Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD29606

Multimodal Learning Paradigms for Open-World Visual Understanding: Advances, Challenges, and Future Directions

Guorun Wang

School of Cyberspace Science and Technology, Beijing Institute of Technology, Beijing, China
1120221727@bit.edu.cn

Abstract. In open-world visual scenes, since insufficient semantic openness, modality isolation and cognitive limitations, traditional unimodal vision models face significant challenges. For example, these models can not recognize unknown categories beyond the training set. They suffer from generalization collapse under variations such as illumination and viewpoint changes, and struggle to understand the deep implicit semantics of scenes. Multimodal learning addresses these limitations by integrating visual, linguistic, and 3D spatio-temporal information to construct joint embedding spaces. This approach significantly enhances the understanding capabilities for openScene, promotes a paradigm shift in research from closed-set recognition towards open-world cognition. This paper systematically reviewed research progress about multimodal learning for open-world visual understanding and provided researchers with a comprehensive research route in the field. Through critical analysis revealing limitations and core challenges of existing research, this paper proposed future directions to advance the development of more efficient and stable models for open-world visual understanding. These findings established a stronger perceptual and cognitive foundation for applications such as autonomous driving, embodied AI, and content moderation.

Keywords: Open-World Visual Scene Understanding, Multimodal Learning, Vision-Language Models, Open-Vocabulary Object Detection

1. Introduction

Traditional computer vision models face fundamental bottlenecks in dynamic and open real-world environments. These models mainly rely on unimodal perception, operating under the core assumption that the training and test sets share same categories and visual domains. However, the closed world assumption (CWA) is difficult to follow in practical applications, leading to performance degradation when encountering unknown categories, domain shifts, or complex semantic understanding. Traditional models exhibit three core issues: semantic openness deficiency, modality isolation and cognitive limitations. First, semantic openness deficiency, i.e., the inability to recognize object categories that are not present in the training set. Second, modality isolation means a lack of cross-modal alignment capability, leading to generalization collapse under variations such as illumination changes and viewpoint shifts. Third, because of cognitive limitations,

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.LD29606

it is difficult to understand the potential implicit semantics of scenes, such as a phone that is charging [1].

In recent years, multimodal learning, particularly the rise of Vision-Language Models(VLMs), has provided a revolutionary solution to overcome these bottlenecks. VLMs are trained on a large number of image-text pairs, learning powerful cross-modal alignment capabilities. This enables them to leverage the rich language semantics as prior knowledge, guiding the generalization of visual concepts. This inflection point of technology has driven a paradigm shift in research from closed-set recognition to open-world cognition. However, most of existing studies treat category generalization and domain generalization as independent problems. Real-world challenges are often the result of complex interplay between both, resulting in a new problem of Open-Domain Open-Vocabulary(ODOV) object detection. ODOV aims to address the compound generalization challenges across both categories and domains simultaneously [1]. Meanwhile, recent systematic surveys on Open-World Object Detection(OWOD) have emphasized the need for models to get dynamic learning capabilities, such as detecting unknown categories and incorporating them into the knowledge base(KB) through incremental learning [2]. This evolution marks a key transformation from static recognition to dynamic cognition.

VLMs have made a significant progress in open-world perception, but their application of complex tasks still faces kinds of challenges. Current methods usually utilize large language models to generate richer text prompts, but ignore their original reasoning capabilities. This leads to hallucinations or reasoning errors when faced with ambiguous or complex natural language queries [3,4].

Furthermore, contemporary multimodal models will exhibit vulnerability to logical composition and semantic perturbations with inadequate evaluation frameworks [5]. Current models also tend to learn spurious correlations of the data rather than the underlying causal mechanisms [6]. For instance, if sky and bird frequently co-occurrence in the training data, a model might incorrectly infer that sky is a sufficient condition for bird, resulting in unreliable predictions in the open world.

In the field of three-dimensional(3D) open-scene understanding, existing methods are often constrained by modality isolation. Most approaches rely on two-dimensional(2D) depth maps as intermediate representations, ignoring the intrinsic spatial properties of 3D point clouds [7]. At the level of fine-grained perception, the MultiModal Large Language Models((MM-LLMs) lose key details due to visual token compression, weakening their ability to detect subtle industrial defects [8].

This study systematically reviews the major technological developments of multimodal learning, the recent advances for open-world visual understanding, and analyzes in-depth the core challenges faced by these approaches. The essay hope to provide a clear and comprehensive research development guide for researchers. Moreover, the paper promotes the development of more efficient and generalizable open-world visual understanding models, enabling the system to be updated more intelligently and reliably.

2. Theoretical foundations of open-world multimodal understanding

2.1. Definition and challenges

The core challenge of open-world visual understanding lies in models' capacity to process unseen categories and domains during deployment. Zhang et al. [1] introduced the novel problem of ODOV object detection, which addresses the composite generalization challenge arising from concurrent category shift and domain shift. They contend that existing studies often treat these issues in

DOI: 10.54254/2755-2721/2025.LD29606

isolation, whereas real-world scenarios typically manifest synergistic variations of both—constituting the fundamental impediment to open-world robustness.

2.1.1. Causal perspective on open-world challenges

Through the lens of causal inference, open-world challenges stem from models' overreliance on spurious correlations. Liu et al. [6] demonstrate that conventional models tend to fit the joint data distribution, thereby capturing inter-modal spurious associations rather than underlying causal mechanisms. For instance, given frequent co-occurrence of birds and sky in training data, a model may erroneously establish sky as a strong predictor of birds. When confronted with an image containing sky without birds (e.g., balloons), it may still produce false positives—a critical vulnerability in open-world settings where unseen domains/categories readily disrupt training-phase spurious dependencies.

Gokhale [5] points out this argument with the perspective of system reliability. He asserts that the vulnerability of current multimodal models when handling logical compositions, semantic perturbations, and complex instructions stems from their insufficient reasoning capabilities. This is limited to processing at the recognition level. This recognition-centered paradigm makes the model susceptible to adversarial inputs and out-of-distribution(OOD) data. Therefore, achieving truly robust open generalization requires not only disentangling spurious correlations but also constructing systems capable of knowledge-guided reasoning, adversarial discovery, and reliable assessment capabilities. It drives a paradigm shift from recognition to reasoning.

2.2. Multimodal-driven open generalization mechanisms

Multimodal learning provides critical mechanisms for addressing these challenges, especially large-scale vision-language pretrained models like CLIP. Zhang et al. [1] illuminate this potential through their Domain-Agnostic Category Prompts framework. They utilized large language models to generate domain-invariant category descriptions as textual prompts of category embeddings. These prompts intentionally emphasize intrinsic visual attributes while avoiding domain-sensitive properties. By inputting the cleaned textual representations into a frozen CLIP text encoder, they obtained domain-independent, generalizable category embeddings.

The theoretical significance of this approach lies in its decoupling of category generalization and domain adaptation mechanisms. The linguistic modality provides powerful cross-domain semantic representations to resolve category generalization, while the visual modality captures input-specific domain characteristics.

However, this approach is still limited to the static "frozen VLM + prompt engineering" paradigm. Huang et al. [4] advanced this mechanism with their LLM2CLIP framework, marking a theoretical evolution from prompt generation to representation enhancement. They hypothesize that LLMs should not merely serve as external prompt generators, but actively optimize representational capacity of VLM. Their caption-to-caption contrastive fine-tuning strategy significantly enhanced discriminability of textual features generated by LLM, and then used these refined outputs to post-train the text encoder of CLIP. This approach confirms that LLMs are private tutors for visual encoders, achieving reverse empowerment where the linguistic modality actively enhances the visual modality, thus establishing a deeper theoretical pathway for open generalization.

DOI: 10.54254/2755-2721/2025.LD29606

2.3. Theoretical frameworks

Li et al. [2] systematically formalized the evaluation paradigm for Open-World Object Detection(OWOD). They pointed out that OWOD evaluation is a dynamic, multi-stage process which includes class-agnostic region proposals, unknown-aware classification, and incremental learning of unknown class. This unified evaluation framework, which integrates unknown detection and incremental learning into a continuous decision-making process, distinguishes OWOD from conventional object detection.

Zhang et al. [1] further contributed a novel benchmarking perspective through their OD-LVIS benchmark The benchmark contains 46,949 images covering18 complex real-world domains and 1,203 categories. The theoretical contribution of OD-LVIS lies in its pioneering quantification of both category diversity and domain diversity within a unified benchmark, providing a standardized platform for evaluating models under complex distribution shifts.

Wang et al. [9] proposed the UniPLV framework, a unified theoretical architecture, for open-world 3D understanding based on this. They observed that although point cloud-text pairs are scarce, point clouds and images often co-occur with spatial alignment. Thus, images can serve as a bridging modality to jointly embed pre-aligned image-text features and point cloud features in a shared latent space. Through logit distillation and feature distillation, this framework transfers region-level semantic knowledge from 2D foundation models to the 3D point cloud branch, enabling unified learning without point cloud-text pairs. This image-text-point cloud tripartite joint embedding paradigm, not only mitigates high annotation costs for 3D data, but also establishes a label-efficient framework for open-world understanding, expanding the frontiers of multimodal theory.

3. Key technological route

3.1. Evolution of open-vocabulary perception techniques

The core of open-vocabulary perception lies in leveraging zero-shot capabilities of pretrained large models. Wu et al. [10] explained this approach in their Open-Vocabulary Video Anomaly Detection framework, which decouples the task into class-agnostic detection and class-specific classification. Regarding classification, they adopt CLIP's cross-modal alignment mechanism, while acknowledging the limited generalization ability of zero-shot methods. To address this issue, they introduced a Novel Anomaly Synthesis module that harnesses LLMs and AIGC models to generate pseudo-anomalous videos for fine-tuning, substantially enhancing generalization for unseen anomalies.

In 2D image domains, Ma et al. [11] propose the APOVIS framework for efficient Open-Vocabulary Instance Segmentation. Its cornerstone is automated pseudo-label generation: first employing pretrained VLMs (e.g., ALBEF) for image-text matching, then extracting instance activation maps via Grad-CAM to derive coordinate-based pseudo-labels, and finally guiding foundation segmentation models (e.g., SAM) for pixel-level segmentation. This "VLM localization + foundation model segmentation" paradigm achieves high-efficiency, high-accuracy OVIS.

Nevertheless, existing methods struggle with industrial-grade fine-grained anomaly detection, often producing binary classifications or coarse masks lacking granular anomaly interpretation. Deng et al. [8] elevate open-vocabulary perception through their VMAD framework, which not only achieves pixel-level anomaly localization but also generates natural language explanations via Multimodal Large Language Models addressing "what anomaly", "impact

DOI: 10.54254/2755-2721/2025.LD29606

severity", and "remediation guidance". Core innovations include Defect-Sensitive Structure Learnin, which is to inject visual discriminative knowledge into LLM space by contrasting patch-similarity distributions between normal/anomalous samples, and Locality-Enhanced Token Compression, which mining multi-level local-context features in visual projectors to significantly enhance micro-defect sensitivity.

VMAD's state-of-the-art performance on MVTec-AD validates that visually augmented MLLMs constitute an effective pathway to fine-grained, interpretable open-vocabulary perception.

3.2. Embodied intelligence and 3D open-scene understanding

3D open-scene understanding is fundamental to embodied intelligence. Jiang et al. [12] tackle open-vocabulary 3D semantic segmentation via their OV3D framework, addressing the critical challenge of scarce point cloud-text pairs. Their solution utilizes multi-view 2D images as a bridging modality: first employing Large Vision-Language Models to generate rich textual descriptions (EntityText) for image entities, then establishing pixel-to-EntityText correspondences via vision foundation models, and finally transferring these relationships to 3D points for fine-grained point-entity text alignment.

Diverging from pure perception, Liu et al. [13] propose Uni3D-LLM to unify 3D perception, generation, and editing. Its core innovation is an LLM-based control center featuring an LLM-to-Generator mapping block that directly channels LLMs' semantic comprehension into controllable 3D content generation or editing, which signifying a paradigm shift from isolated perception to integrated perception-generation-editing.

Advancing further, Chen et al. [14] introduce Reasoning3D, establishing a "cognition-interaction" paradigm for 3D understanding. This framework pioneers reasoning-driven 3D segmentation, requiring fine-grained localization from complex instructions (e.g., "segment the part essential for detonation") with natural language explanations. Methodologically, it executes instructions on multi-view images via pretrained 2D reasoning segmentation models (e.g., LISA), then generates 3D segmentation through face-ID mapping and multi-view fusion. This "instruction-segmentation-explanation" closed loop enables models not merely to "see" but to "comprehend" and "explain", creating novel possibilities for human-robot collaboration in AR/VR and robotic operations.

Concurrently, Jiang et al. [7] contribute the MCL-3D-VCNN framework as a critical technical complement. It constructs a pluggable multimodal contrastive learning module that dynamically co-trains V-LLM (CLIP) and 3D CNN to achieve real-time alignment of 2D depth maps, 3D point clouds, and textual descriptions. Its multi-scale contrastive learning mechanism further enhances geometric detail perception, delivering significant improvements on ModelNet40 and ScanObjectNN – demonstrating that dynamic, scalable joint learning is a key technical pathway for robust 3D open understanding.

4. Typical failure scenario problem solving direction

4.1. Text-rich image comprehension scenarios

Contemporary open-world understanding models exhibit significant deficiencies when processing specialized scenarios, with text-rich image comprehension representing a notable failure mode. Hu et al. [15] identify that standard Vision-Language Models (VLMs) encapsulate image context through a fixed set of query embeddings—an approach that invariably discards critical textual details in densely textual images due to inherent token count constraints. To address this, they propose the BLIVA model, whose core innovation is a 'visual assistant' module that directly projects

DOI: 10.54254/2755-2721/2025.LD29606

encoded image patch embeddings into the Large Language Model (LLM). This architecture significantly enhances performance on OCR-VQA benchmarks, demonstrating unequivocally that preserving raw visual granularity is paramount for resolving such fine-grained alignment challenges.

4.2. Fine-grained alignment and long-tail distribution challenges

Current methodologies remain inadequate for addressing fine-grained alignment and long-tail distribution problems. Ma et al. [11] substantiate that entity-level textual descriptions—as opposed to image-level annotations—enable superior alignment precision. Their experiments achieve 69.7% mAP on COCO, markedly outperforming coarse-grained annotation approaches.

Jiang et al. [12] further corroborate the criticality of fine-grained alignment. By establishing point-entity-level correspondences, they resolve inherent coarse alignment limitations in conventional methods, achieving granular point cloud-text alignment. Their approach attains 57.3% zero-shot mIoU on ScanNet and delivers a +2.5% performance gain in long-tail scenarios (ScanNet200), directly evidencing that fine-grained alignment is indispensable for advancing open-world understanding, particularly under long-tail distributions.

Deng et al. [8] uncover an additional dimension of fine-grained alignment: joint semantic-structural alignment. The Defect-Sensitive Structure Learning (DSSL) module in VMAD reveals that neither visual nor linguistic modalities alone can reliably distinguish subtle anomalies. However, by aligning patch-similarity distributions of normal versus anomalous samples in cross-modal space, discriminative capability is substantially augmented. This insight pioneers a new direction for generalizing to long-tail categories and unseen compositions: future alignment mechanisms must transcend pixel-text matching and evolve toward unified structure-semantic modeling.

5. Core challenges

5.1. Challenges in dynamic open-world modeling

5.1.1. Catastrophic forgetting

Models exhibit significant degradation in prior task performance during incremental learning of new tasks. Lin et al. [16] systematically investigate this challenge in their MoE-LLaVA framework, revealing that naïvely integrating sparse Mixture-of-Experts (MoE) architectures into Large Vision-Language Models (LVLMs) compromises multimodal alignment capabilities. To mitigate this, they design a three-stage MoE-Tuning strategy employing phase-wise training anchored by prior knowledge—demonstrating that dynamic learning necessitates meticulously designed progressive training protocols.

5.1.2. Absence of open generalization theory

Current models lack theoretically grounded generalization guarantees. Dong et al. [17] observe in their InternLM-XComposer2 study that despite competitive performance, generalization remains limited when processing "complex compositional instructions beyond training coverage". This indicates models primarily generalize through heuristic pattern matching rather than genuine compositional reasoning, rendering their generalization behavior neither interpretable nor predictable.

DOI: 10.54254/2755-2721/2025.LD29606

This limitation is empirically substantiated by Gokhale [18], who asserts that model failures frequently originate from deficient reasoning beyond pattern recognition. His VQA-LOL benchmark reveals catastrophic failure modes when queries involve logical transformations (e.g., negation, conjunction), indicating that existing models' generalization resembles black-box pattern matching rather than knowledge- and rule-based compositional reasoning—a direct manifestation of theoretical deficits in open generalization.

Gokhale [5] further attributes this theoretical gap to systemic architectural limitations: current multimodal models' fragility under logical compositions and semantic/syntactic perturbations stems from inadequate knowledge-guided regularization and adversarial discovery mechanisms. Prevailing data-driven pattern matching approaches fail to incorporate external knowledge (e.g., commonsense, physical laws) into the generalization process. Thus, constructing theoretically sound frameworks with explainable generalization is paramount for resolving this challenge.

5.2. Multimodal adversarial vulnerabilities

Multimodal systems exhibit expanded attack surfaces across modalities. Dong et al. [17] note that InternLM-XComposer2's CLIP-ViT encoder may overlook fine details at fixed resolutions, creating vulnerabilities to adversarial samples. Lin et al. [16] concurrently identify that MoE-LLaVA's expert routers can be maliciously manipulated to activate suboptimal expert pathways, degrading output integrity.

5.3. Deficiencies in model robustness evaluation

5.3.1. Quantification gap

A critical shortcoming in current research is the pronounced quantification gap in robustness assessment. Neither Dong et al. [17] nor Lin et al. [16] evaluate their models under structured adversarial scenarios (e.g., texture attacks), undermining confidence in real-world safety.

Gokhale [5] elevates this issue to an evaluation paradigm crisis, contending that prevailing metrics over-rely on accuracy while neglecting performance under complex reasoning, adversarial perturbations, and open instructions. This prescribes a clear remediation path: evaluations must evolve from assessing "whether predictions are correct" to explaining "why they are correct" and determining "under what conditions they remain correct".

5.3.2. Robustness-complexity trade-offs

Current robustness assessments are reductively simplistic, whereas real-world challenges are inherently multidimensional. Gokhale [18] exposes intrinsic robustness-complexity trade-offs: his empirical analysis shows that data-filtering methods for eliminating spurious correlations may enhance in-distribution performance but impair adversarial robustness. This fundamental tension warns that enhancing one robustness dimension may compromise others.

6. Future directions

6.1. Cognitive paradigm shift: from recognition to reasoning

Contemporary multimodal systems predominantly operate at rudimentary "perception-classification" levels, lacking profound semantic comprehension and explanatory capabilities. The Reasoning3D

DOI: 10.54254/2755-2721/2025.LD29606

framework by Chen et al. [14] and VMAD framework by Deng et al. [8] collectively herald a transformative direction: deep integration of LLM reasoning capacities into perception systems to achieve a cognitive leap from recognition to reasoning. Future models should not merely identify "what" but elucidate "why" and "how to utilize". Research should systematize this "instruction-segmentation-explanation" closed-loop mechanism, extending it to human-AI collaborative decision-making scenarios like embodied intelligence and autonomous driving, thereby constructing cognitively capable agents.

6.2. Knowledge-guided reliable evaluation paradigm

Current multimodal evaluation over-relies on singular metrics (e.g., accuracy, mIoU), failing to holistically assess robustness and interpretability in open-world settings. This reductive paradigm neglects models' "cognitive abilities", resulting in fragility against logical compositions, semantic perturbations, and complex instructions [5]. Therefore, future research should pioneer knowledge-guided evaluation frameworks, which incorporate external knowledge into assessment protocols. This paradigm will employ adversarial discovery mechanisms to generate challenging test samples, as well as construct novel benchmarks evaluating spatial reasoning, conceptual compositionality, and automated critique generation.

This approach will expose model vulnerabilities while catalyzing the evolution from "black-box recognition" to "white-box reasoning", ultimately achieving trustworthy open-world understanding.

6.3. Unified multimodal joint learning paradigm

Prevailing "frozen VLM + fine-tuned vision model" decoupled learning constrains dynamic cross-modal interaction. Jiang et al.'s MCL-3D-VCNN [7] and Huang et al.'s LLM2CLIP [4] frameworks collectively illuminate a new pathway: unified multimodal joint learning. This paradigm emphasizes real-time alignment of linguistic, visual, and 3D modalities in shared representational spaces through dynamic joint training. MCL-3D-VCNN achieves end-to-end optimization of V-LLM and 3D CNN via pluggable contrastive modules, while LLM2CLIP employs caption-to-caption contrastive fine-tuning to transform LLMs into "private tutors" for CLIP visual encoders. This dual evolution—from "prompt generation" to "representation enhancement", and from "static fine-tuning" to "dynamic joint learning" – provides foundational insights for overcoming modal isolation.

6.3.1. Integrating structural priors: multimodal graph learning

Current joint learning overemphasizes feature alignment while ignoring inherent structural relationships. Liu et al.'s Graph-MLLM [19] demonstrates that fine-tuning MLLMs as predictors yields significant performance gains in node classification without explicit graph information, suggesting MLLMs internalize structural knowledge.

Future work should inject explicit graph structures (e.g., adjacency relations) into unified frameworks. For example, incorporating graph attention mechanisms into MCL-3D-VCNN's contrastive modules or fusing knowledge graph information during LLM2CLIP's text encoding. This will advance joint learning from "feature alignment" to "joint structure-semantic modeling", unlocking potential in social network analysis and knowledge graph reasoning.

DOI: 10.54254/2755-2721/2025.LD29606

6.3.2. Intelligent optimization: reinforcement learning with chain-of-thought

Current joint learning relies excessively on supervised/instruction fine-tuning, lacking explicit guidance for decision logic. Zhang et al.'s HOID-R1 [3] introduces a novel optimization paradigm: integrating Chain-of-Thought (CoT) with Reinforcement Learning (RL). The core methodology employs a dual-phase optimization strategy: first, supervised fine-tuning enables the model to generate structured reasoning processes; subsequently, the GRPO algorithm optimizes policy parameters by leveraging multiple reward signals – including format compliance, detection precision, interaction classification accuracy, and CoT coherence.

Future research should embed this RL+CoT paradigm into joint learning frameworks. For example, formulating MCL-3D-VCNN training as policy optimization with RL-adjusted contrastive loss weights or guiding LLM2CLIP to generate more discriminative prompts via RL. This will enhance both performance and decision interpretability.

6.3.3. Sustainable evolution: efficient continual learning

To adapt to open worlds' dynamic evolution, unified learning requires sustainable continual learning mechanisms. Zhao et al. [20] introduce the LLaVA-CMoE framework, delivering an efficient continual learning solution: its Probe-Guided Knowledge Expansion (PGKE) mechanism dynamically allocates task-specific experts for novel scenarios while preventing parametric bloat; concurrently, the Probabilistic Task Locator (PTL) coupled with VAE-based feature reconstruction effectively addresses routing conflicts and catastrophic forgetting.

Future work should integrate these mechanisms. For example, designing MCL-3D-VCNN's pluggable modules as MoE architectures with PGKE or incorporating PTL into LLM2CLIP post-training for conflict-free task routing.This establishes technical foundations for lifelong learning embodied agents.

In summary, the future unified multimodal joint learning paradigm must be a comprehensive framework integrating structural priors, intelligent optimization, and continual evolution. By synthesizing multimodal graph learning, RL-enhanced CoT, and efficient continual learning, this paradigm will achieve a fundamental transition from "static feature fusion" to "dynamic cognitive synergy", ultimately realizing genuine multimodal unification.

7. Conclusion

Open-world visual comprehension is progressing toward an era of enhanced intelligence and heightened reliability. Nevertheless, formidable challenges persist: catastrophic forgetting in dynamic environments, the absence of open generalization theory, multimodal adversarial vulnerabilities, and quantification gaps in evaluation frameworks constitute critical unresolved issues. This work demonstrates how multimodal learning, particularly through Vision-Language Models (VLMs), constructs joint embedding spaces that surmount three fundamental limitations of conventional unimodal systems: semantic confinement, modal isolation, and cognitive constraints. Critical technical pathways, including LLM2CLIP's representation enhancement, VMAD's fine-grained perception, and Reasoning3D's cognitive reasoning—collectively epitomize the evolutionary trajectory from simple recognition to complex comprehension. Despite substantial progress, the field confronts persistent challenges: catastrophic forgetting in dynamic environments, the absence of open generalization theory, multimodal adversarial vulnerabilities, and significant quantification gaps in current evaluation frameworks.

DOI: 10.54254/2755-2721/2025.LD29606

Future research should prioritize three foundational paradigms: A recognition-to-reasoning cognitive transition to achieve interpretable interaction; Unified multimodal joint learning integrating structural priors, intelligent optimization, and sustainable evolution; Knowledge-guided reliable evaluation to advance models from black-box pattern matching to white-box reasoning. Collectively, these directions will propel the field toward more robust, efficient, and generalizable futures.

References

[1] Y. Zhang, R. Han, F. Zhou, S. Wang, W. Feng, and L. Wan, “ODOV: Towards Open-Domain Open-Vocabulary Object Detection, ” arXiv.org, 2025. https: //www.arxiv.org/abs/2508.01253

[2] Y. Li, Y. Wang, W. Wang, D. Lin, B. Li and K. -H. Yap, "Open World Object Detection: A Survey, " in IEEE Transactions on Circuits and Systems for Video Technology, vol. 35, no. 2, pp. 988-1008, Feb. 2025, doi: 10.1109/TCSVT.2024.3480691.

[3] Z. Zhang et al., “HOID-R1: Reinforcement Learning for Open-World Human-Object Interaction Detection Reasoning with Multimodal Large Language Model, ” arXiv.org, 2025. https: //arxiv.org/abs/2508.11350 .

[4] W. Huang et al., “LLM2CLIP: Powerful Language Model Unlock Richer Visual Representation, ” arXiv.org, 2024. https: //arxiv.org/abs/2411.04997

[5] T. Gokhale, “Towards Robust Visual Understanding: from Recognition to Reasoning”, AAAI, vol. 38, no. 20, pp. 22665-22665, Mar. 2024.

[6] Y. Liu, Y. Wei, H. Yan, G. Li, and L. Lin, “Causal Reasoning Meets Visual Representation Learning: A Prospective Study, ” arXiv.org, 2022. https: //arxiv.org/abs/2204.12037 .

[7] J. Jiang et al., “Multimodal Contrastive Learning for 3D Object Classification and Part-Segmentation by Leveraging V-LLM and CNNs, ” The Photogrammetric Record, vol. 40, no. 190, Apr. 2025, doi: https: //doi.org/10.1111/phor.70006.

[8] H. Deng, H. Luo, W. Zhai, Y. Guo, Y. Cao and Y. Kang, "VMAD: Visual-enhanced Multimodal Large Language Model for Zero-Shot Anomaly Detection, " in IEEE Transactions on Automation Science and Engineering, doi: 10.1109/TASE.2025.3591656.

[9] Y. Wang et al., “UniPLV: Towards Label-Efficient Open-World 3D Scene Understanding by Regional Visual Language Supervision, ” arXiv.org, 2024. https: //arxiv.org/abs/2412.18131 .

[10] [P. Wu et al., "Open-Vocabulary Video Anomaly Detection, " 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2024, pp. 18297-18307, doi: 10.1109/CVPR52733.2024.01732.

[11] L. Jiang, S. Shi and B. Schiele, "Open-Vocabulary 3D Semantic Segmentation with Foundation Models, " 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2024, pp. 21284-21294, doi: 10.1109/CVPR52733.2024.02011.

[12] D. Liu et al., “Uni3D-LLM: Unifying Point Cloud Perception, Generation and Editing with Large Language Models, ” arXiv.org, 2024. https: //arxiv.org/abs/2402.03327

[13] Q. Ma et al., “APOVIS: Automated pixel-level open-vocabulary instance segmentation through integration of pre-trained vision-language models and foundational segmentation models, ” Image and Vision Computing, vol. 154, p. 105384, Dec. 2024, doi: https: //doi.org/10.1016/j.imavis.2024.105384.

[14] T. Chen et al., “Reasoning3D -- Grounding and Reasoning in 3D: Fine-Grained Zero-Shot Open-Vocabulary 3D Reasoning Part Segmentation via Large Vision-Language Models, ” arXiv.org, 2024. https: //arxiv.org/abs/2405.19326 .

[15] W. Hu, Y. Xu, Y. Li, W. Li, Z. Chen, and Z. Tu, “BLIVA: A Simple Multimodal LLM for Better Handling of Text-Rich Visual Questions”, AAAI, vol. 38, no. 3, pp. 2256-2264, Mar. 2024.

[16] B. Lin et al., “MoE-LLaVA: Mixture of Experts for Large Vision-Language Models, ” arXiv.org, 2024. https: //arxiv.org/abs/2401.15947

[17] X. Dong et al., “InternLM-XComposer2: Mastering Free-form Text-Image Composition and Comprehension in Vision-Language Large Model, ”arXiv.org, 2024. https: //arxiv.org/abs/2401.16420.

[18] T. Gokhale, “Towards robust visual understanding: A paradigm shift in computer vision from recognition to reasoning, ” AI Magazine, vol. 45, no. 3, pp. 429–435, Sep. 2024, doi: https: //doi.org/10.1002/aaai.12194.

[19] J. Liu, D. Fan, J. Shen, C. Ji, D. Zha, and Q. Tan, “Graph-MLLM: Harnessing Multimodal Large Language Models for Multimodal Graph Learning, ” arXiv.org, 2025. https: //arxiv.org/abs/2506.10282 .

DOI: 10.54254/2755-2721/2025.LD29606

[20] H. Zhao et al., “LLaVA-CMoE: Towards Continual Mixture of Experts for Large Vision-Language Models, ” arXiv.org, 2025. https: //arxiv.org/abs/2503.21227 .