) SAATT Nav: A socially aware autonomous transparent transportation navigation framework for wheelchairs. https://doi.org/10.48550/arXiv.2603.13698, arXiv:2603.13698

Zhou C, Miao MC, Chen XR, et al (2022) Human-behaviour-based social locomotion model improves the humanization of social robots. Nat Mach Intell 4(11):1040–1052. https://doi.org/10.1038/s42256-022-00542-z

Zhou G, Hong Y, Wu Q (2024) NavGPT: Explicit reasoning in vision-and-language navigation with large language models. Proc AAAI Conf Artif Intell 38(7):7641–7649. https://doi.org/10.1609/aaai.v38i7.28597

Zitkovich B, Yu T, Xu S, et al (2023) RT-2: Vision-language-action models transfer web knowledge to robotic control. In: Proc 7th Conf Robot Learn, Proc Mach Learn Res, vol 229. PMLR, pp 2165–2183, URL https://proceedings.mlr.press/v229/zitkovich23a.html