# A Survey of Open Vocabulary Scene Understanding and Object Detection with Vision Large Language Models




## Abstract

In this survey, we explore the integration of vision and language models to advance open vocabulary scene understanding and object detection, overcoming the limitations of traditional methods restricted by predefined categories. By leveraging frameworks such as the Scene-Graph-Based Discovery Network (SGDN) and Open-Vocabulary SAM, we highlight significant advancements in multimodal AI that enhance interactive segmentation and recognition capabilities. This survey systematically evaluates the role of Large Language Models (LLMs) in processing 3D spatial data, contributing to the development of context-aware AI systems. We address challenges such as taxonomy conflicts and fore/background inconsistencies, which hinder traditional approaches, and propose innovative solutions like query-based knowledge sharing frameworks and volumetric neural representations. Real-world applications in autonomous vehicles, healthcare, and e-commerce are examined, demonstrating the transformative potential of these technologies. We also evaluate performance on benchmarks like LVIS, ScanNet200, and nuScenes, emphasizing the importance of overcoming ineffective alignment between visual and textual embeddings. Future research directions include improving model scalability, enhancing robustness, and addressing bias and fairness considerations to develop more versatile AI systems capable of operating in diverse environments. Ultimately, this survey provides a comprehensive overview of current advancements and proposes pathways for future exploration in open vocabulary scene understanding and object detection.


## 1 Introduction

### 1.1 Objectives of the Paper

This survey investigates the integration of vision and language models to enhance open vocabulary scene understanding and object detection, addressing the limitations of traditional methods constrained to predefined categories. It provides a comprehensive review of advancements in open-vocabulary detection and segmentation, enabling models to classify and detect objects beyond annotated datasets. Methodologies such as visual-semantic mapping and scene graph utilization are examined, leveraging weak supervision signals to improve recognition of arbitrary objects in diverse scenes. The survey aims to offer insights into future research directions in this evolving field [1, 2, 3, 4, 5]. By harnessing the synergy between these models, it seeks to enhance multimodal interactions, exemplified by frameworks like the Scene-Graph-Based Discovery Network (SGDN), which integrate visual and textual information for improved open vocabulary scene understanding.

Additionally, the survey explores innovative paradigms for object detection methodologies that facilitate few-shot learning without extensive retraining. The integration of vision and language models is crucial for linking language units to their physical referents, enhancing our understanding of how words acquire grounded meanings in the real world. This grounding process not only aids in

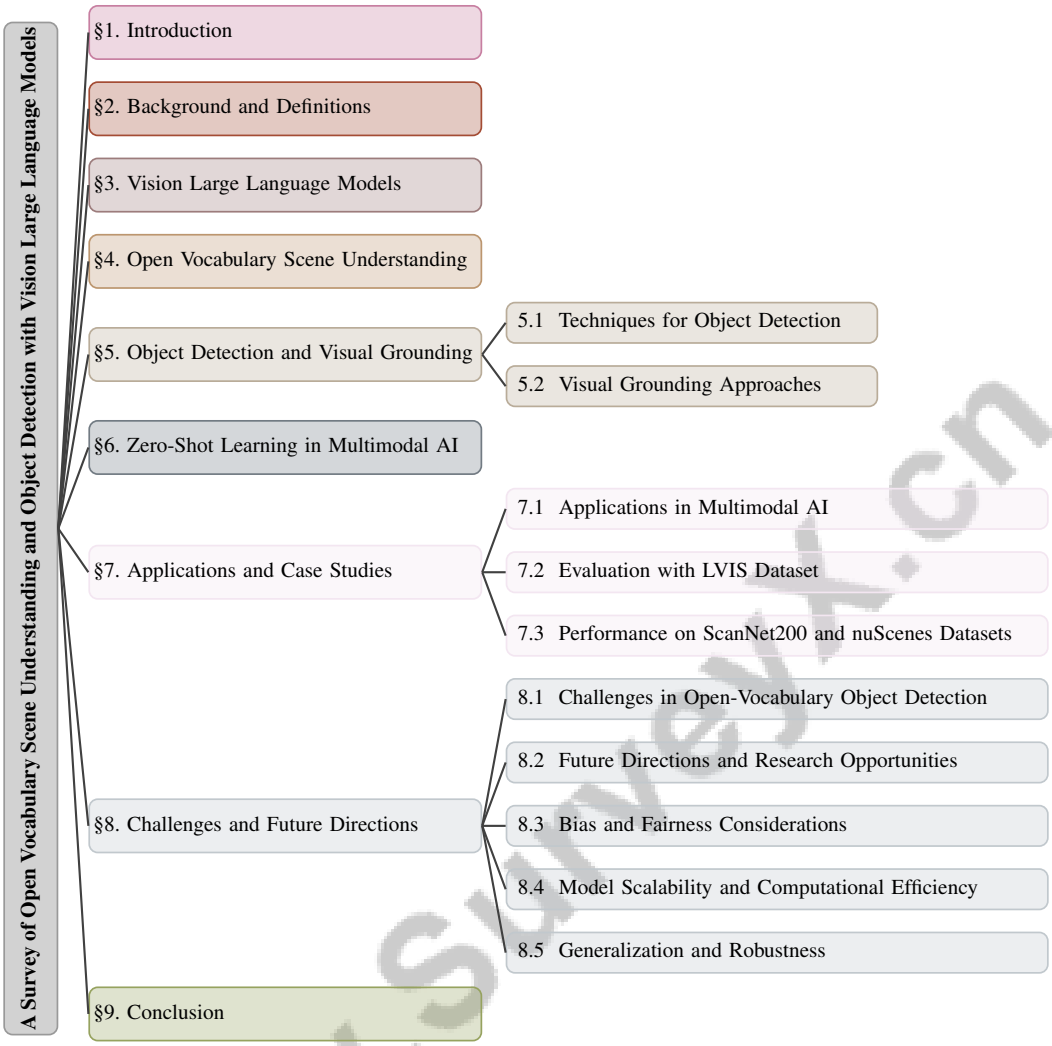



Figure 1: chapter structure

understanding existing vocabulary but also supports rapid learning of new words through mechanisms like fast mapping, as demonstrated by models such as OctoBERT and Grounded 3D-LLM. These models utilize visually-grounded representations and scene referent tokens to enhance coherence and speed in word learning, unifying various vision tasks within a generative framework. Advanced pre-training techniques and large-scale grounded datasets are paving the way for more robust interactions between language and visual information, leading to a deeper understanding of semantic concepts derived from visual inputs [6, 7, 8, 9]. Moreover, the survey addresses challenges in keypoint detection in images, particularly in a zero-shot context, which is essential for applications in mobile robotics that do not rely on prior annotations.

The potential of Large Language Models (LLMs) in processing and understanding 3D spatial data is also investigated, contributing to the development of a unified multimodal system that integrates perceptual data into reasoning and planning processes. The research highlights the critical need to address misalignment between visual and textual embeddings in open vocabulary object detection (OVOD), which undermines zero-shot learning performance due to limitations in training data. Ineffective alignment can result in lower detection accuracy for novel classes not represented during training, as traditional methods often fail to leverage pre-trained cross-modal models effectively. Enhancing feature alignment through innovative techniques, such as augmenting text embeddings, modifying detection architectures, and employing self-training methods, can significantly improve the robustness of OVOD systems, thereby enhancing their capabilities in recognizing and localizing unseen objects [10, 11, 12, 13]. By systematically evaluating these advancements, the survey aims





AI-generated, for reference only.

to provide a comprehensive overview of the current state of the field and propose future research directions, ultimately contributing to the development of versatile and context-aware AI systems.

The limitations of current object detection methods that rely on fixed vocabulary sizes are addressed, proposing breakthroughs in detection capabilities through the utilization of Vision and Language Models (VLMs) [14]. The development of the Open-Vocabulary SAM, which integrates interactive segmentation and recognition capabilities by leveraging knowledge transfer between the Segment Anything Model (SAM) and CLIP, is highlighted as a significant advancement toward achieving open vocabulary scene understanding [15].

### 1.2 Structure of the Survey

This survey is structured to provide an in-depth exploration of open vocabulary scene understanding and object detection, focusing on the integration of vision large language models. It begins with an introduction outlining the objectives and significance of combining vision and language models, followed by a background section that offers essential definitions and explores the interrelations of key concepts such as open vocabulary, scene understanding, and object detection.

The survey discusses Vision Large Language Models, emphasizing their role in synthesizing visual and textual data and showcasing advancements in multimodal AI, including frameworks like 3DVLP [16]. The section on Open Vocabulary Scene Understanding examines methodologies and challenges, featuring innovative frameworks such as O2V-Mapping [17] and benchmarks established by Ren et al. [18].

Subsequently, the survey delves into Object Detection and Visual Grounding, analyzing techniques within open vocabulary paradigms and the role of language models in enhancing detection systems, as demonstrated by approaches proposed by Kaul et al. [19]. The section on Zero-Shot Learning in Multimodal AI explores applications facilitated by vision large language models, featuring frameworks like CastDet [20] and UOVN [21].

Real-world applications and case studies illustrate the practical implementation of these technologies, evaluating models using datasets such as LVIS, ScanNet200, and nuScenes. The concluding sections identify current challenges, propose future research directions, and discuss issues of bias, fairness, scalability, and robustness, as highlighted in surveys like the one by Zhu et al. [1]. The survey concludes by synthesizing key findings and emphasizing the transformative impact of integrating vision and language models in AI applications.The following sections are organized as shown in Figure 1.

## 2 Background and Definitions

### 2.1 Challenges and Interrelations of Core Concepts

Integrating open vocabulary scene understanding and object detection within vision large language models involves complex challenges, primarily due to the diverse nature of visual instances and textual queries. A significant issue is the taxonomy conflict and fore/background inconsistency across datasets with varied annotation schemes, complicating the functionality of traditional detectors [22]. Additionally, the reliance on closed-set semantic labels restricts the adaptability of segmentation methods in dynamic settings [23].

Methodologies often falter at pixel-level localization, as simplistic text prompts lack the semantic depth needed for effective segmentation [24]. The combination of models like SAM and CLIP reveals inefficiencies in recognizing small objects and adapting representations across datasets [15]. Furthermore, the extensive manual annotation required for bounding boxes and instance masks limits vocabulary size and detection systems' scalability [14].

In mobile manipulation, robots struggle without prior knowledge or accurate user hints, complicating the integration of open vocabulary with scene understanding [25]. Existing benchmarks often rely on synthetic images and simple object categories, restricting their relevance to complex real-world tasks [26].

These challenges necessitate innovative approaches for dynamically integrating multimodal data and adapting to various contexts. Developing comprehensive benchmarks and methodologies is crucial

for enhancing the understanding of novel concepts in visual contexts, advancing open vocabulary scene understanding and object detection. Moreover, aligning visual features with language models is problematic due to noise in raw image-text pairs, resulting in unreliable language outputs [27]. Improved methods are needed to capture detailed visual concepts and contexts articulated in linguistic descriptions, thereby enhancing the accuracy of target object identification [28].

In recent years, the field of artificial intelligence has witnessed significant advancements, particularly in the realm of multimodal AI. These developments have notably influenced vision large language models (LLMs), which are increasingly capable of understanding and interpreting complex visual information. Figure 2 illustrates these advancements, highlighting key innovations in open vocabulary scene understanding and 3D object detection. The figure underscores the transformative impact these innovations have on the functionality and effectiveness of vision LLMs, providing a visual representation that complements the discussion of their evolution and application in various domains. This integration of visual and textual data not only enhances the models' capabilities but also paves the way for more sophisticated interactions between humans and machines.

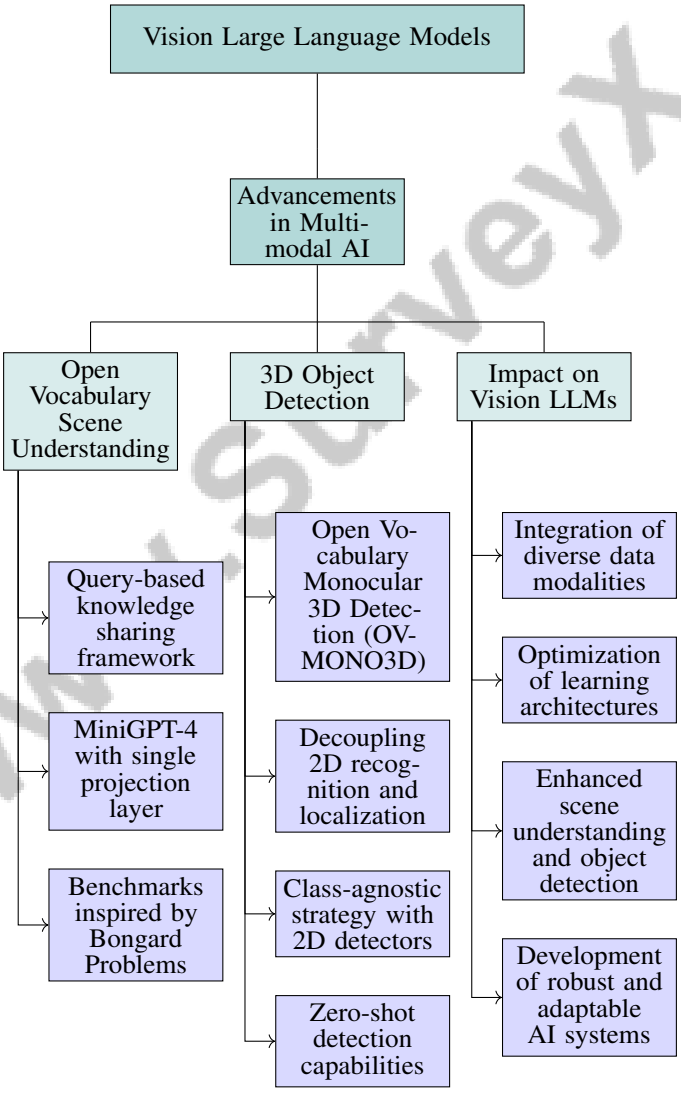



Figure 2: This figure illustrates the advancements in multimodal AI impacting vision large language models, highlighting key innovations in open vocabulary scene understanding, 3D object detection, and their overall transformative impact on vision LLMs.

# 3 Vision Large Language Models

## 3.1 Advancements in Multimodal AI

Recent progress in multimodal AI has significantly enhanced the capabilities of vision large language models (LLMs) in open vocabulary scene understanding and object detection. The introduction of a query-based knowledge sharing framework marks a substantial improvement over traditional knowledge distillation methods, facilitating more effective extraction and sharing of visual information. This framework, exemplified by MiniGPT-4, employs a single projection layer to align visual features with LLMs, offering a streamlined alternative to complex architectures [29, 27].

Benchmarks inspired by classical Bongard Problems present models with open-world concepts and real images, emphasizing the need for systems capable of complex visual reasoning in dynamic environments [26]. These benchmarks highlight the critical importance of generalization across diverse and previously unseen categories in open vocabulary scene understanding.

In 3D object detection, the Open Vocabulary Monocular 3D Detection (OVMONO3D) approach has transformed object recognition and localization. By decoupling 2D recognition and localization from the estimation of 3D bounding boxes, this method allows models to generalize to unseen object categories without extensive retraining. OVMONO3D utilizes a class-agnostic strategy, leveraging open-vocabulary 2D detectors to project 2D bounding boxes into 3D space, thereby enhancing adaptability in varied real-world environments. A target-aware evaluation protocol addresses inconsistencies in existing datasets, improving performance assessment reliability. This approach demonstrates robust zero-shot detection capabilities across novel object categories, highlighting its potential in advancing open-vocabulary detection models [13, 30, 31].

These innovations underscore the transformative impact of multimodal AI on vision LLMs, driving progress by integrating diverse data modalities and optimizing learning architectures. They enhance scene understanding and object detection accuracy and efficiency, enabling machine vision systems to recognize a wide array of visual concepts, textures, and relationships within images. Furthermore, they contribute to developing more robust and adaptable AI systems capable of effective operation in complex, real-world environments, as evidenced by technologies like Unified Perceptual Parsing, dynamic input scaling in Multimodal Large Language Models (MLLMs), and goal-driven text generation for images [32, 33, 8, 34].

# 4 Open Vocabulary Scene Understanding

## 4.1 Open Vocabulary Learning in Scene Understanding

Open vocabulary learning is essential for AI systems to interpret diverse categories beyond fixed vocabularies, with a key challenge being the alignment of visual and textual representations to enhance zero-shot detection. The Scene-Graph-Based Discovery Network (SGDN) addresses this by utilizing scene graph cues in open vocabulary object detection, highlighting structured information's role in scene comprehension [3]. The Open Vocabulary Scene Parsing (OVSP) method exemplifies segmenting objects using open vocabulary labels by integrating word concepts and image features into a joint high-dimensional space, facilitating seamless multimodal data alignment and generalization across categories [2]. Similarly, LMSeg improves segmentation accuracy by combining linguistic prompts from large language models with visual features from CLIP and SAM models [24].

Innovative frameworks, such as those by Blomqvist et al., integrate volumetric neural representations with vision-language features for real-time segmentation without domain-specific tuning, demonstrating potential for flexible scene understanding systems [23]. The Bongard-OpenWorld benchmark challenges models with dynamic visual reasoning tasks beyond fixed structures or synthetic images [26]. The F-VLM methodology employs a frozen vision-language model for feature extraction, combining detection scores with VLM predictions for open-vocabulary recognition, exemplifying robust feature extraction integration with open vocabulary frameworks [14].

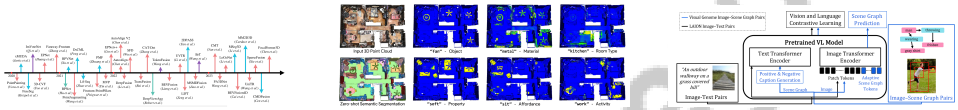
Zhu et al. propose a query-based knowledge sharing framework to enhance performance in recognizing unseen labels, addressing limitations of knowledge distillation in exploring multimodal knowledge in vision-language pre-training models [29]. This highlights the need for innovative strategies to harness and integrate multimodal knowledge, advancing open vocabulary scene understanding.

### 4.2 Innovative Frameworks for Scene Understanding

Innovative frameworks and benchmarks have emerged to enhance scene understanding, expanding AI systems' interpretation of complex visual environments. The V2P method significantly reduces word error rates compared to prior approaches and scales effectively to large vocabularies, offering a robust solution for open vocabulary scene understanding [35]. The O2V-Mapping framework integrates online open vocabulary mapping with real-time scene analysis, allowing dynamic adaptation to new visual categories without extensive retraining, thereby enhancing accuracy and speed of scene interpretation [17].

Blomqvist et al. demonstrate the use of neural implicit vision-language features for real-time segmentation and understanding of scenes without domain-specific tuning, integrating volumetric representations with vision-language models for comprehensive scene understanding [23]. Benchmarks like Bongard-OpenWorld introduce challenges for AI models to manage open vocabulary and free-form visual concepts in dynamic environments, vital for assessing models' generalization capabilities across diverse, unseen categories [26].

Zhu et al. highlight the potential for improving recognition of unseen labels through query-based knowledge sharing frameworks, effectively utilizing multimodal knowledge embedded in vision-language pre-training models [29]. This underscores the importance of innovative methodologies that harness diverse data modalities, enhancing scene understanding systems' adaptability and accuracy.




(a) Timeline of 3D Object Detection and Segmentation Techniques[36]

(b) Zero-shot Semantic Segmentation of 3D Point Clouds: A Comparative Study[37]

(c) Scene Graph Prediction with Contrastive Learning and Pretrained Vision-Language Model[38]

Figure 3: Examples of Innovative Frameworks for Scene Understanding

As depicted in Figure 3, innovative frameworks in open vocabulary scene understanding demonstrate significant advancements. The first framework traces the evolution of 3D object detection and segmentation techniques from 2020 to 2023, indicating rapid field development. The second framework evaluates zero-shot semantic segmentation of 3D point clouds, emphasizing flexible scene understanding by classifying objects, materials, room types, and activities without prior exposure to specific data categories. The third framework presents scene graph prediction through a flowchart integrating contrastive learning with a pretrained vision-language model, highlighting the synergy between visual and linguistic information in enhancing scene comprehension. Collectively, these frameworks represent a forward-thinking approach to scene understanding, leveraging cutting-edge techniques to expand AI's capabilities in interpreting complex visual data [36, 37, 38].

## 5 Object Detection and Visual Grounding

### 5.1 Techniques for Object Detection

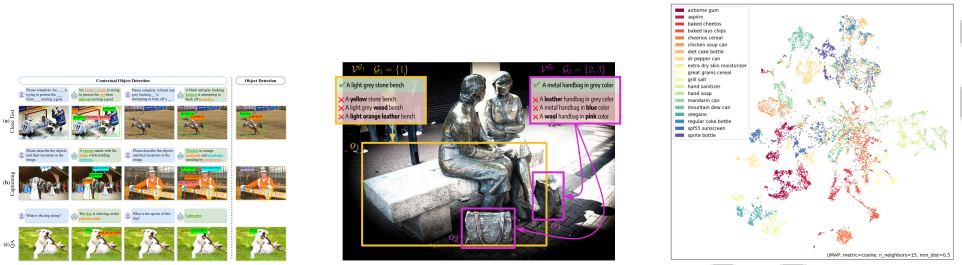
The evolution of open vocabulary object detection has led to numerous techniques that enhance adaptability and precision. Transforming detection models into few-shot learning frameworks, such as the Target-Driven Instance Detection (TDID), leverages image embeddings for few-shot capabilities, effectively adapting the YOLO-World model [39]. Standard metrics like Mean Intersection over Union (mIoU), Pixel Accuracy, and Stroke Accuracy are vital for evaluating model efficacy across diverse categories and tasks [40].

Lightweight modular systems, utilizing pre-trained models like ResNet and RoBERTa, enhance detection accuracy by freezing weights and training a Universal Projection (UP) module with a learnable modality token for image-text feature switching [41]. CondHead designs further improve performance by parameterizing bounding box regression and segmentation heads with semantic embeddings, enhancing adaptability to new vocabularies [42]. The Enhanced Object Detection

Method (EODM) refines model structures and loss functions for efficient detection across numerous classes [43].

The Grounding Everything Model (GEM) uses self-attention to cluster similar tokens, aligning them with language embeddings to improve localization precision in open vocabulary contexts [44]. New benchmarking protocols, including dynamic vocabulary generation and challenging negative examples, assess detection system robustness in fine-grained scenarios [45].

Advanced techniques in open-vocabulary detection (OVD) address challenges in multimodal data fusion and vocabulary adaptability, enhancing localization and classification accuracy. Utilizing pre-trained vision-language models and innovative methods like proposal mining and prediction equalization further strengthens OVD frameworks [1, 46, 13].



(a) Contextual Object Detection[47]

(b) Scene of a statue of two women sitting on a stone bench with a metal handbag[45]

(c) UMAP: Metric = Cosine, $N_neighbors = 15, Min_dist = 0.5$[39]

Figure 4: Examples of Techniques for Object Detection

As depicted in Figure 4, object detection and visual grounding are crucial for machine interpretation of visual data. "Contextual Object Detection" highlights tasks like cloze tests and QA, emphasizing object identification in complex scenes. The "Scene of a statue" example illustrates nuanced detection in real-world settings where visual and contextual elements are pivotal. The "UMAP" example shows dimensionality reduction aiding in visualizing high-dimensional data, underscoring advancements in object detection integral to intelligent systems [47, 45, 39].

### 5.2 Visual Grounding Approaches

Visual grounding, essential in multimodal AI, aligns visual elements in images with textual descriptions to enhance understanding. Vision and language model integration has led to approaches that improve grounding effectiveness. Attention mechanisms, particularly in Transformer architectures, enhance alignment between image regions and textual phrases for fine-grained localization tasks [28]. These models use self-attention layers to capture contextual dependencies, focusing on relevant visual features.

Large-scale pre-trained models like CLIP align visual and textual embeddings in a shared space, enabling zero-shot grounding capabilities [15]. This method leverages pre-training knowledge, allowing models to generalize across diverse, unseen categories without fine-tuning. Contrastive learning further refines grounding by distinguishing positive and negative image-text pairs, enhancing visual-linguistic associations.

Frameworks incorporating volumetric representations with vision-language features, as demonstrated by Blomqvist et al., show potential in real-time segmentation and grounding without domain-specific tuning [23]. These approaches leverage volumetric data's spatial and semantic richness, improving grounding accuracy in dynamic environments.

Benchmarks like Bongard-OpenWorld challenge models with open vocabulary and free-form visual concepts, crucial for advancing visual grounding [26]. These benchmarks assess models' adaptability to novel visual contexts, emphasizing robust system development.

Recent advancements in visual grounding methodologies highlight the integration of vision and language models in multimodal AI, significantly improving grounding task accuracy and flexibility. Transformer-based models enhance target localization with text-conditioned features and multi-stage

reasoning. The Grounding LMM (GLaMM) model enables natural language generation linked with object segmentation, enriching user interaction. The Grounding Everything Module (GEM) demonstrates zero-shot object localization without fine-tuning, showcasing potential for broader applications in diverse environments. These advancements refine grounding task performance and expand multimodal AI applications [7, 48, 28, 44, 8].

# 6 Zero-Shot Learning in Multimodal AI

## 6.1 Conceptual Foundations of Zero-Shot Learning

Zero-shot learning (ZSL) is a transformative approach in AI, enabling models to identify and categorize objects or actions absent from the training dataset, which is vital in areas with limited labeled data like remote sensing and dynamic environments [49]. ZSL employs semantic embeddings to bridge seen and unseen categories, thus facilitating generalization to new instances without direct training. A significant challenge in ZSL is semantic misalignment due to lexical variations, which the Synonymous Semantic Space (S3) method addresses by constructing a semantic space for each image class using multiple synonymous concepts, thereby enhancing generalization capabilities [50].

In multimodal AI, integrating ZSL with vision-language models enhances the interpretation of complex visual scenes. For instance, integrating visual object detection into language frameworks significantly improves decision-making in applications like autonomous driving, where recognizing unseen objects is crucial [51]. The Query-Based Knowledge Sharing (QKS) framework exemplifies an innovative approach to filter out irrelevant visual information, emphasizing crucial features to enhance the recognition of unseen labels [29]. The application of ZSL in multi-object navigation tasks, as demonstrated by the OneMap framework, underscores the value of high-quality semantic representations in improving navigation performance [52]. Furthermore, leveraging multiple remote sensing modalities enhances model understanding and classification capabilities by addressing performance gaps in existing methods [53].

The integration of ZSL with multimodal AI has led to significant advancements, including models that generate descriptive text from visual inputs, dynamically adapt to varying image resolutions, and utilize innovative frameworks for zero-shot semantic segmentation. These developments enable AI systems to function more efficiently in diverse environments, enhancing human-like understanding and interaction capabilities [54, 32, 8, 34]. This integration not only enhances the versatility of AI systems but also paves the way for robust, context-aware applications across various domains.

## 6.2 Zero-Shot and Few-Shot Learning

Zero-shot and few-shot learning are pivotal in AI, enabling models to recognize and classify new categories with minimal or no explicit training data. These methodologies are especially beneficial in contexts where labeled data is scarce or costly, such as remote sensing and dynamic environments. Vision-language models (VLMs), pre-trained on extensive image-text pairs, significantly enhance zero-shot classification accuracy by leveraging semantic embeddings to connect seen and unseen categories [49]. ZSL generalizes to new classes using auxiliary information like class descriptions or attributes, which is vital in fields like autonomous navigation and remote sensing, where unseen objects frequently occur.

The integration of VLMs within ZSL frameworks improves the alignment of visual and textual modalities, enhancing the model's ability to interpret complex scenes and make informed decisions. Techniques such as dynamic input scaling optimize image processing at varying resolutions based on content, while structured representations like scene graphs enrich the model's understanding of object attributes and relationships. Employing fine-grained textual descriptors and hierarchical prompts achieves precise region-text alignment, essential for effective open-vocabulary object detection and autonomous driving applications [38, 51, 55, 34, 56].

Few-shot learning extends this concept by enabling models to learn from a small number of examples, thus reducing reliance on large labeled datasets. This approach is particularly effective in domains requiring rapid adaptation to new tasks, such as personalized recommendation systems and adaptive user interfaces. Few-shot learning techniques often employ meta-learning strategies to enhance models' ability to learn from limited data, improving flexibility across various applications. For instance, models like Flamingo leverage architectural innovations to bridge vision and language

modalities, facilitating rapid adaptation to tasks such as visual question-answering and captioning with minimal annotated examples. Similarly, AdaptVision employs dynamic input scaling for efficient image interpretation based on content density, while frameworks like Open-VCLIP++ adapt existing models for zero-shot video classification, achieving state-of-the-art performance with minimal additional training data [57, 58, 34].

The distinctions between zero-shot and few-shot learning lie in their methodologies for addressing novel categories. ZSL utilizes pre-trained models to recognize unseen categories without additional examples, while few-shot learning adapts to new tasks using a limited number of annotated examples, requiring some prior exposure to the task domain [57, 54, 59, 60]. While ZSL relies on semantic embeddings and auxiliary information for generalization, few-shot learning leverages limited training examples for rapid adaptation. Both paradigms aim to enhance the versatility and robustness of AI systems, enabling effective operation in complex environments.

The integration of zero-shot and few-shot learning with vision-language models represents a significant advancement in artificial intelligence, driving progress in applications that necessitate adaptive and context-aware decision-making capabilities. These methodologies not only enhance the efficiency and accuracy of classification tasks across various domains but also contribute to developing robust AI systems capable of adapting and performing effectively in diverse environments. By leveraging techniques such as goal-driven text generation from visual inputs and improved feature alignment in multimodal models, these approaches deepen the understanding of cross-modal interactions and the generation of meaningful outputs, ultimately propelling progress toward achieving Artificial General Intelligence (AGI) [10, 32, 8].

# 7 Applications and Case Studies

## 7.1 Applications in Multimodal AI

Multimodal AI is revolutionizing various sectors by integrating visual and linguistic data to improve decision-making and user experiences. In autonomous vehicles, the fusion of visual object detection with language models enhances the interpretation of complex driving scenes, thereby improving navigation safety and efficiency [51]. This capability is essential for recognizing diverse road scenarios, including unseen objects, thus advancing autonomous driving technologies.

In healthcare, the combination of medical imaging and textual patient data enables more accurate diagnoses and personalized treatment plans [49]. By integrating diagnostic images with clinical histories, these systems provide comprehensive insights that improve patient outcomes and streamline healthcare workflows.

The entertainment industry leverages multimodal AI for interactive media and virtual assistants, where vision-language models enable systems to understand and respond contextually to user queries, enhancing engagement. Virtual reality environments use multimodal AI to create immersive experiences by combining visual elements with narrative content, thereby improving user interaction [58].

In e-commerce, multimodal AI enhances product recommendation systems by analyzing visual product features alongside customer reviews, providing personalized recommendations and improving the shopping experience [50].

Moreover, in remote sensing, the integration of multimodal AI facilitates the analysis of satellite imagery with geographical and environmental data, improving natural resource monitoring and management [53]. This is crucial in agriculture, urban planning, and disaster response, where timely information is vital for effective decision-making.

The diverse applications of multimodal AI highlight its transformative potential in enhancing the functionality and adaptability of AI systems. By integrating various data modalities, these advanced systems deliver comprehensive insights, optimize operational efficiency, and create tailored experiences across domains such as object detection, 3D scene understanding, and document comprehension. This multimodal approach enriches data interpretation and ensures robustness in challenging environments, leading to superior performance across various applications [61, 32, 36, 62].

| Benchmark | Size | Domain | Task Format | Metric |
|---|---|---|---|---|
| OneMap[52] | 236 | Object Navigation | Multi-Object Navigation | Success Rate, Success weighted by Path Length |
| CLIP-Openness[18] | 1,000,000 | Visual Recognition | Image Classification | Extensibility, Stability |
| OSR-VLM[63] | 50,000 | Object Detection | Open-Set Recognition | AuPR, mAP |
| SceneVerse[64] | 2,500,000 | 3D Vision-Language | Visual Grounding | Acc@0.5 |
| ZVLB[60] | 25,215 | Computer Vision | Image Classification | AP, ECE |
| VALUE[8] | 1,531,448 | Visual Relation Detection | Visual Coreference Resolution | NMI |
| PRISM[55] | 558,000 | Visual Question Answering | Visual Question Answering | Accuracy, F1-score |
| OO3D-9D[65] | 5,000,000 | Robotics | Object Pose Estimation | Rot |
| Trans precision, 3D IoU | | | | |

Table 1: This table presents a comprehensive overview of various benchmarks utilized in the evaluation of open vocabulary models across different domains. It details the size, domain, task format, and performance metrics for each benchmark, providing a resource for assessing the capabilities of models in object navigation, visual recognition, object detection, and more.

### 7.2 Evaluation with LVIS Dataset

The LVIS dataset serves as a crucial benchmark for evaluating open vocabulary object detection models, particularly in their ability to generalize to novel and rare categories. The ViLD model demonstrated its effectiveness in open vocabulary settings when tested against baseline methods on the LVIS dataset [66]. Similarly, the MIC method excelled in generalizing to unseen classes, outperforming state-of-the-art models [67].

Performance metrics such as average precision (AP) and mean average precision (mAP) are commonly used to assess object detection and instance segmentation capabilities. For example, the F-VLM model showed improvements in Mask AP and training efficiency on LVIS and COCO benchmarks compared to existing methods [68]. The LaMiDeT model also demonstrated its performance on rare classes within the LVIS dataset using mAP [69].

The LVIS dataset's extensive categorization, featuring 337 rare novel categories, provides a robust framework for testing models like the one proposed by Li et al., on both OV-COCO and OV-LVIS tasks [70]. This comprehensive evaluation allows for a detailed assessment of a model's detection capabilities across a wide range of categories, including less frequently encountered ones.

Moreover, the method proposed by Chen et al. was evaluated on both COCO and LVIS datasets, using metrics like mean Average Precision (mAP) and AP50 to benchmark its performance against leading open vocabulary detection methods [46]. The extensive evaluation across multiple datasets, including LVIS, underscores the importance of diverse benchmarks in validating the effectiveness and generalization capabilities of open vocabulary object detection models. Table 1 provides a detailed overview of representative benchmarks critical for evaluating the performance and generalization capabilities of open vocabulary models across diverse domains.

The LVIS dataset is pivotal for advancing open vocabulary object detection by providing a diverse array of categories that challenge existing models, fostering innovation in detection methodologies. By incorporating a wide range of object classes, including rare and fine-grained categories, the LVIS dataset enables researchers to develop and test novel approaches, such as those utilizing vision-language models and knowledge distillation techniques, which enhance alignment between visual and textual representations. This evolution in detection strategies not only pushes the boundaries of current capabilities but also facilitates the transfer of learned models across various datasets, broadening the scope of object detection applications [46, 66, 71, 56].

### 7.3 Performance on ScanNet200 and nuScenes Datasets

Evaluating open vocabulary scene understanding and object detection models on the ScanNet200 and nuScenes datasets provides critical insights into their performance in complex 3D environments. The ScanNet200 dataset is widely used for assessing 3D scene understanding, serving as a benchmark for comparing open vocabulary models against closed-vocabulary methods. For instance, the OpenMask3D framework demonstrated substantial improvements over baseline methods, including closed-vocabulary fully supervised models, showcasing its effectiveness in open vocabulary settings [72]. Similarly, OpenIns3D achieved state-of-the-art results on ScanNet200 by focusing exclusively on 3D inputs, highlighting the potential of 3D-centric approaches in open vocabulary tasks [73].

The PoVo method was also evaluated on ScanNet200, demonstrating its effectiveness in both open-vocabulary and vocabulary-free settings, underscoring the versatility of open vocabulary frameworks in 3D scene understanding [74]. The OV-SAM3D model, tested on both ScanNet200 and nuScenes datasets, exhibited competitive performance against existing methods, validating its applicability in open vocabulary 3D detection tasks [75].

The nuScenes dataset, known for its comprehensive representation of urban driving scenarios, provides a robust platform for evaluating the adaptability and generalization capabilities of open vocabulary models in dynamic environments. The UniMoV3D framework, assessed on public 3D benchmarks including ScanNet, ScanNet200, and nuScenes, utilized metrics such as hIoU, mIoUB, and mAP to evaluate performance, highlighting its ability to effectively integrate unimodal and open vocabulary approaches [76].

Furthermore, the dual-path integration framework evaluated on ScanNet200 demonstrated proficiency in surpassing existing closed-vocabulary methods, thereby advancing the state-of-the-art in open vocabulary scene understanding [77]. The role of language models in 3D scene understanding, as explored in SceneGPT, further emphasizes the importance of semantic-rich embeddings in enhancing model performance on datasets like ScanNet [78].

The performance analysis on the ScanNet200 and nuScenes datasets highlights the significant advancements achieved by open vocabulary models in 3D scene understanding and object detection. These models transcend the limitations of traditional closed-set approaches, enabling the identification and recognition of categories not explicitly included in the training data. This capability enhances the robustness and adaptability of AI systems, equipping them to navigate and operate effectively in complex, real-world environments, thus paving the way for more versatile applications in open-world scenarios [75, 5].

## 8 Challenges and Future Directions

In open vocabulary object detection (OVD), recognizing the challenges inherent in current methodologies is essential for advancing the field. This section delves into the specific obstacles faced by open-vocabulary systems, focusing on the complexities of identifying and localizing objects without predefined categories. By identifying these challenges, we can pinpoint critical areas for improvement and innovation, paving the way for more effective solutions.

### 8.1 Challenges in Open-Vocabulary Object Detection

Open-vocabulary object detection (OVD) presents significant challenges due to the need to recognize and localize objects beyond fixed categories. Unlike traditional systems, OVD leverages pre-trained cross-modal vision-language models (VLMs) like CLIP and ALIGN to identify unseen objects using an unbounded vocabulary. However, these systems often struggle with weak supervision and zero-shot learning limitations, necessitating strategies for accurately detecting novel categories without extensive annotations. Recent advancements focus on integrating multi-modal classifiers and addressing misclassification issues, aiming for robust, scalable solutions [1, 11, 19, 79, 13]. A key issue is the reliance on high-quality labels and mask proposals, which can lead to mislabeling and hinder accurate classification, especially for novel categories tied to their contexts. This highlights the need for adaptable methods that enhance detection systems' flexibility in open-vocabulary settings.

Another challenge is the computational burden of existing methods, which often require significant resources for retraining or fine-tuning [39]. This limitation affects their applicability in personalized detection scenarios, where efficiency and scalability are crucial. Moreover, the reliance on generated superpoints, as seen in frameworks like OV-SAM3D, can impact performance in varied scenes [75].

Calibration issues also arise, particularly in out-of-domain tasks where models may struggle to generalize due to dataset-specific training [80]. This underscores the need for improved calibration techniques to enhance robustness across diverse environments. Integrating vision-language models for dense prediction tasks, such as semantic segmentation and object detection, remains an area requiring further exploration [81].

Furthermore, limitations of methods like FOCUS complicate open-vocabulary detection in complex scenes with overlapping objects [82]. Addressing overfitting to base categories and reliance on CLIP's

text embeddings is essential for improving generalization to novel classes [1]. The computational overhead and input resolution constraints of large language models (LLMs) further hinder real-time applications, necessitating more efficient algorithms [51].

Current models also struggle with open vocabulary concepts and multi-image reasoning, limiting their effectiveness on benchmarks like Bongard-OpenWorld [26]. Vulnerability to membership inference attacks in open-vocabulary video models raises privacy concerns [58]. Additional hurdles include recognizing small objects and dependency on visual prompts for segmentation [15].

Future research should focus on expanding datasets, refining class embedding techniques, and developing robust models for challenging scenarios, such as camouflaged object segmentation [83]. Incorporating diverse datasets and task types, along with strategies to mitigate the impact of unannotated objects during training, could enhance model performance [84]. Addressing these challenges will propel the field towards more robust and adaptable open-vocabulary detection systems capable of dynamic operation.

### 8.2 Future Directions and Research Opportunities

Future research in open vocabulary scene understanding and object detection is poised to explore several promising avenues to enhance AI systems' robustness and adaptability. A critical direction involves refining language model prompts and optimizing visual aggregators to improve detection capabilities while expanding datasets for comprehensive training [19]. Additionally, expanding benchmarks to include additional tasks and modalities, along with improving evaluation metrics, will facilitate a more holistic assessment of model performance [85].

Enhancing model robustness in extreme conditions and integrating diverse data sources could significantly improve models like OV-UNI3DETR, particularly in challenging environments [31]. Exploring object-level discrete representations using Neural Radiance Fields (NeRF) and improving performance on complex indoor datasets could advance understanding in 3D open vocabulary contexts [86].

Refining the proposal mining process and addressing class imbalance in open-vocabulary detection remain crucial areas for development, enhancing model accuracy and generalization [46]. Investigating few-shot learning settings for open-vocabulary domain adaptation and adaptation methods for coupled open vocabulary systems could lead to more versatile models capable of transferring knowledge across domains [87].

Exploring enhancements to segment matching processes and applying methods like ZeroSeg to other vision tasks could broaden the applicability of open vocabulary frameworks [88]. Incorporating richer visual features and adapting Abstract Meaning Representations (AMRs) for better performance in capturing complex visual semantics could further refine model capabilities [6].

Scaling models like STIC to larger architectures and leveraging greater amounts of unlabeled data could significantly improve performance, particularly in settings with limited labeled data [89]. Additionally, exploring the synergy between Large Language Models (LLMs) and Vision Language Models (VLMs) for various open vocabulary dense prediction tasks could unlock new capabilities and enhance model effectiveness [56].

By addressing these research opportunities, the field can progress toward more robust and adaptable open-vocabulary scene understanding and object detection systems capable of effectively operating in diverse environments. Future directions include exploring more powerful vision-language models and enhancing fine-grained semantic segmentation capabilities [90]. Improving models' capabilities to handle complex datasets and ensuring robustness in diverse real-world scenarios is crucial [91]. Additionally, investigating sophisticated language parsing techniques and expanding the range of language data to enhance 3D feature learning are promising avenues [92]. Methods to improve the integration of heterogeneous annotations and enhance model performance in recognizing complex visual relationships should also be explored [33]. Enhancements in temporal modeling and the integration of advanced LLMs to broaden the range of recognizable concepts are suggested [93]. Developing a unified backbone that merges the capabilities of SAM and CLIP could enhance performance while reducing model complexity [94]. Future work could also explore enhancements in scene graph extraction techniques and the integration of additional modalities to further improve detection capabilities [3]. Improving text prompt encoding methods and exploring new pre-training strategies to

enhance models' zero-shot detection performance are also critical [22]. Addressing feature extraction challenges, dynamic scenes, and optimizing SLAM integration with the segmentation process are potential research directions [23]. Lastly, developing models capable of holistic perception and reasoning across multiple images is essential for future exploration [26].

### 8.3 Bias and Fairness Considerations

The deployment of multimodal AI systems, particularly in open vocabulary scene understanding and object detection, necessitates a careful examination of bias and fairness issues. Research by Ma et al. highlights potential biases inherent in datasets used for training these models, which can lead to skewed outcomes in AI applications [9]. These biases often stem from imbalanced data representation, where certain categories or demographics are underrepresented, resulting in models that may not perform equitably across diverse user groups.

Addressing these biases requires a multifaceted approach, starting with diversifying training datasets to include a broader range of scenarios and demographics. Curating datasets representative of diverse contexts in which AI systems are deployed is essential. Advancements in bias detection and mitigation algorithms are crucial for ensuring fairness and transparency in the rapidly evolving field of vision-and-language models, especially as these systems increasingly influence decision-making processes across various applications [32, 8]. These algorithms can identify and correct imbalances during training, ensuring that model predictions are not unduly influenced by skewed data distributions.

In addition to technical advancements, addressing the ethical implications of deploying AI models in sensitive domains such as surveillance, healthcare, and autonomous decision-making is vital, as issues of privacy, accountability, and bias can significantly impact individuals and society [95, 6, 96, 32, 8]. Ensuring transparency in model development and decision-making processes can help build trust and accountability, including providing clear documentation on model training and validation processes, as well as potential limitations and biases.

Incorporating fairness constraints into model training is another promising avenue. Strategically implementing these constraints can ensure equitable model performance across various demographic groups, minimizing the risk of biased outcomes, such as the gender-based disparities observed in zero-shot vision-language models. This approach addresses calibration differences in model accuracy and confidence related to perceived gender and mitigates the propagation of existing biases from foundational models, ensuring a more fair and inclusive application of AI technologies across diverse populations [32, 60, 34]. Regular audits and evaluations of AI systems post-deployment can help identify and rectify emerging biases, ensuring that models remain fair and equitable in their applications.

Addressing bias and fairness in multimodal AI remains a complex and ongoing challenge, necessitating sustained efforts to ensure that models like large vision-language models (LVLMs) do not perpetuate or exacerbate existing biases, particularly as they increasingly integrate visual perception and language generation capabilities. Recent studies highlight issues such as object hallucination, where models generate descriptions of non-existent objects, and the tendency of pre-trained models to prioritize textual over visual information during inference, underscoring the critical need for continuous evaluation and refinement of these systems to promote fairness and accuracy in their outputs [97, 98, 32, 8]. By adopting comprehensive strategies encompassing dataset diversification, algorithmic fairness, and ethical considerations, the AI community can work towards developing systems that are effective and equitable, ultimately enhancing the trust and reliability of AI technologies in society.

### 8.4 Model Scalability and Computational Efficiency

Model scalability and computational efficiency are pivotal concerns in developing open vocabulary scene understanding and object detection systems. Integrating Large Language Models (LLMs) presents significant challenges, particularly in processing and understanding intricate 3D details without sufficient contextual information [99]. This limitation underscores the need for enhanced data diversity and improved training methodologies to ensure comprehensive model comprehension and performance.

The computational demands of current models often necessitate substantial resources, hindering their applicability in real-time and resource-constrained environments. The complexity of model architectures, coupled with the need for extensive data processing, can lead to inefficiencies that limit scalability. The necessity for precise and efficient alignment of visual and textual information becomes particularly pronounced in scenarios involving multimodal data integration, such as in advanced AI systems like InternLM-XComposer, which facilitates coherent article generation combining images and text while emphasizing meaningful descriptions and diverse captions to enhance human-computer interaction [100, 32].

To address these challenges, research must focus on optimizing model architectures to reduce computational overhead while maintaining performance. This includes exploring lightweight model designs and efficient training algorithms that leverage distributed computing resources. Recent advancements in hardware acceleration, particularly through Graphics Processing Units (GPUs) and Tensor Processing Units (TPUs), have significantly improved the performance of multimodal large language models (MLLMs) by enhancing processing speed and scalability. These improvements are crucial for efficiently handling complex tasks involving both visual and textual data, as seen in innovations like AdaptVision, which dynamically adjusts input processing based on image resolution and content, and frameworks like LAMM, which facilitate integrating diverse modalities for more effective human-AI interaction [85, 100, 34, 101].

Moreover, developing adaptive models that dynamically adjust their complexity based on task requirements and available resources is essential. These advanced multimodal models facilitate deploying sophisticated vision-language capabilities across a diverse array of devices, from high-performance servers equipped for intensive processing to edge devices with constrained computational resources, enabling seamless integration of text and image comprehension in various applications [100, 32, 34, 27]. By prioritizing scalability and efficiency, the field can advance towards more robust and versatile AI systems capable of operating effectively in diverse and dynamic environments.

### 8.5 Generalization and Robustness

The generalization and robustness of models in open vocabulary scene understanding and object detection are critical for ensuring reliable performance across diverse and unseen environments. Generalization refers to the model's ability to apply learned knowledge to novel scenarios, while robustness pertains to its resilience against variations and perturbations in input data. The challenges in achieving these attributes are underscored by the complexities inherent in dynamic and heterogeneous environments, where models must adapt to new visual categories and contexts without extensive retraining [26].

A significant factor influencing generalization is the quality and diversity of training data. Models trained on datasets encompassing a wide range of categories and scenarios are more likely to generalize effectively to new contexts. This is particularly important in open vocabulary settings, where models must recognize and understand categories beyond those explicitly seen during training. Techniques such as domain adaptation and transfer learning can enhance generalization by leveraging knowledge from related tasks or domains [87].

Robustness is crucial for maintaining model performance in varying input conditions, such as changes in lighting, occlusions, or noise. Integrating robust feature extraction methods, such as those utilizing neural implicit vision-language features, can improve the model's ability to handle variations [23]. Additionally, developing models capable of holistic perception and reasoning across multiple images can further enhance robustness by providing a more comprehensive understanding of complex scenes [26].

Future research should focus on developing adaptive models that dynamically adjust their complexity and processing strategies based on task requirements and input conditions. This adaptability will be key to improving both generalization and robustness, enabling models to operate effectively in a wide range of real-world scenarios. Furthermore, exploring more sophisticated language parsing techniques and integrating additional data modalities can provide richer contextual information, thereby enhancing the model's ability to generalize and remain robust in diverse environments [92]. By addressing these aspects, the field can advance towards more reliable and versatile AI systems capable of performing consistently across various applications and conditions.

## 9 Conclusion

The survey highlights the transformative potential of integrating vision and language models in open vocabulary scene understanding and object detection. These models enable AI systems to interpret complex visual scenes using an unrestricted vocabulary, significantly improving their capacity to generalize to novel categories and diverse contexts. The Open-Vocabulary SAM exemplifies this progress, showcasing substantial advancements in interactive segmentation and recognition, and establishing a solid baseline for future research in vision foundation model integration [15].

Unified frameworks incorporating vision-language tasks have yielded notable enhancements in classification and detection. Models such as V-HOI MLCR and PosSAM improve reasoning and segmentation performance by collaborating with large language models to generate class and instance-aware masks. This integration underscores the potential of vision-language synergy to revolutionize AI applications, facilitating versatile, context-aware interactions across multiple modalities.

Additionally, methodologies like LMSeg and OpenOcc demonstrate significant advancements in open vocabulary semantic segmentation and 3D scene understanding, respectively. By leveraging knowledge graphs and integrating 3D scene reconstruction, these approaches enhance scene comprehension and achieve competitive benchmark performance. The robust results in real-world tasks, including open vocabulary multi-label video classification and sketch segmentation, further illustrate the transformative impact of these technologies.

Future research should focus on integrating detection and perception networks to bolster safety and robustness in the face of various challenges. Advancing these technologies will enable AI systems to develop more adaptable and resilient solutions, paving the way for innovative applications across diverse fields.

## References


[1] Chaoyang Zhu and Long Chen. A survey on open-vocabulary detection and segmentation: Past, present, and future, 2024.

[2] Hang Zhao, Xavier Puig, Bolei Zhou, Sanja Fidler, and Antonio Torralba. Open vocabulary scene parsing, 2017.

[3] Hengcan Shi, Munawar Hayat, and Jianfei Cai. Open-vocabulary object detection via scene graph discovery, 2023.

[4] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection with vision transformers, 2022.

[5] Jianzong Wu, Xiangtai Li, Shilin Xu, Haobo Yuan, Henghui Ding, Yibo Yang, Xia Li, Jiangning Zhang, Yunhai Tong, Xudong Jiang, Bernard Ghanem, and Dacheng Tao. Towards open vocabulary learning: A survey, 2024.

[6] Mohamed Ashraf Abdelsalam, Zhan Shi, Federico Fancellu, Kalliopi Basioti, Dhaivat J. Bhatt, Vladimir Pavlovic, and Afsaneh Fazly. Visual semantic parsing: From images to abstract meaning representation, 2022.

[7] Yilun Chen, Shuai Yang, Haifeng Huang, Tai Wang, Runsen Xu, Ruiyuan Lyu, Dahua Lin, and Jiangmiao Pang. Grounded 3d-llm with referent tokens, 2024.

[8] Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 565–580. Springer, 2020.

[9] Ziqiao Ma, Jiayi Pan, and Joyce Chai. World-to-words: Grounded open vocabulary acquisition through fast mapping in vision-language models, 2023.

[10] Relja Arandjelović, Alex Andonian, Arthur Mensch, Olivier J. Hénaff, Jean-Baptiste Alayrac, and Andrew Zisserman. Three ways to improve feature alignment for open vocabulary detection, 2023.

[11] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions, 2021.

[12] Yanhao Zheng and Kai Liu. Training-free boost for open-vocabulary object detection with confidence aggregation, 2024.

[13] Jincheng Li, Chunyu Xie, Xiaoyu Wu, Bin Wang, and Dawei Leng. What makes good open-vocabulary detector: A disassembling perspective, 2023.

[14] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-vlm: Open-vocabulary object detection upon frozen vision and language models, 2023.

[15] Haobo Yuan, Xiangtai Li, Chong Zhou, Yining Li, Kai Chen, and Chen Change Loy. Open-vocabulary sam: Segment and recognize twenty-thousand classes interactively, 2024.

[16] Taolin Zhang, Sunan He, Dai Tao, Bin Chen, Zhi Wang, and Shu-Tao Xia. Vision-language pre-training with object contrastive learning for 3d scene understanding, 2023.

[17] Muer Tie, Julong Wei, Zhengjun Wang, Ke Wu, Shansuai Yuan, Kaizhao Zhang, Jie Jia, Jieru Zhao, Zhongxue Gan, and Wenchao Ding. O2v-mapping: Online open-vocabulary mapping with neural implicit representation, 2024.

[18] Shuhuai Ren, Lei Li, Xuancheng Ren, Guangxiang Zhao, and Xu Sun. Delving into the openness of clip, 2023.

[19] Prannay Kaul, Weidi Xie, and Andrew Zisserman. Multi-modal classifiers for open-vocabulary object detection, 2023.

[20] Yan Li, Weiwei Guo, Xue Yang, Ning Liao, Dunyun He, Jiaqi Zhou, and Wenxian Yu. Toward open vocabulary aerial object detection with clip-activated student-teacher learning, 2024.

[21] Hengcan Shi, Munawar Hayat, and Jianfei Cai. Unified open-vocabulary dense visual prediction, 2023.

[22] Tiancheng Zhao, Peng Liu, and Kyusong Lee. Omdet: Large-scale vision-language multi-dataset pre-training with multimodal detection network, 2024.

[23] Kenneth Blomqvist, Francesco Milano, Jen Jen Chung, Lionel Ott, and Roland Siegwart. Neural implicit vision-language feature fields, 2023.

[24] Huadong Tang, Youpeng Zhao, Yan Huang, Min Xu, Jun Wang, and Qiang Wu. Lmseg: Unleashing the power of large-scale models for open-vocabulary semantic segmentation, 2024.

[25] Dicong Qiu, Wenzong Ma, Zhenfu Pan, Hui Xiong, and Junwei Liang. Open-vocabulary mobile manipulation in unseen dynamic environments with 3d semantic maps, 2024.

[26] Rujie Wu, Xiaojian Ma, Zhenliang Zhang, Wei Wang, Qing Li, Song-Chun Zhu, and Yizhou Wang. Bongard-openworld: Few-shot reasoning for free-form visual concepts in the real world, 2025.

[27] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[28] Li Yang, Yan Xu, Chunfeng Yuan, Wei Liu, Bing Li, and Weiming Hu. Improving visual grounding with visual-linguistic verification and iterative reasoning, 2022.

[29] Xuelin Zhu, Jian Liu, Dongqi Tang, Jiawei Ge, Weijia Liu, Bo Liu, and Jiuxin Cao. Query-based knowledge sharing for open-vocabulary multi-label classification, 2024.

[30] Jin Yao, Hao Gu, Xuweiyi Chen, Jiayun Wang, and Zezhou Cheng. Open vocabulary monocular 3d object detection, 2024.

[31] Zhenyu Wang, Yali Li, Taichi Liu, Hengshuang Zhao, and Shengjin Wang. Ov-uni3detr: Towards unified open-vocabulary 3d object detection via cycle-modality propagation, 2024.

[32] Ruotian Luo. Goal-driven text descriptions for images, 2021.

[33] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding, 2018.

[34] Yonghui Wang, Wengang Zhou, Hao Feng, and Houqiang Li. Adaptvision: Dynamic input scaling in mllms for versatile scene understanding, 2024.

[35] Brendan Shillingford, Yannis Assael, Matthew W. Hoffman, Thomas Paine, Cían Hughes, Utsav Prabhu, Hank Liao, Hasim Sak, Kanishka Rao, Lorrayne Bennett, Marie Mulville, Ben Coppin, Ben Laurie, Andrew Senior, and Nando de Freitas. Large-scale visual speech recognition, 2018.

[36] Yinjie Lei, Zixuan Wang, Feng Chen, Guoqing Wang, Peng Wang, and Yang Yang. Recent advances in multi-modal 3d scene understanding: A comprehensive survey and evaluation, 2023.

[37] Songyou Peng, Kyle Genova, Chiyu "Max" Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. Openscene: 3d scene understanding with open vocabularies, 2023.

[38] Roei Herzig, Alon Mendelson, Leonid Karlinsky, Assaf Arbelle, Rogerio Feris, Trevor Darrell, and Amir Globerson. Incorporating structured representations into pretrained vision language models using scene graphs, 2023.

[39] Ben Crulis, Barthelemy Serres, Cyril De Runz, and Gilles Venturini. Few-shot target-driven instance detection based on open-vocabulary object detection models, 2024.

[40] Ahmed Bourouis, Judith Ellen Fan, and Yulia Gryaditskaya. Open vocabulary semantic scene sketch understanding, 2024.

[41] Bilal Faye, Binta Sow, Hanane Azzag, and Mustapha Lebbah. A lightweight modular framework for low-cost open-vocabulary object detection training, 2024.

[42] Tao Wang and Nan Li. Learning to detect and segment for open vocabulary object detection, 2024.

[43] Peixi Wu, Bosong Chai, Xuan Nie, Longquan Yan, Zeyu Wang, Qifan Zhou, Boning Wang, Yansong Peng, and Hebei Li. Enhanced object detection: A study on vast vocabulary object detection track for v3det challenge 2024, 2024.

[44] Walid Bousselham, Felix Petersen, Vittorio Ferrari, and Hilde Kuehne. Grounding everything: Emerging localization properties in vision-language transformers, 2023.

[45] Lorenzo Bianchi, Fabio Carrara, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. The devil is in the fine-grained details: Evaluating open-vocabulary object detectors for fine-grained understanding, 2024.

[46] Peixian Chen, Kekai Sheng, Mengdan Zhang, Mingbao Lin, Yunhang Shen, Shaohui Lin, Bo Ren, and Ke Li. Open vocabulary object detection with proposal mining and prediction equalization, 2022.

[47] Yuhang Zang, Wei Li, Jun Han, Kaiyang Zhou, and Chen Change Loy. Contextual object detection with multimodal large language models, 2024.

[48] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018, 2024.

[49] Mohamad Mahmoud Al Rahhal, Yakoub Bazi, Hebah Elgibreen, and Mansour Zuair. Vision-language models for zero-shot classification of remote sensing images. *Applied Sciences*, 13(22):12462, 2023.

[50] Xiaojie Yin, Qilong Wang, Bing Cao, and Qinghua Hu. $s^3$: Synonymous semantic space for improving zero-shot generalization of vision-language models, 2024.

[51] Linfeng He, Yiming Sun, Sihao Wu, Jiaxu Liu, and Xiaowei Huang. Integrating object detection modality into visual language model for enhanced autonomous driving agent, 2024.

[52] Finn Lukas Busch, Timon Homberger, Jesús Ortega-Peimbert, Quantao Yang, and Olov Andersson. One map to find them all: Real-time open-vocabulary mapping for zero-shot multi-object navigation, 2024.

[53] Angelos Zavras, Dimitrios Michail, Begüm Demir, and Ioannis Papoutsis. Mind the modality gap: Towards a remote sensing vision-language model via cross-modal alignment, 2024.

[54] Jialei Chen, Daisuke Deguchi, Chenkai Zhang, Xu Zheng, and Hiroshi Murase. Clip is also a good teacher: A new learning framework for inductive zero-shot semantic segmentation, 2024.

[55] Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. Prismatic vlms: Investigating the design space of visually-conditioned language models, 2024.

[56] Sheng Jin, Xueying Jiang, Jiaxing Huang, Lewei Lu, and Shijian Lu. Llms meet vlms: Boost open vocabulary object detection with fine-grained descriptors, 2024.

[57] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[58] Zuxuan Wu, Zejia Weng, Wujian Peng, Xitong Yang, Ang Li, Larry S. Davis, and Yu-Gang Jiang. Building an open-vocabulary video clip model with better architectures, optimization and data, 2023.

[59] Shaunak Halbe, Junjiao Tian, K J Joseph, James Seale Smith, Katherine Stevo, Vineeth N Balasubramanian, and Zsolt Kira. Grounding descriptions in images informs zero-shot visual recognition, 2024.

[60] Melissa Hall, Laura Gustafson, Aaron Adcock, Ishan Misra, and Candace Ross. Vision-language models performing zero-shot tasks exhibit gender-based disparities, 2023.

[61] Yifan Xu, Mengdan Zhang, Xiaoshan Yang, and Changsheng Xu. Exploring multi-modal contextual knowledge for open-vocabulary object detection, 2023.

[62] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, et al. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*, 2023.

[63] Dimity Miller, Niko Sünderhauf, Alex Kenna, and Keita Mason. Open-set recognition in the age of vision-language models, 2024.

[64] Baoxiong Jia, Yixin Chen, Huangyue Yu, Yan Wang, Xuesong Niu, Tengyu Liu, Qing Li, and Siyuan Huang. Sceneverse: Scaling 3d vision-language learning for grounded scene understanding, 2024.

[65] Junhao Cai, Yisheng He, Weihao Yuan, Siyu Zhu, Zilong Dong, Liefeng Bo, and Qifeng Chen. Ov9d: Open-vocabulary category-level 9d object pose and size estimation, 2024.

[66] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation, 2022.

[67] Zhao Wang, Aoxue Li, Fengwei Zhou, Zhenguo Li, and Qi Dou. Open-vocabulary object detection with meta prompt representation and instance contrastive optimization, 2024.

[68] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-vlm: Open-vocabulary object detection upon frozen vision and language models. *arXiv preprint arXiv:2209.15639*, 2022.

[69] Penghui Du, Yu Wang, Yifan Sun, Luting Wang, Yue Liao, Gang Zhang, Errui Ding, Yan Wang, Jingdong Wang, and Si Liu. Lami-detr: Open-vocabulary detection with language model instruction, 2024.

[70] Jiaming Li, Jiacheng Zhang, Jichang Li, Ge Li, Si Liu, Liang Lin, and Guanbin Li. Learning background prompts to discover implicit knowledge for open vocabulary object detection, 2024.

[71] Sheng Liu, Kevin Lin, Lijuan Wang, Junsong Yuan, and Zicheng Liu. Ovis: Open-vocabulary visual instance search via visual-semantic aligned representation learning, 2021.

[72] Ayça Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. Openmask3d: Open-vocabulary 3d instance segmentation, 2023.

[73] Zhening Huang, Xiaoyang Wu, Xi Chen, Hengshuang Zhao, Lei Zhu, and Joan Lasenby. Openins3d: Snap and lookup for 3d open-vocabulary instance segmentation, 2024.

[74] Guofeng Mei, Luigi Riz, Yiming Wang, and Fabio Poiesi. Vocabulary-free 3d instance segmentation with vision and language assistant, 2024.

[75] Hanchen Tai, Qingdong He, Jiangning Zhang, Yijie Qian, Zhenyu Zhang, Xiaobin Hu, Xiangtai Li, Yabiao Wang, and Yong Liu. Open-vocabulary sam3d: Towards training-free open-vocabulary 3d scene understanding, 2024.

[76] Qingdong He, Jinlong Peng, Zhengkai Jiang, Kai Wu, Xiaozhong Ji, Jiangning Zhang, Yabiao Wang, Chengjie Wang, Mingang Chen, and Yunsheng Wu. Unim-ov3d: Uni-modality open-vocabulary 3d scene understanding with fine-grained feature representation, 2024.

[77] Tri Ton, Ji Woo Hong, SooHwan Eom, Jun Yeop Shim, Junyeong Kim, and Chang D. Yoo. Zero-shot dual-path integration framework for open-vocabulary 3d instance segmentation, 2024.

[78] Shivam Chandhok. Scenegpt: A language model for 3d scene understanding, 2024.

[79] Zizhao Li, Zhengkang Xiang, Joseph West, and Kourosh Khoshelham. From open vocabulary to open world: Teaching vision language models to detect novel objects, 2024.

[80] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. Pla: Language-driven open-vocabulary 3d scene understanding, 2023.

[81] Jianyu Zhang, Li Zhang, and Shijian Li. Incorporating feature pyramid tokenization and open vocabulary semantic segmentation, 2024.

[82] Jinwoo Ahn, Hyeokjoon Kwon, and Hwiyeon Yoo. Fine-grained open-vocabulary object recognition via user-guided segmentation, 2024.

[83] Youwei Pang, Xiaoqi Zhao, Jiaming Zuo, Lihe Zhang, and Huchuan Lu. Open-vocabulary camouflaged object segmentation, 2024.

[84] AJ Piergiovanni, Weicheng Kuo, and Anelia Angelova. Pre-training image-language transformers for open-vocabulary tasks, 2022.

[85] Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Xiaoshui Huang, Zhiyong Wang, Lu Sheng, Lei Bai, et al. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *Advances in Neural Information Processing Systems*, 36:26650–26685, 2023.

[86] Kunhao Liu, Fangneng Zhan, Jiahui Zhang, Muyu Xu, Yingchen Yu, Abdulmotaleb El Saddik, Christian Theobalt, Eric Xing, and Shijian Lu. Weakly supervised 3d open-vocabulary segmentation, 2024.

[87] Gonca Yilmaz, Songyou Peng, Marc Pollefeys, Francis Engelmann, and Hermann Blum. Opendas: Open-vocabulary domain adaptation for 2d and 3d segmentation, 2024.

[88] Jun Chen, Deyao Zhu, Guocheng Qian, Bernard Ghanem, Zhicheng Yan, Chenchen Zhu, Fanyi Xiao, Mohamed Elhoseiny, and Sean Chang Culatana. Exploring open-vocabulary semantic segmentation without human labels, 2023.

[89] Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, Quanquan Gu, James Zou, Kai-Wei Chang, and Wei Wang. Enhancing large vision language models with self-training on image comprehension. *arXiv preprint arXiv:2405.19716*, 2024.

[90] Yunheng Li, Yuxuan Li, Quansheng Zeng, Wenhai Wang, Qibin Hou, and Ming-Ming Cheng. Densevlm: A retrieval and decoupled alignment framework for open-vocabulary dense prediction, 2024.

[91] Chaofan Ma, Yuhuan Yang, Yanfeng Wang, Ya Zhang, and Weidi Xie. Open-vocabulary semantic segmentation with frozen vision-language models, 2022.

[92] Junbo Zhang, Guofan Fan, Guanghan Wang, Zhengyuan Su, Kaisheng Ma, and Li Yi. Language-assisted 3d feature learning for semantic scene understanding, 2022.

[93] Rohit Gupta, Mamshad Nayeem Rizve, Jayakrishnan Unnikrishnan, Ashish Tawari, Son Tran, Mubarak Shah, Benjamin Yao, and Trishul Chilimbi. Open vocabulary multi-label video classification, 2024.

[94] Vibashan VS, Shubhankar Borse, Hyojin Park, Debasmit Das, Vishal Patel, Munawar Hayat, and Fatih Porikli. Possam: Panoptic open-vocabulary segment anything, 2024.

[95] Yingdong Hu, Fanqi Lin, Tong Zhang, Li Yi, and Yang Gao. Look before you leap: Unveiling the power of gpt-4v in robotic vision-language planning. *arXiv preprint arXiv:2311.17842*, 2023.

[96] Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. *Advances in neural information processing systems*, 35:35087–35102, 2022.

[97] Weicheng Kuo, AJ Piergiovanni, Dahun Kim, Xiyang Luo, Ben Caine, Wei Li, Abhijit Ogale, Luowei Zhou, Andrew Dai, Zhifeng Chen, Claire Cui, and Anelia Angelova. Mammut: A simple architecture for joint learning for multimodal tasks, 2023.

[98] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*, 2023.

[99] Xianzheng Ma, Yash Bhalgat, Brandon Smart, Shuai Chen, Xinghui Li, Jian Ding, Jindong Gu, Dave Zhenyu Chen, Songyou Peng, Jia-Wang Bian, Philip H Torr, Marc Pollefeys, Matthias Nießner, Ian D Reid, Angel X. Chang, Iro Laina, and Victor Adrian Prisacariu. When llms step into the 3d world: A survey and meta-analysis of 3d tasks via multi-modal large language models, 2024.

[100] Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Haodong Duan, Songyang Zhang, Shuangrui Ding, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023.

[101] Zeju Li, Chao Zhang, Xiaoyan Wang, Ruilong Ren, Yifan Xu, Ruifei Ma, and Xiangde Liu. 3dmit: 3d multi-modal instruction tuning for scene understanding, 2024.

**Disclaimer:**

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.