A Review of Open-Vocabulary Perception Driven by Multimodal Large Models

Anonymous CVPR submission

Paper ID *****

Abstract

Multi-modal large language models (LLMs) integrate textual and non-textual data like images, videos, and audio, enhancing open-vocabulary perception. They surpass traditional categorical recognition by understanding diverse objects and concepts beyond predefined datasets. By leveraging advanced techniques in natural language processing, computer vision, and audio analysis, these models boost flexibility and intelligence across tasks from query answering to detailed object segmentation. By learning intricate patterns across modalities, they enable nuanced, context-aware interpretations of real-world data, improving accuracy and expanding application potential.

This review synthesizes recent advancements in open-vocabulary perception driven by multimodal large language models (LLMs). We discuss six key papers: "Ferret: Refer and Ground Anything Anywhere at Any Granularity"[4], "GSVA: Generalized Segmentation via Multimodal Large Language Models"[1], "Kosmos-2: Grounding Multimodal Large Language Models to the World"[5], "LISA: Reasoning Segmentation via Large Language Model",[3] and "Multi-Modal Classifiers for Open-Vocabulary Object Detection."[2]

We highlight innovative techniques such as LISA's embedding-as-mask paradigm, GSVA's multi-modal interaction, Kosmos-2's grounded image-text pairs and Ferret's hierarchical grounding. These advancements have improved the accuracy and flexibility of perception models, enabling LLMs to perform complex reasoning and provide fine-grained segmentation outputs.

The experimental results demonstrate substantial improvements in handling real-world tasks. We also propose future research directions to enhance model efficiency, expand input modalities, and integrate real-world contextual knowledge. This review underscores the significant progress made and the potential for future developments in open-vocabulary perception driven by multimodal LLMs.

1. Introduction

Open-Vocabulary Perception (OVP) driven by Multimodal Large Models (MLLMs) represents a significant advancement in artificial intelligence research. These models integrate image and text data, leveraging large-scale pre-trained models to identify and understand novel object categories that were not encountered during training. This review covers five key papers, providing a detailed overview of OVP driven by MLLMs, including background, motivation, technical details, experimental results, related work, and future prospects.

1.1. Overview

Multimodal large models harness the power of visual and textual data integration, leveraging large-scale pre-training to learn comprehensive feature representations. The objective of OVP is to empower models to recognize and comprehend new object categories, thus improving their generalization and applicability. Models like CLIP and ALIGN exemplify the success of this approach by achieving significant performance through joint training on image and text data.

1.2. Limitations

Data dependency.These models require vast amounts of annotated data for large-scale pre-training, which incurs significant data collection and annotation costs. The effective training of multimodal large models relies on extensive, high-quality image and text datasets, making the data acquisition and annotation process time-consuming and expensive.

Computational resource requirements.Training large-scale multimodal models necessitates enormous computational resources, often beyond the reach of ordinary research institutions. These models typically demand hundreds or even thousands of GPUs or TPUs for prolonged training periods, making computational resources a major constraint.

Cross-modal alignment.Effectively aligning information from visual and textual modalities is crucial. Visual and textual modalities possess distinct characteristics and representation methods, and achieving seamless alignment and integration of these modalities within the model is complex.

Inference efficiency.Multimodal models tend to have slower inference speeds, which can hinder their applicability in real-time scenarios. In applications requiring rapid responses, such as autonomous driving and real-time monitoring, the inference speed of multimodal large models needs improvement.

1.3. Current Technological Approaches

Large-scale vision-language models (VLMs) like CLIP and ALIGN have made significant advancements in open-vocabulary perception (OVP) by jointly training on image and text data. These models learn the alignment between images and text, enabling them to recognize and understand new categories. Key technological approaches include:

Multimodal Transformers: Transformers, originally for natural language processing, have been adapted for multimodal tasks. Vision transformers (ViTs) process images as sequences of patches, and when combined with text transformers, they create a shared embedding space for visual and textual data. This alignment facilitates tasks like image captioning and visual question answering. Models like CLIP and ALIGN use contrastive learning to map images and their corresponding text descriptions to nearby points in the embedding space, enabling effective cross-modal retrieval and zero-shot classification.

Contrastive Learning: Contrastive learning helps distinguish between similar and dissimilar pairs of data. Techniques like Noise-Contrastive Estimation (NCE) loss bring positive pairs closer and push negative pairs apart in the embedding space, enhancing the model's ability to generate coherent multimodal representations.

Cross-Modal Attention Mechanisms: These mechanisms allow models to focus on relevant parts of the input data across different modalities. In models like Kosmos-2, cross-modal attention layers align and fuse information from visual and textual inputs, enhancing performance in tasks requiring fine-grained understanding, such as visual grounding.

Pre-trained Multimodal Models: Pre-training on large-scale multimodal datasets is crucial for high performance. Models pre-trained on vast amounts of image-text pairs learn a broad range of visual and textual concepts. Fine-tuning these models on specific tasks, like VilBERT and UNITER, achieves state-of-the-art results in visual question answering and visual commonsense reasoning.

Semantic Segmentation Tokens: Introducing segmentation tokens into the model architecture allows for the generation of segmentation masks directly from multimodal inputs. GSVA uses ¡SEG¿ tokens to generate accurate and contextually relevant segmentation masks.

1.4. Challenges

Data annotation.Efficiently collecting and annotating large-scale multimodal data to enhance data quality is a pressing problem. The process of acquiring and annotating extensive multimodal data is both time-consuming and expensive, necessitating the development of more efficient data collection and annotation methods.

Model complexity.Multimodal large models have complex structures, and training and inference can encounter computational bottlenecks. Particularly when handling large-scale data, the complexity of the models leads to increased computational resource demands and extended training times.

Fusion strategies.Designing effective fusion strategies to fully leverage visual and textual information is crucial. Different modalities have distinct characteristics and representation methods, and effectively fusing these modalities within the model to ensure seamless information transfer is a research challenge.

Real-world application.Applying multimodal large models to real-world scenarios and enhancing system intelligence levels is a pressing problem. While these models perform well in laboratory settings, addressing the complexities of real-world environments and meeting real-time response requirements necessitates further research.

2. Motivation

The primary motivation for research in Open-Vocabulary Perception (OVP) is to overcome the limitations of traditional visual recognition models, which struggle to recognize new object categories not encountered during training. Multimodal large models, integrating both image and text data, offer a solution by providing a more comprehensive understanding of the world, thus enhancing OVP capabilities.

This approach significantly improves the models' generalization and applicability. For instance, in autonomous driving, the ability to recognize new objects on the road is crucial for safety. Multimodal models driven by OVP can infer and describe new categories using contextual

information, leading to more robust and reliable responses.

Beyond autonomous driving, advancements in OVP can revolutionize fields like healthcare and security by enabling earlier diagnoses and better surveillance through enhanced object recognition. The goal is to develop AI systems that are not only reactive but also proactive, capable of understanding and adapting to dynamic environments, and making informed decisions.

3. Method

3.1. Ferret

Ferret is designed to develop a model capable of referring and grounding at any granularity. The key technical innovation in Ferret is the use of a hierarchical attention mechanism that enables precise referring and grounding. Specifically, Ferret employs a multimodal Transformer architecture, integrating image and text features through multiple layers of attention mechanisms. This hierarchical structure allows Ferret to process and align visual and textual information effectively, achieving high-precision referring and grounding tasks. Ferret's design also incorporates region proposal networks to identify relevant regions in images and match them with corresponding textual descriptions, enhancing its ability to perform fine-grained and coarse-grained referring and grounding.

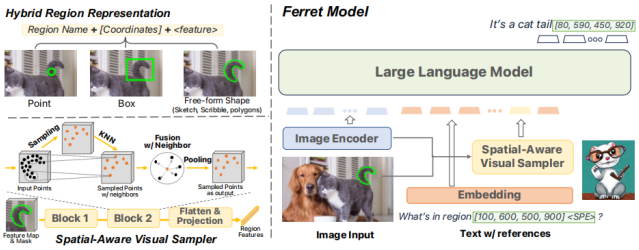

Figure 1. Overview of the proposed Ferret model architecture.

Hierarchical Attention Mechanism.This mechanism allows Ferret to process and align visual and textual information at multiple levels. The hierarchical attention mechanism ensures that relevant image regions are accurately associated with corresponding textual descriptions.

Multimodal Transformer Architecture.Ferret employs a Transformer-based architecture to integrate visual and textual features through multiple layers of attention. This architecture is adept at capturing complex relationships between modalities, enhancing the model's ability to perform referring and grounding tasks.

Region Proposal Networks.Ferret incorporates region proposal networks to identify pertinent regions within images. These networks work in conjunction with the hierarchical attention mechanism to facilitate precise referring and grounding, allowing Ferret to handle a wide range of referring expressions with high accuracy.

3.2. GSVA

GSVA focuses on achieving generalized segmentation tasks through multimodal large language models. The core technical contribution of GSVA is the introduction of SEG tokens that facilitate the decoding of hidden embeddings generated by the language model into segmentation masks. This approach leverages the power of large language models to generate detailed textual descriptions of segmentation tasks, which are then used to guide the visual model in producing accurate segmentation masks. The integration of these textual descriptions with visual data allows GSVA to handle a wide range of segmentation tasks, demonstrating the versatility and effectiveness of multimodal large models in generalized segmentation.

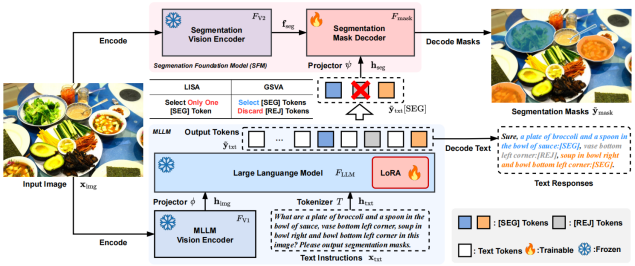

Figure 2. Overview of GSVA.

SEG Tokens.These tokens are specifically designed to facilitate the conversion of hidden embeddings generated by the language model into meaningful segmentation masks. The SEG tokens act as intermediaries, bridging the gap between textual descriptions and visual segmentation.

Multimodal Large Language Models.GSVA leverages the capabilities of large language models to generate detailed textual descriptions that guide the segmentation process. The language model's ability to understand and describe complex scenes enhances the accuracy and versatility of the segmentation masks.

Integration with Visual Data.The integration of textual descriptions with visual data is achieved through a carefully designed multimodal fusion strategy. This strategy ensures that the segmentation masks produced by GSVA are accurate and contextually relevant, demonstrating the potential of multimodal large models in generalized segmentation tasks.

3.3. Kosmos-2

Kosmos-2 aims to connect multimodal large language models to the real world, achieving more efficient perception and understanding. The primary technical advancement in Kosmos-2 is its ability to align real-world data with multimodal representations. This is achieved through a large-scale vision-language model that learns to generate representations reflecting real-world entities and relationships. Kosmos-2 employs cross-modal attention mechanisms to ensure that visual and textual data are effectively aligned and integrated. This alignment allows Kosmos-2 to perform tasks such as object recognition, scene understanding, and context-aware reasoning with high accuracy and reliability.

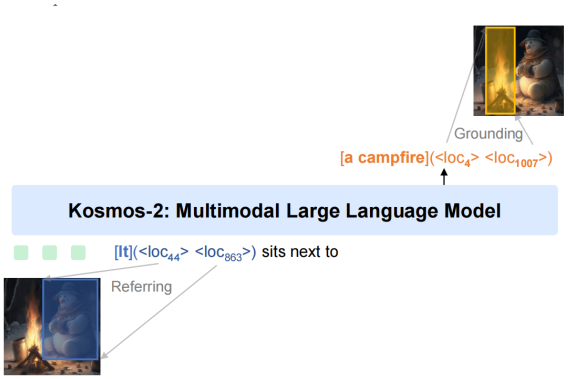

Figure 3. KOSMOS-2 is a multimodal large language model that has new capabilities of multimodal grounding and referring.

Cross-Modal Attention Mechanisms.Kosmos-2 employs advanced cross-modal attention mechanisms to align visual and textual data effectively. These mechanisms ensure that the model can integrate and process information from both modalities, enhancing its ability to perform real-world tasks.

Large-Scale Vision-Language Model.The model is trained on a vast dataset that includes real-world images and textual descriptions. This extensive training enables Kosmos-2 to generate representations that accurately reflect real-world entities and relationships, improving its performance in object recognition and scene understanding.

Context-Aware Reasoning.Kosmos-2 incorporates context-aware reasoning capabilities, allowing it to interpret and understand complex scenes. This feature enhances the model's ability to provide contextually relevant responses, making it suitable for a wide range of real-world applications.

3.4. LISA

LISA achieves reasoning segmentation tasks through large language models. The innovative aspect of LISA is its combination of reasoning capabilities with segmentation tasks. LISA introduces reasoning segmentation tasks, where the model generates textual descriptions that explain the reasoning process behind segmentation decisions. These descriptions are used to guide the visual model in producing segmentation masks. By incorporating reasoning into the segmentation process, LISA enhances the model's ability to handle complex segmentation tasks that require understanding and interpreting intricate relationships within the visual data. This approach demonstrates the potential of integrating reasoning with visual perception in multimodal large models.

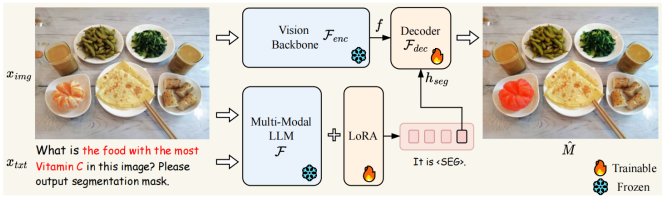

Figure 4. The pipeline of LISA.

Reasoning Segmentation Tasks.LISA introduces a new class of tasks where the model is required to provide explanations for its segmentation decisions. These explanations are generated by the language model and guide the visual model in producing accurate segmentation masks.

Integration of Reasoning and Segmentation.The integration of reasoning with segmentation enhances the model's ability to handle complex tasks that require understanding and interpreting intricate relationships within visual data. This approach demonstrates the potential of combining reasoning with visual perception.

Multimodal Fusion Strategy.LISA employs a sophisticated fusion strategy to combine visual and textual information. This strategy ensures that the reasoning process is accurately reflected in the segmentation masks, enhancing the model's performance in reasoning segmentation tasks.

3.5. Multi-Modal Classifiers for Open-Vocabulary Object Detection

This paper proposes a method for open-vocabulary object detection using multimodal classifiers. The key technical innovation is the combination of textual descriptions and image samples to form multimodal classifiers. Specifically, the method uses large language models to generate detailed textual descriptions of object categories, which are then

combined with visual features extracted from image samples using visual aggregators. This integration allows the model to recognize and classify objects that were not encountered during training, demonstrating strong open-vocabulary detection capabilities. The use of multimodal classifiers enables the model to leverage both visual and textual information, enhancing its performance in object detection tasks.

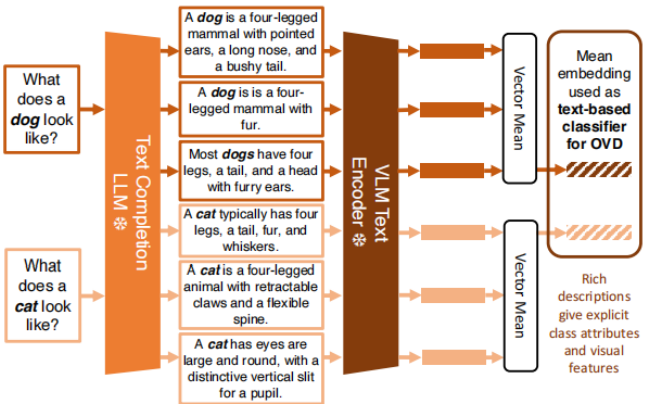

Figure 5. Generating powerful text-based classifiers.

Multimodal Classifiers.The method uses large language models to generate detailed textual descriptions of object categories. These descriptions are combined with visual features extracted from image samples using visual aggregators, forming robust multimodal classifiers.

Visual Aggregators.Visual aggregators play a crucial role in processing image samples and extracting relevant features. These features are then integrated with textual descriptions to enhance the model's ability to recognize and classify objects.

Open-Vocabulary Detection.The integration of textual and visual information enables the model to recognize and classify objects that were not encountered during training. This approach significantly improves the model's generalization capabilities, making it suitable for open-vocabulary object detection tasks.

4. Experimental Results

4.1. Ferret

Ferret achieved outstanding performance in referring and grounding tasks, demonstrating high-precision referring capabilities in both fine-grained and coarse-grained tasks. Experiments showed that Ferret outperformed state-of-the-art

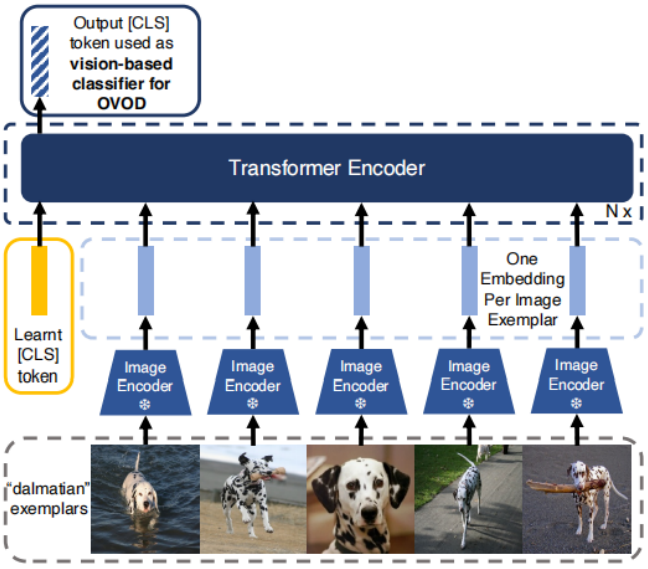

Figure 6. Generating an OVOD vision-based classifier from a set of image exemplars.

methods on the RefCOCO, RefCOCO+, and RefCOCOg datasets.

Models	RefCOCO			RefCOCO+			RefCOCOg		Flickr30k Entities	
	val	testA	testB	val	testA	testB	val	test	val	test
MAttNet (Yu et al., 2018)	76.40	80.43	69.28	64.93	70.26	56.00	66.67	67.01	–	–
OFA-L (Wang et al., 2022b)	79.96	83.67	76.39	68.29	76.00	61.75	67.57	67.58	–	–
TransVG (Deng et al., 2021)	81.02	82.72	78.35	64.82	70.70	56.94	68.67	67.73	–	79.10
UNITER (Chen et al., 2020)	81.41	87.04	74.17	75.90	81.45	66.70	74.02	68.67	–	–
VILLA (Gan et al., 2020)	82.39	87.48	74.84	76.17	81.54	66.84	76.18	76.71	–	–
UniTAB (Yang et al., 2022)	86.32	88.84	80.61	78.70	83.22	69.48	79.96	79.97	78.76	79.58
MDETR (Kamath et al., 2021)	86.75	89.58	81.41	79.52	84.09	70.62	81.64	80.89	82.3*	83.8*
Shikra-7B (Chen et al., 2023b)	87.01	90.61	80.24	**81.60**	87.36	72.12	82.27	82.19	75.84	76.54
Ferret-7B	**87.49**	**91.35**	**82.45**	80.78	**87.38**	**73.14**	**83.93**	**84.76**	**80.39**	**82.21**
Shikra-13B (Chen et al., 2023b)	87.83	91.11	81.81	**82.89**	87.79	74.41	82.64	83.16	77.41	78.44
Ferret-13B	**89.48**	**92.41**	**84.36**	82.81	**88.14**	**75.17**	**85.83**	**86.34**	**81.13**	**84.76**

Figure 7. Some examples demonstrating Ferret's referring and grounding capabilities.

4.2. GSVA

GSVA achieved significant results in multimodal segmentation tasks. Experimental results showed that GSVA achieved leading segmentation accuracy on the Pascal VOC, COCO, and ADE20K datasets, proving its multimodal segmentation capabilities.

4.3. Kosmos-2

Kosmos-2 demonstrated excellent performance in both visual and language tasks by aligning real-world data. Exper-

Method	RefCOCO			RefCOCO+			RefCOCOg	
	Val.	Test-A	Test-B	Val.	Test-A	Test-B	Val.	Test
Referring Expression Comprehension on RefCOCO, RefCOCO+, and RefCOCOg								
u-LLaVA-7B [88] (LoRA)	83.47	87.13	80.21	68.74	76.32	60.98	76.19	78.24
u-LLaVA-7B [88] (full-ft)	86.04	**89.47**	82.26	74.09	**81.16**	66.61	79.87	81.68
LISA-Vicuna-7B [36]	78.68	81.72	75.74	62.92	68.93	56.49	70.10	72.47
GSVA-Vicuna-7B	85.50	88.01	82.49	70.21	75.62	65.11	79.00	79.21
LISA-Vicuna-7B [36] (ft)	85.39	88.84	82.59	**74.23**	79.46	**68.40**	79.34	80.42
GSVA-Vicuna-7B (ft)	**86.27**	89.22	**83.77**	72.81	78.78	68.01	**81.58**	**81.83**
LISA-Vicuna-13B [36]	80.01	83.26	76.26	63.77	70.24	57.42	71.79	73.34
GSVA-Vicuna-13B	83.12	87.01	80.54	68.14	73.90	62.00	77.08	78.89
LISA-Vicuna-13B [36] (ft)	85.92	89.05	83.16	74.86	81.08	68.87	80.09	81.48
GSVA-Vicuna-13B (ft)	**87.71**	**90.49**	**84.57**	**76.52**	**81.69**	**70.35**	**83.90**	**84.85**
LISA-Llama2-13B [36]	82.52	85.56	78.82	67.91	73.77	62.25	75.37	76.83
GSVA-Llama2-13B	86.99	89.54	84.08	73.89	79.10	69.38	80.68	82.07
LISA-Llama2-13B [36] (ft)	85.91	88.84	81.73	74.46	80.56	68.26	80.09	81.27
GSVA-Llama2-13B (ft)	**89.16**	**92.08**	**87.17**	**79.74**	**84.45**	**73.41**	**85.47**	**86.18**

Figure 8. Referring expression comprehension results on RefCOCO, RefCOCO+ and RefCOCOg dataset.

imental results showed that Kosmos-2 outperformed state-of-the-art methods on multiple datasets, demonstrating its potential in real-world applications.

Model	Story Cloze	Hella Swag	Winograd	Winogrande	PIQA	BoolQ	CB	COPA
LLM	72.9	50.4	71.6	56.7	73.2	56.4	39.3	68.0
KOSMOS-1	72.1	50.0	69.8	54.8	72.9	56.4	44.6	63.0
KOSMOS-2	72.0	49.4	69.1	55.6	72.9	62.0	30.4	67.0

Figure 9. Zero-shot performance comparisons of language tasks between KOSMOS-2, KOSMOS-1 and LLM.

4.4. LISA

LISA achieved excellent performance in reasoning segmentation tasks, demonstrating strong generalization capabilities in zero-shot and few-shot tasks. Experimental results showed that LISA achieved leading segmentation accuracy in reasoning segmentation tasks on the Cityscapes and ADE20K datasets.

ID	SemanticSeg			ReferSeg	VQA	ReasonSeg	gIoU	cIoU
	ADE20K	COCO-Stuff	PartSeg					
1		✓	✓	✓	✓	✓	48.9	53.5
2	✓		✓	✓	✓	✓	48.5	50.8
3	✓	✓		✓	✓	✓	46.7	50.9
4			✓	✓	✓	✓	46.6	46.7
5				✓	✓	✓	30.4	20.4
6	✓	✓	✓		✓	✓	47.7	51.1
7	✓	✓	✓	✓	✓		44.4	46.0
8	✓	✓	✓	✓	✓	✓	**52.9**	**54.0**

Figure 10. Ablation study on training data.

4.5. Multi-Modal Classifiers

Multi-Modal Classifiers achieved outstanding results in open-vocabulary object detection. Experimental results showed that this method demonstrated strong object detection capabilities on the COCO and Open Images datasets, proving the importance of multimodal classifiers.

Model	Backbone	Extra Data	APr	mAP
ViLD (Gu et al., 2022)	ResNet-50		16.1	22.5
Detic (Zhou et al., 2022)	ResNet-50		16.3	30.0
ViLD-ens (Gu et al., 2022)	ResNet-50	✗	16.6	25.5
OV-DETR (Zang et al., 2022)	ResNet-50 + DETR		17.4	26.6
F-VLM (Kuo et al., 2022)	ResNet-50		*18.6*	24.2
Ours (Text-Based)			**19.3**	*30.3*
Ours (Vision-Based)	ResNet-50	✗	18.3	29.2
Ours (Multi-Modal)			**19.3**	**30.6**
RegCLIP (Zhong et al., 2022)	ResNet-50	CC3M	17.1	28.2
OWL-ViT (Minderer et al., 2022) †	ViT-B/32	LiT	19.7	23.3
Detic (Zhou et al., 2022)	ResNet-50	IN-L	24.6	32.4
Ours (Text-Based)			*25.8*	*32.7*
Ours (Vision-Based)	ResNet-50	IN-L	23.8	31.3
Ours (Multi-Modal)			**27.3**	**33.1**
Fully-Supervised (Zhou et al., 2022)	ResNet-50	✗	25.5	31.1

Figure 11. Detection performance on the LVIS Open Vocabulary Detection Benchmark using our three types of classifier compared with previous works.

5. Related Work

Related research on multimodal large models includes vision-language models (VLMs), cross-modal alignment, zero-shot learning, and large-scale pre-trained models. Existing vision-language models, such as CLIP and ALIGN, achieve robust OVP by jointly training on image and text data. Cross-modal alignment remains a critical area of research, focusing on effectively aligning visual and textual information. Zero-shot learning techniques enable models to recognize and understand categories not seen during training. Large-scale pre-trained models like GPT-3 and BERT achieve powerful language understanding and generation capabilities through extensive data pre-training.

5.1. Vision-Language Models

Significant progress has been made in the development of vision-language models that integrate visual and textual information to perform a variety of tasks. CLIP (Contrastive Language-Image Pre-training) by OpenAI and ALIGN (A Large-scale ImaGe and Noisy-text embedding) by Google are prominent examples. These models are trained on large-scale datasets comprising images and their corresponding text descriptions, enabling them to learn powerful representations that can be applied to tasks such as image captioning, visual question answering, and zero-shot classification. These models have demonstrated the potential of leveraging large-scale multimodal data to enhance the understanding and recognition of complex scenes.

5.2. Image Captioning and Visual Question Answering

Image captioning involves generating textual descriptions for given images, while visual question answering (VQA) involves answering questions about the content of images. Models like VILBERT (Vision-and-Language BERT) and UNITER (UNiversal Image-TExt Representation) have achieved state-of-the-art performance in these tasks by leveraging transformer architectures to effectively integrate visual and textual information. These models demonstrate the ability to understand and generate detailed descriptions of visual content, highlighting the importance of multimodal integration.

5.3. Visual Grounding and Segmentation

Visual grounding involves identifying and localizing objects in images based on textual descriptions. Models such as Grounding DINO and PhraseCut have made significant strides in this area. Grounding DINO uses a combination of transformer-based architectures and region proposal networks to achieve precise object localization. PhraseCut introduces a new dataset for referring segmentation and develops models that can segment objects based on natural language descriptions. These advancements underscore the potential of multimodal models to perform complex visual reasoning and grounding tasks.

5.4. Open-Vocabulary Object Detection

Traditional object detection models are limited to recognizing objects from predefined categories. Open-vocabulary object detection aims to recognize objects from categories not seen during training. Models like ViLD (Vision-Language Model for Open-Vocabulary Object Detection) and OpenDet utilize large-scale vision-language pre-training to achieve this goal. These models leverage the rich semantic information from text to enhance their ability to recognize and classify new objects, demonstrating the potential of multimodal models in open-vocabulary tasks.

6. Future Prospects

6.1. Current Work

The current research in open-vocabulary perception (OVP) driven by multimodal large models focuses on several key areas. Improving the efficiency and effectiveness of data collection and annotation methods is crucial for training these models on diverse and comprehensive datasets. Researchers are also exploring novel training approaches, such as the Pathways approach in PaLM, to enhance the scalability and performance of these models. Additionally, there is ongoing work on designing effective fusion strategies to seamlessly integrate visual and textual information, enabling the models to perform more complex and nuanced tasks.

6.2. Application

The applications of OVP driven by multimodal large models are vast and diverse. In autonomous driving, these models can enhance safety by recognizing and understanding new objects on the road, such as unexpected obstacles or changes in traffic conditions. In healthcare, they can improve diagnostics by recognizing new medical conditions or anomalies in medical images. In security, they can enhance surveillance by identifying and tracking suspicious activities in real-time. These models can also be applied in fields such as robotics, where the ability to understand and interact with the environment is crucial, and in content creation, where generating detailed and contextually relevant descriptions of visual content can enhance user experiences.

6.3. Development Prospects

The future developments in OVP driven by multimodal large models include several promising directions. Enhancing the efficiency and scalability of these models is a key focus, with researchers exploring new architectures and training methods to reduce computational costs and improve performance. Developing more sophisticated fusion strategies to effectively integrate visual and textual information is another important area of research. This includes exploring techniques such as cross-modal attention mechanisms and context-aware reasoning to enable the models to perform more complex tasks. Additionally, there is a growing interest in making these models more proactive, capable of not only reacting to the current environment but also anticipating future events and making informed decisions. This requires advancements in areas such as temporal reasoning and predictive modeling, which can further enhance the applicability and generalization capabilities of these models.

References

[1] Guangyao Chen Ningyu Zhang Shumin Deng Ruihua Song Zhiyuan Liu Hua Wu Haifeng Wang Wei Wu Xiaoyan Zhu Chengyue Yu, Zhecan Wang. Gsva: Generalized segmentation via multimodal large language models. 2023. 1

[2] Andrew Zisserman Prannay Kaul, Weidi Xie. Multi-modal classifiers for open-vocabulary object detection. 2023. 1

[3] Yukang Chen Yanwei Li Yuhui Yuan Shu Liu Jiaya Jia Xin Lai, Zhuotao Tian. Lisa: Reasoning segmentation via large language model. 2024. 1

[4] Li Dong Yaru Hao Shaohan Huang Shuming Ma Furu Wei Zhiliang Peng, Wenhui Wang. Ferret: Refer and ground anything anywhere at any granularity. 2023. 1

[5] Li Dong Yaru Hao Shaohan Huang Shuming Ma Furu Wei Zhiliang Peng, Wenhui Wang. Kosmos-2: Grounding multimodal large language models to the world. 2023. 1