Bridging Perception, Language, and Action: A Survey and Bibliometric Analysis of VLM & VLA Systems

Roman Dolgopolyi

r.dolgopolyi@acg.edu

The American College of Greece https://orcid.org/0009-0002-9884-8386

Anastasios Tsevas

The American College of Greece

Systematic Review

Keywords: Vision-Language Models, Vision-Language-Action, Multimodal Learning, Biblio- metric Analysis, Embodied AI, Foundation Models, Large Language Models

Posted Date: October 27th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7935378/v1

Additional Declarations: The authors declare no competing interests.

Bridging Perception, Language, and Action: A Survey and Bibliometric Analysis of VLM & VLA Systems

Roman Dolgopolyi
r.dolgopolyi@acg.edu

Anastasios Tsevas
a.tsevas@acg.edu

Abstract

Vision-Language Models (VLMs) and Vision-Language-Action (VLA) models mark a pivotal advancement in multimodal AI, integrating perception, language understanding, and physical action for embodied intelligence in real-world applications. This survey delivers an in-depth bibliometric analysis of 1,798 publications sourced from Scopus and OpenAlex, merging quantitative computational techniques with qualitative examination to examine the field's evolution, key themes, and persistent challenges. Through keyword frequency assessment, co-occurrence networks, temporal trend mapping, author collaboration visualization, and similarity-based clustering of titles and keywords, we uncover exponential publication growth since 2022—spanning a 50-fold increase—with core themes in language models, visual languages, and action recognition driving unified multimodal architectures. The analysis highlights ten keyword clusters centered on multimodal integration, robotic learning, and foundation models, alongside ten title clusters emphasizing applications from robotic navigation and video understanding to generalist agents and web navigation. Drawn author collaboration maps highlight geographic dominance by U.S. and Chinese institutions, which implies risks in technology governance and safety oversight. The qualitative review of seven high-impact papers traces VLA progression from closed-source fine-tuning to open-source transfer learning, rigorous grounding evaluations, cultural bias assessments, and deployments in virtual agents, autonomous driving, and robotic manipulation. Despite technical maturity, critical gaps persist in safety mechanisms, adversarial robustness, and standardized evaluation, urging prioritized research for responsible deployment.

Keywords: Vision-Language Models, Vision-Language-Action, Multimodal Learning, Bibliometric Analysis, Embodied AI, Foundation Models, Large Language Models

1 Introduction

Visual Language Models (VLMs) mark a significant evolution in artificial intelligence, merging computer vision and natural language processing to enable machines to process and generate information across both modalities [1, 2]. Advancements in transformer-based architectures [3, 4], large-scale pretraining, and contrastive learning techniques such as CLIP [2] have propelled VLMs to the forefront of multimodal AI research. Recent extensions into Vision-Language-Action (VLA) models represent a further leap, integrating physical action capabilities that enable embodied agents—such as robots—to interact with their environments based on visual and linguistic inputs [5, 6, 7].

The rapid proliferation of publications in this domain reflects both its technical promise and its broad applicability, spanning robotic manipulation, autonomous navigation, assistive technologies, and beyond [8, 9]. However, this explosive growth also introduces challenges: fragmented research directions, inconsistent evaluation methodologies, and an urgent need for systematic synthesis. While prior surveys have addressed subsets of this landscape—such as specific model architectures or application domains—a comprehensive bibliometric analysis integrating quantitative mapping with qualitative review remains absent from the literature.

This survey addresses this gap by conducting an extensive bibliometric study of 1,798 publications from Scopus and OpenAlex [10], employing computational techniques including keyword frequency analysis, co-occurrence network construction, temporal trend visualization, and K-means clustering [11] to reveal the field's structural and thematic evolution. We complement this quantitative foundation with a qualitative examination of seven influential papers, tracing the trajectory from foundational VLM architectures to state-of-the-art VLA systems. Our dual-method approach illuminates not only what has been achieved but also where critical gaps—particularly in safety, robustness, and evaluation standardization—demand future attention.

The contributions of this work are threefold:

1. We provide the first large-scale computational analysis of VLM and VLA literature, identifying dominant research clusters, geographic distributions, collaboration networks, and temporal patterns through rigorous quantitative methods [10, 11].

2. We systematically review key high-impact publications across development, assessment, and deployment dimensions, contextualizing technical advances within broader research trajectories and highlighting paradigm shifts from prompt-engineering tuning to models of multimodal capabilities based on foundational learning and transfer learning approaches [5, 6, 7].

3. We identify the persistent absence of dedicated safety and security research within VLM and VLA domains, underscoring the need for further developments, such as safety evaluation frameworks [12, 13].

The remainder of this paper is organized as follows: Section 2 details our bibliometric methodology, including data collection, preprocessing, and analytical techniques such as thematic time-series trend analysis, keyword clustering, and author collaboration mapping. Section 3 offers a qualitative review of seminal VLA papers, examining stages of development, assessment, and deployment of VLA and VLM systems. Section 4 provides detailed conslusions of the paper, while section 5 addresses limitations and future directions.

2 Computational Analysis of Literature: Bibliometrics

2.1 Methodology

2.1.1 Data Collection

To ensure comprehensive coverage of the VLM and VLA research landscape, we employed a dual-database approach, querying both Scopus and OpenAlex [10]. This strategy mitigates the limitations inherent to any single database, such as incomplete indexing or access restrictions, while maximizing the breadth and representativeness of our dataset. The query was designed to capture publications explicitly addressing vision-language models, vision-language-action systems, and related multimodal architectures.

The search query was structured as follows:

```
("vision-language" OR "vision-language-action" OR "multimodal vision-language")
AND ("model" OR "system" OR "architecture" OR "framework")
```

We applied no temporal restrictions to the initial search, allowing us to capture the full historical arc of research in this domain. However, subsequent temporal analysis revealed that publication activity accelerated dramatically beginning in 2020, consistent with breakthroughs in machine learning (ML) and deep learning (DL) with introduction of transformer-based deep neural network (DNN) architecture, which serves as a techological backbone for VLM and VLA systems [3].

2.1.2 Data Preprocessing and Cleaning

The raw dataset comprised 800 records from Scopus and 1,400 from OpenAlex, which was preprocessed to ensure data quality and relevance by identifying and removing duplicate entries based on Digital Object Identifiers (DOIs), titles, and author lists, a critical step given the partial overlap between the two databases [10]. Author names, institutional affiliations, and keywords were normalized to account for variations in spelling, formatting, and language—for instance, standardizing terms like Vision-Language Model," vision language model," and “VLM” to a single form—while records with missing values were excluded to maintain consistency and integrity [11]. After these steps, the final dataset consisted of 1,798 unique publications.

2.1.3 Analytical Techniques

Our bibliometric analysis employed multiple computational methods to uncover structural and thematic patterns within the literature [11].

Firsrtly, we extracted author-provided keywords and performed frequency counting to identify the most prevalent terms (Table 1). To capture semantic relationships, we constructed a co-occurrence Table 2 that represent keywords joint appearance within individual papers. This approach enabled us to identify tightly connected keyword instances, reflecting important research directions.

We aggregated publications by year and visualized in the 1 the growth trajectories to assess the field's expansion. Exponential growth fitting was applied to quantify the rate of increase, revealing a 50-fold rise in annual publications between 2020 and 2024. This temporal lens also allowed us to trace the emergence of specific themes—such as ”computer vision”, ”intelligent robotics”, and ”multimodal” approaches—across time.

To uncover latent thematic structures, we applied K-means with $k = 10$ for predefined number of clusters to titles and keywords [11]. Each cluster was visualized (Figure 2, Figure 3) with distinct color, while contextual proximity of individual nodes was expressed by the spacial layout of clusters. Clusters were interpreted by examining their most representative terms and publications, revealing ten distinct research themes.

We constructed a co-authorship map (Figure 4), where nodes represent authors and edges represent collaborative joint publications. This map allows to identify influential researchers and collaborative communities, which provides further geographical and institutional insights of the research field development. This analysis illuminated the field's social structure and highlighted key contributors driving innovation.

2.2 Results

2.2.1 Keyword Occurrence and Co-Occurrence

Table 1 presents the 15 most frequent keywords across the dataset. The top terms—”language models” (244 occurrences), ”visual languages” (192 occurrences), and ”action recognition” (97 occurrences)—underscore the convergence of natural language processing, computer vision, and action prediction as foundational pillars of VLA research [8]. Other prominent keywords include ”deep learning” (67 occurrences), reflecting the architectural backbone of most modern VLMs [3, 4], and ”multimodal” (53 occurrences), highlighting the emphasis on generalization capabilities across the modalities [14].

Table 1: Top Frequent Keywords

Keywords	Occurrences
language model	244
visual languages	192
large language model	106
action recognition	97
computer vision	94
vision-language model	90
deep learning	67
semantics	65
multimodal	53
robot learning	47
computational linguistics	47
video understanding	47
intelligent robots	46
contrastive learning	45
adversarial machine learning	42
Truncated Table. Total rows = 100.	

Presence of ”intelligent robots” and ”robot learning” terms suggests close relation of VLAs and VLMs to robotics research area [6]. High presence of keywords related to the natural language processing (NLP)—”language model”, ”large language model”, ”semantics”—confirms language as a foundational modality for the VLM and VLA systems.

Table 2 visualizes the keyword co-occurrence pairs, revealing that language model and visual languages are not isolated domains but mutually constitutive. Often interconnections between perception-oriented terms (action recognition, computer vision) and linguistic concepts (computational linguistics, semantics) demonstrate that research increasingly emphasizes integrated perception–language pipelines. The presence of multimodal pairs (”action recognition” and ”language model”, ”semantics” and ”visual languages”, ”intelligent robots” and ”language model”) supports transition from unimodal specialization toward architectures integrating understanding of various data types [1, 2]. This provides quantitative evidence that the community converges around common multimodal-learning foundations, positioning VLAs as the evolutionary step from earlier vision-language frameworks [5, 7].

Table 2: Top Co-Occurrences Keywords

Keyword i	Keyword j	Co-Occurrences
visual languages	language model	151
large language model	language model	70
vision-language model	language model	58
computer vision	language model	52
semantics	language model	50
multimodal	language model	48
action recognition	language model	46
computational linguistics	language model	46
vision-language model	visual languages	44
semantics	visual languages	43
multimodal	visual languages	37
action recognition	visual languages	35
intelligent robots	language model	33
computer vision	visual languages	32
adversarial machine learning	visual languages	31
Truncated Table. Total rows = 2422.		

2.2.2 Publication Trends and Growth Dynamics

Figure 1 illustrates the temporal evolution of VLM and VLA publications. The data reveal a modest but steady output prior to 2020, followed by exponential growth beginning in 2022. Specifically, annual publication counts increased from approximately 30 papers in 2020 to over 500 in 2024—a 50-fold increase within four years. This surge coincides with several key developments: the release of large-scale vision-language datasets (e.g., Open X-Embodiment) [15], breakthroughs in transformer-based architectures [3, 4], and the demonstration of VLA capabilities in robotic manipulation tasks [5, 6].

Temporal distribution of keywords from 2017–2026 appears in Figure 1. Total keyword frequency exhibits exponential growth, increasing from baseline levels in 2017–2021 to exceeding 10^3 occurrences by 2024–2025—representing over 50-fold increase within four years. The 2026 decline reflects incomplete indexing.

Language model demonstrates the most dramatic growth trajectory, remaining near-zero until 2021 before exploding from single-digit occurrences to over 100 publications in 2024–2025—a greater than 10× increase in just three years [3]. Visual languages mirrors this explosive pattern, with both curves rising in parallel from 2023 onward, demonstrating synchronized integration of visual and linguistic modalities across the VLA literature [5].

Computer vision and vision-language model exhibit similar mid-tier exponential trajectories, remaining minimal before 2022 before surging 5–7× during 2023–2024, achieving peak counts of approximately 40–70 occurrences during 2024–2025. Semantics and large language models display consistent acceleration starting in 2022, exhibiting 3–4× growth over two years and reaching 2024–2025 peaks near 60–70 occurrences [4]. Multimodal exhibits comparable intensity, inflecting around 2023 with 5× growth, peaking near 50 occurrences by 2025.

Intelligent robots emerges later but explosively, first registering measurable activity in 2024 and immediately reaching 46 occurrences—indicating rapid adoption once foundational multimodal capabilities matured [6] This coincides with action recognition's parallel 2024 surge to approximately 40 occurrences, representing 6× growth from 2023.

All major trajectories share three phases: flat pre-2021 baseline, sharp 2022–2023 inflection, and unified 2024–2025 plateau. Synchronous upward movement across language models, visual languages, multi-modality, computer vision, and action recognition indicates coordinated growth dynamics. The simultaneous rise of intelligent robots and vision-language model after 2024 correlates with transition from perceptual representation learning toward embodied implementations.

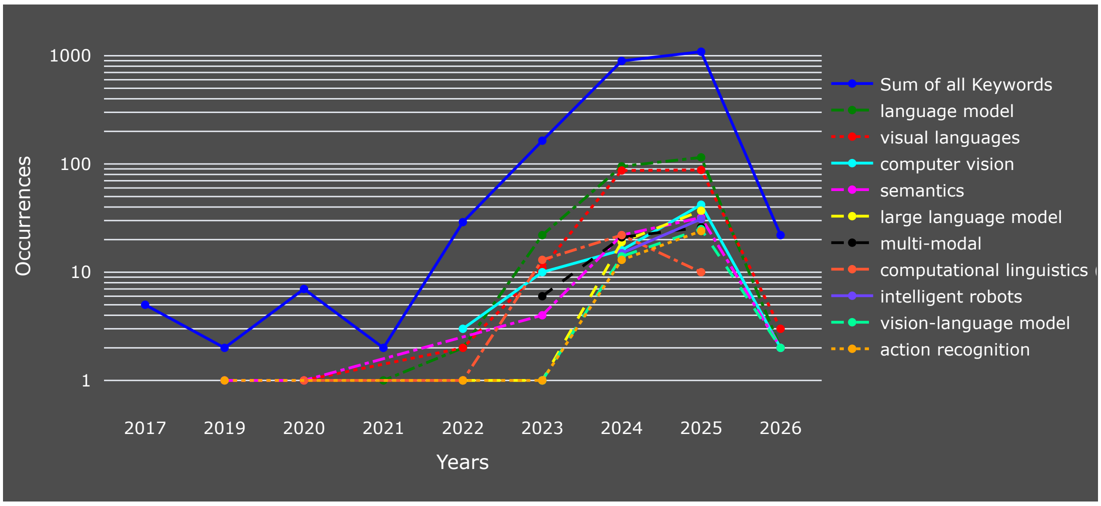

Figure 1: Keywords Timeseries

2.2.3 K-means Clustering

The ten clusters of keywords that are listed in Table 3 and visualized in Figure 2 demonstrate interconnected structure with minimal differentiation between linguistic, perceptual, and robotic domains. Central terms (language model, vision-language model, contrastive learning) occupy compact regions, while applied terms (robot learning, embodied intelligence) appear along periphery. Top keywords map consistently between adjacent clusters: language model to vision-language model (models architectures); deep learning to contrastive learning (learning mechanisms); robot learning to embodied intelligence (execution and control); foundation models to benchmarking (scaling and evaluation). Each cluster retains variants of the conceptual triad—multimodality, learning approaches, high-salience terms—indicating the VLA landscape comprises highly overlapping spheres unified by the goal of integrating perception, language, and action in generalizable systems [6, 8].

Table 3: Keyword Clusters Identified via K-means ($k = 10$)

Cluster	Representative Term	Count	Thematic Focus
1	Language model	244	Multimodal core: linguistic, visual, action integration
2	Vision-language model	90	Model architectures & multimodal representation
3	Deep learning	67	Machine learning foundations
4	Robot learning	47	Embodied robotics applications
5	Contrastive learning	45	Self-supervised representation learning
6	Zero-shot & Foundation models	42	Scaling and transfer learning
7	Benchmarking	35	Evaluation methodologies
8	NLP systems	30	Language-based multimodal integration
9	Modeling languages	24	Formal system design
10	Embodied intelligence	24	Planning and control for physical agents

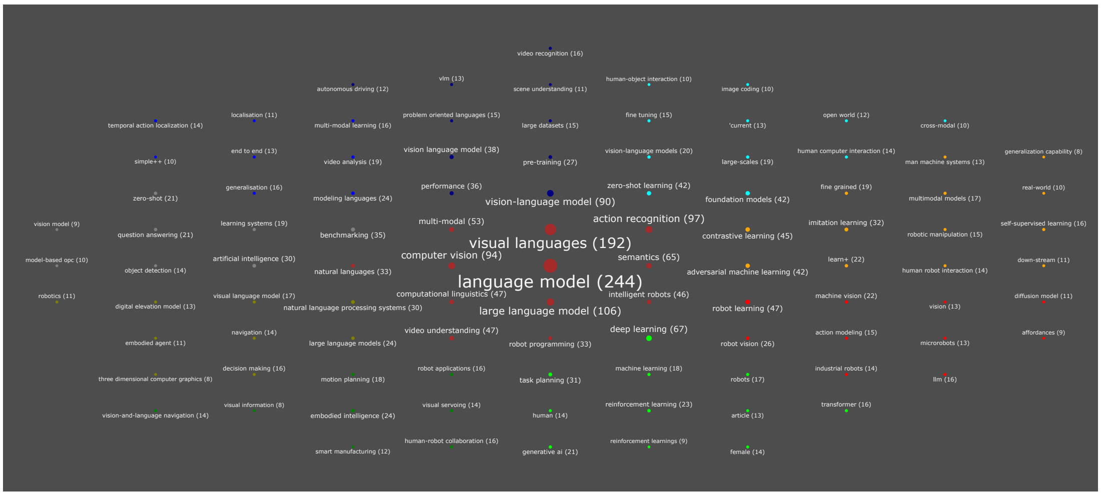

Figure 2: Keywords Clusters

K-means clustering of titles listed in Table 4 and visualized in Figure 3 identified similar ten distinct thematic clusters within the VLA literature, revealing field consolidation around foundational architectures and specialized applications [3]. Cluster 2 with a thematic category of (Video Understanding and Prompting), centered on (Prompting Visual-Language Models for Efficient Video), occupied the highest density region, indicating that prompting strategies the field's methodological core [16].

Strong spatial proximity between Cluster 3 (Pre-training and Foundation Models) and application-oriented clusters demonstrates active knowledge transfer from foundational research to practical implementations. Clusters 1 (Robotic Navigation and Embodiment) and 7 (Object Interaction and Manipulation) form the primary application domain, encompassing socially-aware navigation, quadruped control, and human-object interaction. Their spatial separation from video understanding clusters suggests divergent technical requirements between continuous navigation and video processing [5, 16].

Cluster 8 (Generalist VLA Models) occupies a bridging position between foundation models and specialized applications, reflecting architectural trends toward unified systems capable of zero-shot generalization across diverse embodiments [3, 6]. Peripheral positioning of Cluster 9 (Grounding and Spatial Understanding) indicates emerging interest in multimodal redundancy mechanisms, while Cluster 10 (Web Navigation and Knowledge Integration) remains spatially distant from physical robotics clusters, suggesting distinct architectural paradigms.

Table 4: Title Clusters Identified via Keyword Similarity ($k = 10$)

Cluster	Thematic Category	Description
1	Robotic Navigation & Embodiment	Socially aware robot navigation and embodied control
2	Video Understanding & Prompting	Prompting visual-language models for efficient video analysis
3	Pre-training & Foundation Models	Vision-language pre-training via embodied learning
4	Multimodal Agents & Intelligence	Next-generation intelligent assistants with multimodal capabilities
5	Temporal Action Recognition	Zero-shot temporal action localization and understanding
6	Language Planning & Reasoning	Unified representation for language-based planning
7	Object Interaction & Manipulation	Human-object interaction and manipulation tasks
8	Generalist VLA Models	Foundation models for generalist vision-language-action systems
9	Grounding & Spatial Understanding	Visually-grounded planning and spatial reasoning
10	Web Navigation & Knowledge	Generalist web agents and knowledge-based navigation

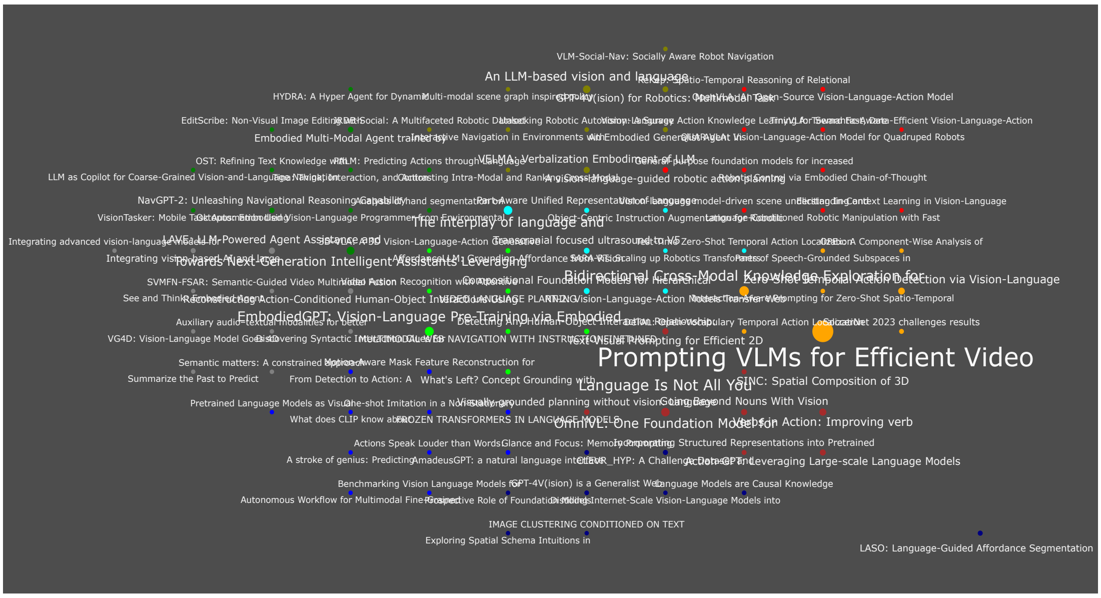

Figure 3: Titles Clusters

2.2.4 Author Collaboration Networks

Author collaboration network (Figure 4) revealed a highly concentrated research landscape dominated by tightly interconnected institutional networks. The most productive authors—such as Joshua B. Tenenbaum and Oier Mees—anchor clusters centered around major research institutions [17]. The central cluster, led by Sergey Levine, encompasses Berkeley-affiliated researchers including Mees, Pertsch, Finn, and Abbeel, exhibiting the highest collaboration density [6]. A parallel cluster, centered on Tenenbaum, unites MIT researchers such as Du, Kaelbling, and Xia, characterized by strong intra-institutional cohesion. These U.S.-based clusters demonstrate intensive internal collaboration but limited cross-institutional connectivity.

Chinese research groups form distinct and comparably dense sub-networks, notably around Wen Junjie, who maintains close ties with collaborators including Zhu Yichen, Zhu Minjie, Li Jinming, Xu Zhiyuan, and Shen Chaomin [7]. These networks show collaboration intensities on par with U.S. academic clusters, suggesting parallel trajectories in research development. European contributions remain more peripheral, while individual efforts indicate isolated research paths.

Overall, the network topology underscores a stark geographic concentration, with Visual-Language-Action (VLA) research primarily consolidated within U.S. and Chinese institutions. European and other international contributions remain fragmented and regionally isolated. This U.S.–China duopoly in foundational VLA research raises concerns about technological control, governance, and equitable access [12]. Given the potential applications of VLA systems in autonomous vehicles, robotics, and physical-world interaction, such concentration of expertise risks narrowing safety research perspectives, reinforcing proprietary dominance, and constraining the development of diverse ethical and regulatory frameworks.

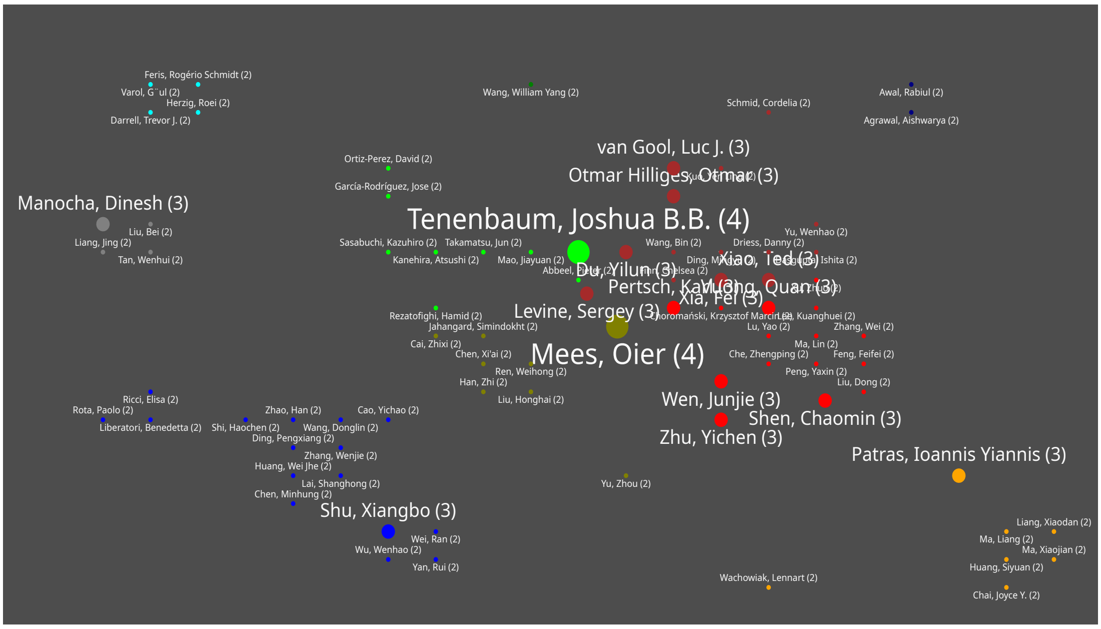

Figure 4: Authors Collaborative Map

3 Qualitative Review of Key Publications

3.1 Methodology

To complement our quantitative bibliometric analysis, we conducted a qualitative review of high-impact publications representing pivotal advances in VLA research. Selection criteria included citation count, publication venue prestige, methodological novelty, and thematic coverage across development, assessment, and deployment dimensions [18]. This approach ensured a clear representation of the field's trajectory, from model architectures to real-world applications.

3.2 Development: Main Approaches

3.2.1 GPT-4V for Robotics: Multimodal Task Planning

The integration of GPT-4V into robotic task planning represents a significant step toward leveraging large-scale vision-language models for action planning [16]. This work demonstrates how a pretrained VLM can be adapted—without fine-tuning—to generate high-level task plans from natural language instructions and visual observations. The system decomposes complex user commands (e.g., ”prepare a meal”) into sequences of executable sub-tasks (e.g., ”locate ingredients,” ”grasp utensil”), leveraging GPT-4V's world knowledge and reasoning capabilities.

Key contributions include a prompt-based architecture uniting perception, reasoning, and action execution. However, the closed-source nature of GPT-4V limits reproducibility and raises concerns about dependency on proprietary systems. Moreover, the system's reliance on high-level task planning leaves low-level motor control and error recovery largely unaddressed [16].

3.2.2 OpenVLA: Open-Source Vision-Language-Action Model

In contrast to closed-source approaches, OpenVLA pioneered open-source VLA development, training a 7-billion-parameter model on the Open X-Embodiment dataset [6, 15]. By leveraging diverse robotic demonstrations spanning 22 robot embodiments and over 1 million trajectories, OpenVLA achieves impressive zero-shot generalization to novel objects, environments, and tasks [14].

The model architecture combines a vision transformer for image encoding with a language model backbone (LLaMA-2) for joint vision-language reasoning, followed by an action decoder predicting low-level control commands [6, 19, 20]. Crucially, OpenVLA's open-source release—including model weights, training code, and evaluation scripts—has democratized access to state-of-the-art VLA capabilities, fostering community-driven innovation and enabling independent audits of model behavior.

Despite its strengths, OpenVLA inherits limitations from its training data, including potential biases toward specific robot morphologies and task distributions. Additionally, its computational requirements (7B parameters) pose barriers for deployment on resource-constrained platforms [6].

3.3 Assessment: Performance Evaluation Methodologies

3.3.1 LEFT Framework: Concept Grounding Evaluation

The LEFT (Logic-Enhanced Foundation Model Testing) framework addresses a critical gap in VLA evaluation: rigorous assessment of concept grounding—the ability to map linguistic concepts to visual referents [17]. Traditional benchmarks often conflate superficial pattern matching with genuine semantic understanding. LEFT mitigates this through logic-enhanced test generation, constructing compositional queries that probe fine-grained grounding capabilities (spatial relationships, attribute binding, numerical reasoning).

By systematically varying query complexity and introducing distractors, LEFT exposes failure modes in state-of-the-art VLMs, revealing brittleness in tasks requiring multi-step reasoning or disambiguation. The framework's emphasis on compositional generalization and adversarial robustness provides a blueprint for more rigorous VLA evaluation [17].

3.3.2 Cultural Bias Benchmarking

Recent work has highlighted cultural biases in vision-language models, demonstrating systematic performance disparities across geographic regions, languages, and cultural contexts [21]. Benchmarks such as CulturalVQA curate image-question pairs reflecting diverse cultural artifacts, traditions, and norms, then evaluate VLM accuracy and fairness metrics.

Results reveal pronounced biases favoring Western-centric content, with models achieving 15-25% lower accuracy on non-Western cultural references [21]. These findings underscore the urgency of diversifying training data and developing fairness-aware evaluation protocols. Without such efforts, VLA deployment risks perpetuating and amplifying existing inequities, particularly in global and multicultural contexts.

3.4 Deployment: Real-World Implementation

3.4.1 See and Think: Embodied Agents in Virtual Environments

The "See and Think" framework demonstrates VLA deployment in complex virtual environments, where agents must navigate, reason, and interact based on visual observations and natural language goals [22]. By coupling a VLM with a reinforcement learning policy, the system achieves human-level performance on tasks such as object retrieval, multi-step navigation, and interaction with dynamic environments.

Key innovations include a hierarchical policy architecture separating high-level planning (handled by the VLM) from low-level control (handled by a learned motor policy), as well as a memory module enabling long-horizon task execution [9, 22]. The virtual environment setting facilitates rapid iteration and scalable evaluation.

3.4.2 Autonomous Driving with VLMs

VLMs have also been explored for autonomous driving, where they provide natural language interfaces for human-vehicle interaction and support interpretable decision-making [23]. For example, systems integrate VLMs to generate textual explanations for driving maneuvers ("slowing down because pedestrian is crossing"), enhancing transparency and trust.

However, safety-critical deployment introduces critical requirements for robustness, real-time performance, and failure mitigation—areas where current VLMs exhibit significant shortcomings [24]. Adversarial inputs, such as occluded traffic signs or ambiguous visual scenes, can induce erroneous predictions with potentially catastrophic consequences. Addressing these vulnerabilities demands advances in adversarial training, uncertainty quantification, and fail-safe mechanisms.

3.4.3 Robotic Manipulation: From Lab to Real-World

Translating VLA capabilities from controlled laboratory settings to real-world robotic manipulation introduces challenges including sensor noise, object variability, and unstructured environments. Projects such as OpenVLA, RT-2, and RoboFlamingo and subsequent work have demonstrated successful deployment in domestic settings, where robots perform tasks like table setting, object sorting, and assistive manipulation [5, 6, 7].

Critical to these successes are domain adaptation techniques, closed-loop feedback mechanisms, and failure recovery strategies. For instance, RoboFlamingo employs visual servoing to refine grasp poses in real-time, compensating for perception errors [7]. Nonetheless, long-horizon tasks requiring multi-step reasoning and dynamic action planning remain as active research directions, pointing towards wide integration and adoption of VLA systems [5, 6].

4 Conclusion

This survey has provided a comprehensive bibliometric and qualitative analysis of VLM and VLA systems, synthesizing insights from 1,798 publications through quantitative mapping—including keyword frequency analysis, co-occurrence networks, temporal trends, and K-means clustering [11]—which illuminated the field's exponential growth, thematic structure, and geographic concentration, alongside a qualitative review tracing the trajectory from pioneering VLM architectures to state-of-the-art VLA systems, highlighting paradigm shifts toward open-source foundational models [6], rigorous evaluation frameworks [17], and real-world deployment [5, 7]. Key findings encompass a 50-fold increase in annual publications since 2020, driven by breakthroughs in transformer architectures [3, 4] and large-scale datasets [15]; convergence of vision, language, and action via multimodal transformer architectures [2, 25] with action tokenization [5]; pronounced geographic concentration in U.S. and Chinese institutions, raising concerns about technology governance and cultural bias [21]; the transition from closed-source models with prompt-based architectures (GPT-4V) to multimodal alternatives (OpenVLA, RT-2) reflecting growing emphasis on foundational learning and balanced unification of of diverse modalities [6]; frameworks like LEFT and cultural bias benchmarks signaling maturation of the field with increased attention to systematic evaluation, compositional generalization, and fairness [17, 21]; and persistent gaps in adversarial robustness [24, 26], safety mechanisms, hallucination mitigation [13], and standardized evaluation, which constrain real-world applicability, particularly in safety-critical domains, despite most successful VLA systems converging on transformer-based architectures.

As VLA systems transition from research prototypes to deployed technologies, addressing these gaps becomes imperative. The field stands at a critical juncture: continued technical maturation must be accompanied by equally rigorous attention to safety, fairness, responsible deployment, and international academic community involvement.

5 Limitations and Future Directions

5.1 Methodological Limitations

Our bibliometric analysis, while comprehensive, carries inherent limitations. First, reliance on Scopus and OpenAlex may exclude relevant work published in non-indexed venues or available only as preprints on platforms like arXiv [10]. Second, keyword-based clustering and co-occurrence analysis depend on author-provided keywords, which may not fully capture semantic nuances or interdisciplinary connections. Third, our qualitative review, though systematic, represents a selective snapshot; the rapid pace of publication means that emerging work may not yet be reflected.

Future bibliometric studies could incorporate citation network analysis, author trajectory tracking, and dynamic topic modeling to capture temporal evolution more granularly. Additionally, expanding coverage to include preprint literature and workshop papers would provide a more complete picture of the field's development.

5.2 Future Research Directions

As VLA systems evolve from research prototypes to deployed technologies, addressing key technical and ethical challenges is imperative, including enhancing adversarial robustness against visual and linguistic perturbations through robust training paradigms, uncertainty quantification, and fail-safe mechanisms [24, 26]; mitigating hallucinations via grounding mechanisms, external knowledge integration, and confidence calibration [13]; countering cultural biases by diversifying training data, incorporating fairness constraints, and conducting bias audits [21]; establishing standardized benchmarks and evaluation metrics for cross-study comparisons [17]; and reducing environmental and computational costs through model compression, efficient architectures, and green AI practices [6]. High-priority future directions encompass developing robust and safe VLAs with architectural innovations for safety guarantees [13, 24, 26]; creating culturally aware models with diverse data and fairness-focused evaluations [21]; advancing efficient and scalable architectures via knowledge distillation and hardware-aware designs; extending capabilities to long-horizon and hierarchical planning for multi-step tasks [5, 6]; standardizing benchmarks covering robustness, fairness, and interpretability [17, 21]; and fostering interdisciplinary collaborations among AI researchers, roboticists, ethicists, and domain experts to align VLA development with societal values [21]. By pursuing these directions, the research community can advance VLA systems to unlock their potentials in robotics, autonomous systems, and embodied intelligence further.

References

[1] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, "Contrastive learning of medical visual representations from paired images and text," in *Machine Learning for Healthcare Conference*, PMLR, 2022, pp. 2–25.

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, 2021, pp. 8748–8763.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 5998–6008.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[5] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, P. Florence, C. Fu, M. Arenas, K. Gopalakrishnan, K. Han, K. Hausman, A. Herzog, J. Hsu, B. Ichter, A. Irpan, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, L. Lee, T.-W.E. Lee, S. Levine, Y. Lu, H. Michalewski, I. Mordatch, K. Pertsch, K. Rao, K. Reymann, M. Ryoo, G. Salazar, P. Sanketi, P. Sermanet, J. Singh, A. Singh, R. Soricut, H. Tran, V. Vanhoucke, Q. Vuong, A. Wahid, S. Welker, P. Wohlhart, J. Wu, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich, "RT-2: Vision-language-action models transfer web knowledge to robotic control," in *Proceedings of the 6th Conference on Robot Learning*, 2023, pp. 2165–2183.

[6] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, Q. Vuong, T. Yu, T. Kollar, A. Finn, S. Levine, and C. Finn, "OpenVLA: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[7] X. Li, M. Liu, H. Zhang, C. Yu, X. Xu, H. Wu, C. Cheang, Y. Jing, W. Zhang, H. Song, W. Zhang, H. Xu, and L. Kong, "Vision-language foundation models as effective robot imitators," in *International Conference on Learning Representations*, 2024.

[8] Y. Liu, W. Zhang, and H. Chen, "Vision-language-action models: Concepts, progress, and future directions," *arXiv preprint arXiv:2505.04769*, 2025.

[9] L. Wang and J. Chen, "A survey: Learning embodied intelligence from physical interaction," *arXiv preprint arXiv:2507.00917*, 2025.

[10] R. Pranckutė, "Web of Science (WoS) and Scopus: The titans of bibliographic information in today's academic world," *Publications*, vol. 9, no. 1, p. 12, 2021.

[11] N. Bakas, S. Lavdas, K. Vavousis, C. Christodoulou, and A. Langousis, "Automated Machine Learning in a Multi-agent Environment," in *Information Systems: 21st European, Mediterranean, and Middle Eastern Conference, EMCIS 2024, Proceedings*, M. Themistocleous, N. Bakas, G. Kokosalakis, and M. Papadaki, Eds. Cham, Switzerland: Springer, pp. 47–57, 2025.

[12] IBM Research, "AI governance: Principles, challenges, and best practices," IBM Think, 2024. [Online]. Available: https://www.ibm.com/think/topics/ai-governance

[13] Y. Zhou, S. Li, R. Zhang, J. Wang, and M. Chen, "Investigating VLM hallucination from a cognitive psychology perspective," *arXiv preprint arXiv:2507.03123*, 2024.

[14] J. Rocamonde, V. Montesinos, E. Nino, E. Cercós, and N. R. Ke, "Vision-language models are zero-shot reward models for reinforcement learning," *arXiv preprint arXiv:2310.12921*, 2023.

[15] *Open X-Embodiment Collaboration*, "Open X-embodiment: Robotic learning datasets and RT-X models," *arXiv preprint arXiv:2310.08864*, 2023.

[16] N. Wake, A. Kanehira, K. Sasabuchi, J. Takamatsu, and K. Ikeuchi, "GPT-4V(ision) for robotics: Multimodal task planning from human demonstration," *IEEE Robotics and Automation Letters*, vol. 9, no. 4, pp. 3772–3779, 2024.

[17] J. Hsu, J. Mao, J. B. Tenenbaum, and J. Gao, "What's left? concept grounding with logic-enhanced foundation models," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 70248–70267.

[18] N. Donthu, S. Kumar, D. Mukherjee, N. Pandey, and W. M. Lim, "How to conduct a bibliometric analysis: An overview and guidelines," *Journal of Business Research*, vol. 133, pp. 285–296, 2021.

[19] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "DINOv2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[20] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11816–11826.

[21] H. Liu, Y. Li, Z. Chen, Y. Zhao, J. Zhang, and M. Wang, "Benchmarking vision language models for cultural understanding," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 5421–5438.

[22] W. Chai, C. Zheng, H. Wu, J. Zhang, L. Guo, and B. An, "See and think: Embodied agent in virtual environment," in *European Conference on Computer Vision*, 2024, pp. 154–171.

[23] J. Wu, B. Gao, J. Gao, J. Yu, H. Chu, Q. Yu, X. Gong, Y. Chang, H. E. Tseng, H. Chen, and J. Chen, "Prospective Role of Foundation Models in Advancing Autonomous Vehicles," *Res*, vol. 7, Art. no. 0399, 2024.

[24] E. Mozhegova, M. J. Khan, S. Patel, A. Kumar, and R. Singh, "Assessing the adversarial robustness of multimodal foundation models across modalities," *Frontiers in Medicine*, vol. 12, p. 1606238, 2025.

[25] B. Mustafa, C. Riquelme, J. Puigcerver, L. Beyer, S. Houlsby, N. Houlsby, S. Gelly, J. Keysers, P. Lucic, and X. Zhai, "Multimodal contrastive learning with LIMoE: The language-image mixture of experts," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 9564–9576.

[26] C. Jiang, Z. Wang, M. Dong, and J. Gui, "Survey of adversarial robustness in multimodal large language models," *arXiv preprint arXiv:2503.13962*, 2025.