X-SAM: From Segment Anything to Any Segmentation

Hao Wang[1,2], Limeng Qiao[3], Zequn Jie[3], Zhijian Huang[1], Chengjian Feng[3], Qingfang Zheng[2], Lin Ma[3], Xiangyuan Lan[2 *], Xiaodan Liang[1 *]

[1]Sun Yat-sen University
[2]Peng Cheng Laboratory
[3]Meituan Inc.
{wanghao9610, xdliang328}@gmail.com, lanxy@pcl.ac.cn

Abstract

Large Language Models (LLMs) demonstrate strong capabilities in broad knowledge representation, yet they are inherently deficient in pixel-level perceptual understanding. Although the Segment Anything Model (SAM) represents a significant advancement in visual-prompt-driven image segmentation, it exhibits notable limitations in multi-mask prediction and category-specific segmentation tasks, and it cannot integrate all segmentation tasks within a unified model architecture. To address these limitations, we present X-SAM, a streamlined Multimodal Large Language Model (MLLM) framework that extends the segmentation paradigm from *segment anything* to *any segmentation*. Specifically, we introduce a novel unified framework that enables more advanced pixel-level perceptual comprehension for MLLMs. Furthermore, we propose a new segmentation task, termed Visual GrounDed (VGD) segmentation, which segments all instance objects with interactive visual prompts and empowers MLLMs with visual grounded, pixel-wise interpretative capabilities. To enable effective training on diverse data sources, we present a unified training strategy that supports co-training across multiple datasets. Experimental results demonstrate that X-SAM achieves state-of-the-art performance on a wide range of image segmentation benchmarks, highlighting its efficiency for multimodal, pixel-level visual understanding.

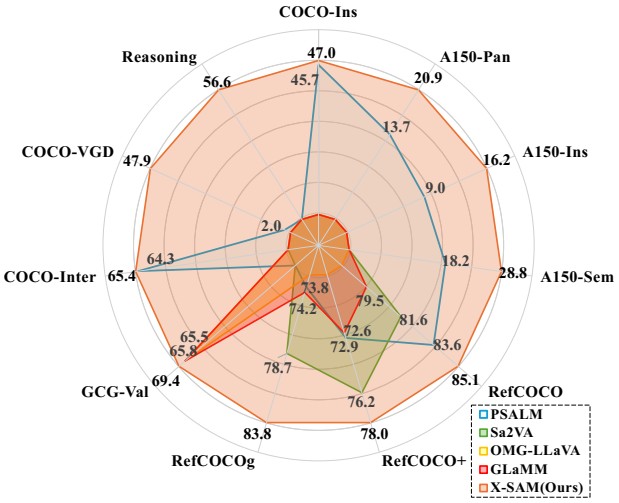


Figure 1: Illustration of Performance of X-SAM on Image Segmentation Benchmarks. X-SAM consistently surpasses existing Multimodal Large Language Models (MLLMs) across all evaluated segmentation benchmarks.

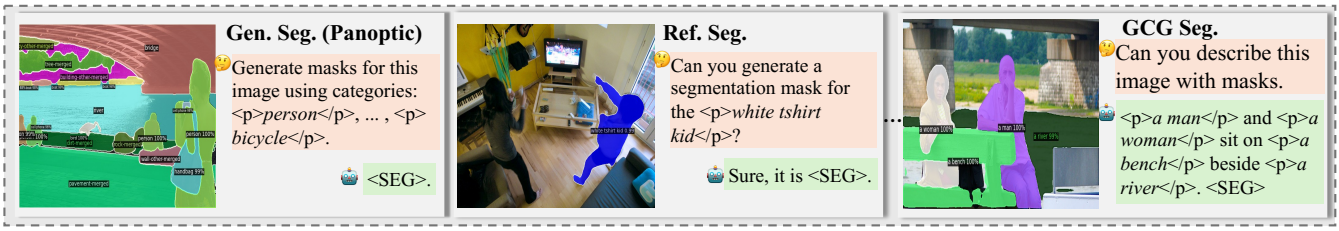
Introduction

Multi-modal Large Language Models (MLLMs) have exhibited substantial advancements alongside the rapid development of Large Language Models (LLMs) (Bai et al. 2023; Touvron et al. 2023) and multi-modal pre-training methods (Radford et al. 2021; Jia et al. 2021). These models have shown remarkable effectiveness in a wide range of applications, including image captioning (Xu et al. 2015), VQA (Antol et al. 2015), and visual editing (Chen et al. 2018). However, current MLLMs lack the capability to generate dense pixel-level outputs for precise spatial understanding. This limitation poses a considerable challenge in directly addressing tasks that require pixel-level comprehension of visual data, such as image segmentation, which is the most critical task in the field of computer vision.

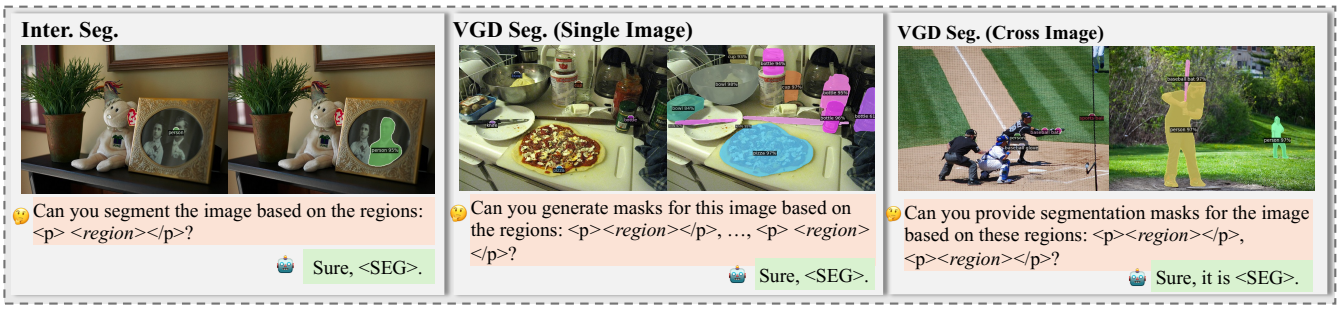
The Segment Anything Model (SAM) represents a foundational segmentation model that demonstrates exceptional efficacy in generating dense segmentation masks and has inspired the development of various segmentation tasks, such as high-quality segmentation (Ke et al. 2023), matching anything (Li et al. 2024a), and tracking anything (Rajič et al. 2025). Nevertheless, SAM's architecture is fundamentally constrained by its dependency on visual prompts, which significantly limits its direct applicability to a wide range of image segmentation tasks, including generic (semantic, instance, panoptic) segmentation, referring segmentation, and open-vocabulary (OV) segmentation, among others. Achieving a unified framework capable of addressing various image segmentation tasks remains a challenging problem.

In this work, we introduce X-SAM, an innovative framework that unifies diverse image segmentation tasks, expanding the segmentation paradigm from *segment anything* to *any segmentation*. To accomplish this objective, our approach addresses three critical technical challenges: (1) *Task formulation*: Transforming SAM into a versatile segmen-

*Corresponding author.



(a) Text Query



(b) Vision Query

Figure 2: Illustration of the capabilities of X-SAM. (a). Text query tasks: Generic (Gen.), Referring(Ref.), Reasoning(Rea.), and Grounded Conversation Generation(GCG) segmentation, *etc.*. (b). Vision query tasks: Interactive(Inter.) and Visual GrounDed (VGD) segmentation for single and cross-image. Additional examples of supported tasks can be found in (Wang et al. 2025).

tation architecture with cross-task applicability. (2) *Modality enhancement*: Augmenting LLMs with multimodal input processing capabilities. (3) *Unified framework*: Developing a cohesive approach to effectively facilitate comprehensive segmentation applications across diverse domains.

First, we develop a unified segmentation MLLM architecture that incorporates a unified mask decoder capable of generating segmentation masks suitable for generalized image segmentation tasks. Second, we expand the multimodal capabilities of MLLMs to process not only textual queries but also visual queries. Specifically, we introduce a novel task termed Visual GrounDed (VGD) segmentation, which segments all instance objects with interactive visual prompts in an image. This task introduces visual guide modalities into large language models (LLMs). Moreover, we propose a unified input format and training methodology that reformulates segmentation tasks within a unified framework, thus optimizing the adaptation of MLLMs to diverse image segmentation tasks.

As shown in Figure 2 and Table 1, we present the comprehensive capabilities of X-SAM and compare them with those of other methods. Our proposed framework exhibits capabilities in processing text query-based tasks, such as generic segmentation and referring segmentation, while simultaneously accommodating vision query-based tasks such as interactive segmentation (Zhang et al. 2024c) and our novel VGD segmentation, which functions effectively in both single-image and cross-image contexts. Furthermore, X-SAM leverages the reasoning and generative capacities of LLMs, thereby enabling advanced reasoning segmentation and Grounded Conversation Generation (GCG) (Rasheed et al. 2024) segmentation.

X-SAM undergoes co-training with a diverse range of datasets. We perform a comprehensive evaluation on more than twenty segmentation datasets across seven distinct image segmentation tasks, even including the image conversion task. X-SAM achieves the state-of-the-art performance across all image segmentation benchmarks, and establishes a robust new baseline for unified pixel-level image understanding, as illustrated in Figure 1. In summary, our contributions are as follows:

- We introduce X-SAM, a novel unified framework that extends the segmentation paradigm from *segment anything* to *any segmentation*. Our approach formulates diverse image segmentation tasks into a standardized segmentation format.
- We propose a new image segmentation benchmark, Visual GrounDed (VGD) segmentation, which provides visual grounded prompts for MLLMs to segment instance objects in images. The benchmark introduces user-friendly inputs to ground the segmentation objects and guide the MLLMs to output the segmentation masks.
- We present a unified multi-stage training strategy to co-train X-SAM with a diverse range of datasets, and conduct extensive evaluations on more than twenty image segmentation benchmarks, achieving state-of-the-art performance on all of them. This establishes a new strong baseline for unified pixel-level perceptual understanding in MLLMs.

Related Work

Multi-modal Large Language Model. Multi-modal learning has evolved from early models focused on task-specific fusion and feature extraction (Li et al. 2022b), to leveraging

Method	Text Query					Vision Query	
	Gen. Seg.	OV Seg.	Ref. Seg.	Rea. Seg	GCG Seg.	Inter. Seg.	VGD Seg.
SAM(Kirillov et al. 2023a)						✓	
Mask2Former(Cheng et al. 2022a)	✓						
ODISE(Xu et al. 2023)	✓	✓					
UNINEXT(Yan et al. 2023)	✓		✓			✓	
SEEM(Zou et al. 2023)	✓	✓	✓			✓	
OMG-Seg(Li et al. 2024b)	✓	✓				✓	
LISA(Lai et al. 2024)			✓	✓			
GLaMM(Rasheed et al. 2024)			✓		✓		
PixelLM(Zhang et al. 2024c)			✓				
OMG-LLaVA(Zhang et al. 2024b)	✓		✓		✓		
Sa2VA(Yuan et al. 2025)			✓	✓	✓		
PSALM(Zhang et al. 2024c)	✓	✓	✓			✓	✓
X-SAM (Ours)	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparison of Capability. We compare different methods on both segmentation-specific (Gray) and MLLM-based.

large language models (Brown et al. 2020; Touvron et al. 2023) for generalized, instruction-tuned multi-task benchmarks (Liu et al. 2024b; Hudson et al. 2019). LLaVA (Liu et al. 2023a, 2024a) introduced visual feature tokenization, inspiring advances in visual representation (Yuan et al. 2024b), specialized vision extensions (Lai et al. 2024; Lin et al. 2023), and language-guided segmentation (Li et al. 2024b; Zhang et al. 2024a). However, most progress remains task-specific. To our knowledge, we are the first to successfully implement a comprehensive approach, opening new directions for image segmentation.

Multi-modal Grounded Segmentation. Recent works (Pan et al. 2024; Wang et al. 2023) explore visual initiation methods in vision, including learnable tokens (Zhou et al. 2022a), mask-visual-modeling (Fang et al. 2023), and visual prompting encoders (Yuan et al. 2024a). SAM (Kirillov et al. 2023b) and its extensions (Xu et al. 2024) introduce visual grounding signals to segmentation models, greatly improving performance. Interactive segmentation (Li et al. 2024b) further enhances user-guided segmentation for MLLMs. However, existing methods cannot freely treat grounded input as textual input for segmentation. To address this, we propose Visual GrounDed (VGD) segmentation, enabling more diverse multi-modal grounded segmentation.

Unified Segmentation Model. Vision transformers (Dosovitskiy et al. 2020) have advanced universal segmentation, with recent works (Xu et al. 2024; Cheng et al. 2022b) developing end-to-end mask classification frameworks that outperform earlier models (Zhou et al. 2022b; Wang et al. 2021) across various applications. Research has expanded to open-world and open-vocabulary segmentation (Yuan et al. 2024a; Qi et al. 2022a,b), as well as unified architectures for multiple tasks (Athar et al. 2023; Jain et al. 2023; Li et al. 2024b). However, most methods focus solely on visual segmentation and lack interactive textual and visual prompts found in MLLMs. To address this, we combine SAM with MLLMs, extending SAM from *segment anything* to *any segmentation*, and introduce a unified framework adaptable to all image segmentation tasks, establishing a new strong baseline.

Method

To achieve unified image segmentation, we present X-SAM, a novel multi-modal segmentation MLLM. We design a versatile input format and a unified framework to integrate diverse segmentation tasks into a single model. Additionally, we introduce an innovative training strategy that enables SAM to handle any segmentation task. The following sections detail our methodology.

Formulation

The development of a unified segmentation model is fraught with challenges stemming from the diverse nature of segmentation tasks and the variability in input format. To address these issues, we introduce a versatile input format tailored to support a wide range of image segmentation tasks, laying the groundwork for the unified framework of X-SAM. We delineate the input format into two primary categories: text query input and vision query input. The text query input consists exclusively of linguistic prompts derived from user requests, the vision query input integrates both linguistic prompts and visual prompts provided by the user.

Text Query Input. The majority of existing image segmentation tasks can be conceptualized as text query inputs, including generic segmentation (Kirillov et al. 2019), referring segmentation, open-vocabulary (OV) segmentation (Li et al. 2022a), GCG segmentation (Rasheed et al. 2024), and reasoning segmentation (Lai et al. 2024). A text query input encapsulates the user's request along with the specific category or object to be segmented, which may be embedded within the user's prompt or generated by a large language model (LLM). To facilitate the GCG segmentation task, inspired by GLaMM (Rasheed et al. 2024), we incorporate two special phrase tokens, and , into the tokenizer to denote the beginning and end of a phrase, respectively. For each category in generic segmentation and GCG segmentation, phrase in referring segmentation, or sentence in reasoning segmentation, the format is standardized as

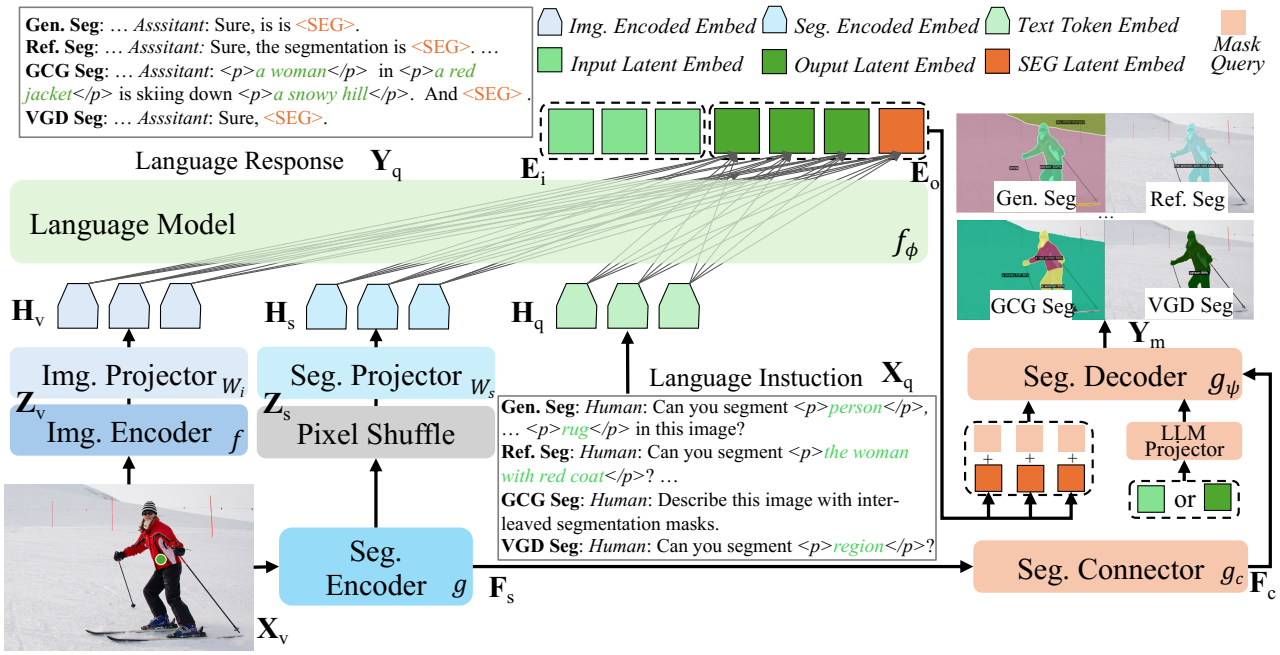


Figure 3: The Overview of X-SAM. X-SAM comprises dual encoders, dual projectors, a language model, a segmentation connector, and a segmentation decoder. The dual encoders process the image and project features to match text embedding dimensions, which are then input to the language model with tokenized text for instruction-guided understanding. The SAM features are connected to the segmentation decoder, which uses the LLM's <SEG> token to generate segmentation masks.

" category/phrase/sentence ". Specifically, the and tokens are not only encoded in the input tokens but also generated in the output tokens, ensuring consistency across different tasks. Additionally, for the output, we introduce a special token <SEG> into the tokenizer to signify the segmentation result, following the approach in (Lai et al. 2024).

Vision Query Input. Beyond text query inputs, some tasks necessitate vision query inputs, such as interactive segmentation (Zhang et al. 2024c) and the Visual GrounDed segmentation proposed in this work. In contrast to text query inputs, vision query inputs incorporate a visual prompt from the user, which may take the form of points, scribbles, boxes, or masks. To denote the visual prompt, we employ a dedicated token, <region>, within the input format. Analogous to the text query input, the visual prompt is formatted as " <region> " and the segmentation output is similarly indicated by the <SEG> token. The <region> token serves as a placeholder for the visual prompt and will be replaced by the region feature extracted from the segmentation encoder.

Unified Formulation. The latent language embeddings between the and tokens are used as the condition embeddings for the segmentation decoder to compute the classification scores. Based on this formulation, we achieve a unified framework for all image segmentation tasks. Given an input image $\mathbf{X}_v \in \mathbb{R}^{H\times W\times 3}$ and a language instruction $\mathbf{X}_q \in \mathbb{R}^{P\times 1}$, the model takes the image and language instruction as inputs and outputs a language response $\mathbf{Y}_q \in \mathbb{R}^{L\times 1}$ and a segmentation mask $\mathbf{Y}_m \in \mathbb{R}^{H\times W}$. Here,

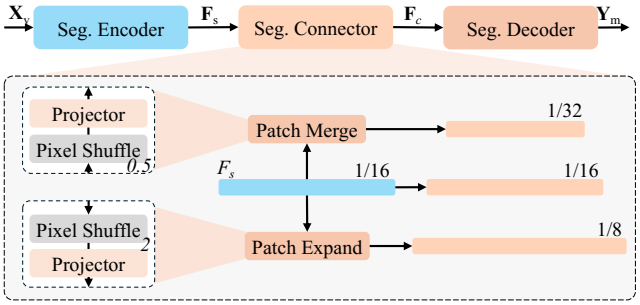


Figure 4: The Architecture of Segmentation Connector.

P is the length of the input text tokens, and L is the total length of the input and output text tokens. H and W denote the height and width of the image, respectively. Detailed input format examples can be found in Figure 2 (a) and (b).

Architecture

In this section, we propose X-SAM, a unified segmentation MLLM for any segmentation. As shown in Figure 3, it includes dual encoders, dual projectors, an LLM, a segmentation connector, and a segmentation decoder.

Dual Encoders. There are two encoders in X-SAM, an image encoder and a segmentation encoder. The image encoder f is used to extract the global image feature $\mathbf{Z}_v = f(\mathbf{X}_v)$, while the segmentation encoder g extracts the fine-grained image feature $\mathbf{Z}_s = g(\mathbf{X}_v)$. The feature from the image encoder is global and benefits image understanding tasks,

Method	Gen. Seg. Pan. / Ins. / Sem.	OV Seg. Pan. / Ins. / Sem.	Ref. Seg. RefCOCO / + / g	Rea. Seg. Val / Test	GCG Seg. Val / Test	Inter. Seg. Point / Box	VGD Seg. Point / Box
SAM-L(Kirillov et al. 2023a)	✗	✗	✗	✗	✗	51.8 / 76.6	12.8 / 31.7
Mask2Former-L(Cheng et al. 2022a)	57.8 / 48.6 / 67.4	✗	✗	✗	✗	✗	✗
SEEM-B(Zou et al. 2023)	56.1 / 46.4 / 66.3	✗	- / - / 65.6	✗	✗	47.8 / 44.9	✗
ODISE(Xu et al. 2023)	55.4 / 46.0 / 65.2	22.6 / 14.4 / 29.9	✗	✗	✗	✗	✗
OMG-Seg(Li et al. 2024b)	53.8 / - / -	✗	✗	✗	✗	-	✗
LISA-7B(Lai et al. 2024)	✗	✗	74.9 / 65.1 / 67.9	52.9 / 47.3	✗	✗	✗
GLaMM(Rasheed et al. 2024)	✗	✗	79.5 / 72.6 / 74.2	✗	65.8 / 64.6	✗	✗
PixelLM-7B(Ren et al. 2024)	✗	✗	73.0 / 66.3 / 69.3	✗	✗	✗	✗
OMG-LLaVA-7B(Zhang et al. 2024b)	53.8 / - / -	✗	78.0 / 69.1 / 72.9	✗	65.5 / 64.7	✗	✗
Sa2VA-8B(Yuan et al. 2025)	✗	✗	81.6 / 76.2 / 78.7	- / -	- / -	✗	✗
PSALM(Zhang et al. 2024c)	**55.9** / 45.7 / **66.6**	13.7 / 9.0 / 18.2	83.6 / 72.9 / 73.8	✗	✗	64.3 / 67.3	2.0 / 3.7
X-SAM (Ours)	54.7 / **47.0** / 66.5	**20.9 / 16.2 / 28.8**	**85.1 / 78.0 / 83.8**	**56.6 / 57.8**	**69.4 / 69.0**	**65.4 / 70.0**	**47.9 / 49.5**

Table 2: Comprehensive Performance Comparison. We compare X-SAM to segmentation-specific models (Gray) and MLLMs. "✗" denotes unsupported tasks. "–" indicates unreported results. X-SAM achieves state-of-the-art performance across all segmentation tasks with a single model. Best results are in **bold**, second-best are underlined.

whereas the feature from the segmentation encoder is fine-grained and benefits image segmentation tasks. We adopt SigLIP2-so400m (Tschannen et al. 2025) as the image encoder and SAM-L (Ke et al. 2023) as the segmentation encoder.

Dual Projectors. To enhance the LLM's understanding of the image, we concatenate the features from the image encoder and the segmentation encoder before passing them to the LLM. Specifically, the feature from the segmentation encoder is too large to be processed directly by the LLM, so we utilize a pixel-shuffle operation to reduce its spatial size. We then project the reduced feature into the language embedding space $\mathbf{H}_q$ via an MLP projector W_s. For the feature from the image encoder, we directly project it into the language embedding space via an MLP projector W_i, such that $\mathbf{H}_v = W_i \cdot \mathbf{Z}_{\mathrm{v}}$ and $\mathbf{H}_s = W_s \cdot \mathbf{Z}_{\mathrm{s}}$. We then concatenate the features from dual projectors and the language embeddings, and input them into the LLM f_ϕ.

Segmentation Connector. For image segmentation tasks, fine-grained multi-scale features are crucial for the segmentation decoder to accurately predict segmentation masks. The output of the segmentation encoder in SAM is single-scale (1/16) with reduced spatial resolution. To obtain multi-scale features, we design a segmentation connector g_{c}, to bridge the segmentation encoder and decoder. As shown in Figure 4, we perform patch-merge using a pixel-shuffle (Chen et al. 2024) with a scale of 0.5 to reduce the spatial size of the last feature in the encoder to a smaller scale (1/32). We also perform patch-expand with a pixel-shuffle of scale 2.0 to increase the spatial size of the last feature to a larger scale (1/8), resulting in multi-scale features for the segmentation decoder.

Segmentation Decoder. The Segment Anything Model (SAM) can segment a single object based on input text or visual prompts, but it fails to segment all objects in a single inference. To segment all objects at once, we replace its original segmentation decoder with a new decoder, following the approach in (Cheng et al. 2022a; VS et al. 2024). The segmentation decoder g_ψ predicts masks and their category probabilities from either the input latent embedding $\mathbf{E}_{\mathrm{i}}$ or the output latent embedding $\mathbf{E}_{\mathrm{o}}$, multi-scale segmentation features $\mathbf{F}_{\mathrm{c}}$, and a set of mask query tokens plus the <SEG> token embedding, which bridges the LLM output with the segmentation decoder. Notably, we introduce a latent background embedding to represent the "ignore" category for all tasks, thereby unifying all image segmentation tasks with one model.

Training

To improve the performance on diverse image segmentation tasks, we propose a novel multi-stage training strategy. The training strategy consists of three stages: segmentor fine-tuning, alignment pre-training, and mixed fine-tuning.

Stage 1: Segmentor Fine-tuning. As the segmentation decoder is redesigned, we need to train the segmentor to adapt to segment all objects in a single forward pass. We follow the training pipeline in (Cheng et al. 2022a), which trains the model on the popular COCO-Panoptic (Kirillov et al. 2019) dataset. To enable faster convergence during training, we unfreeze all the parameters in the segmentor while training the segmentation encoder with a lower learning rate. The training objective, $\mathcal{L}_{\mathrm{seg}}$, is the same as in (Cheng et al. 2022a), and is defined as the sum of the classification loss $\mathcal{L}_{\mathrm{cls}}$, the mask loss $\mathcal{L}_{\mathrm{mask}}$, and the dice loss $\mathcal{L}_{\mathrm{dice}}$:

$$\mathcal{L}_{\mathrm{seg}} = \mathcal{L}_{\mathrm{cls}} + \mathcal{L}_{\mathrm{mask}} + \mathcal{L}_{\mathrm{dice}} \tag{1}$$

Stage 2: Alignment Pre-training. To align the language embeddings and visual embeddings, we perform alignment pre-training on the LLaVA-558K dataset, following (Liu et al. 2023b). We keep the dual encoders and the LLM parameters frozen, and only train the dual projectors. In this way, the image embeddings and segmentation embeddings can be aligned with the pre-trained LLM word embeddings. The training objective for alignment pre-training is an auto-regressive loss $\mathcal{L}_{\mathrm{regressive}}$:

$$\mathcal{L}_{\mathrm{regressive}} = -\sum_{i=1}^{N} \log p_\theta \left(\mathcal{Y}_q^{[P+i]} | \mathcal{Y}_q^{[:i-1]}, \mathcal{X}_q^{[:i-1]} \right), \tag{2}$$

where $\mathcal{X}_q$ is the input sequence $\mathcal{X}_q = [x_1, x_2, ..., x_p] \in \mathbb{R}^{P \times D}$, $\mathcal{Y}_q$ is the output sequence $\mathcal{Y}_q = [y_1, y_2, ..., y_l] \in$

Method	(M)LLM	RefCOCO			RefCOCO+			RefCOCOg	
		val	testA	testB	val	testA	testB	val	test
SEEM-L (Zou et al. 2023)	-	-	-	-	-	-	-	65.6	-
UNINEXT-L (Yan et al. 2023)	-	80.3	82.6	77.8	70.0	74.9	62.6	73.4	73.7
UNINEXT-H (Yan et al. 2023)	-	82.2	83.4	81.3	72.5	76.4	66.2	74.7	76.4
GLaMM (Rasheed et al. 2024)	Vicuna-7B	79.5	83.2	76.9	72.6	78.7	64.6	74.2	74.9
OMG-LLaVA (Zhang et al. 2024b)	InternLM-7B	77.2	79.8	74.1	68.7	73.0	61.6	71.7	71.9
Sa2VA(Yuan et al. 2025)	InternVL2-8B	81.6	-	-	76.2	-	-	78.7	-
PSALM (Zhang et al. 2024c)	Phi-1.5-1.3B	83.6	84.7	81.6	72.9	75.5	70.1	73.8	74.4
HyperSeg (Wei et al. 2024)	Mipha-3B	84.8	85.7	**83.4**	**79.0**	**83.5**	**75.2**	79.4	78.9
X-SAM (Ours)	Phi-3-3.8B	**85.1**	**87.1**	**83.4**	78.0	81.0	74.4	**83.8**	**83.9**

Table 3: Comparison of Referring Segmentation. We evaluate methods on referring segmentation benchmarks by (M)LLMs.

$\mathbb{R}^{L\times D}$, where $L = P + N$ represents the length of output sequence, D represents the hidden size of LLM. θ is a trainable parameter in LLM, and we only calculate the loss for the generated text.

Stage 3: Mixed Fine-tuning. X-SAM is co-trained on multiple datasets across diverse tasks in an end-to-end manner. For the image conversation task, we adopt the auto-regressive loss $\mathcal{L}_{\text{regressive}}$ as is common in MLLM training. For the segmentation tasks, we not only use the segmentation loss as in segmentor training, but also add the auto-regressive loss to the training objective. Benefiting from the unified formulation and simple training objective, end-to-end mixed fine-tuning across diverse tasks can be performed within a unified framework. The training objective for mixed fine-tuning can be formulated as:

$$\mathcal{L}_{\text{total}} = \begin{cases} \mathcal{L}_{\text{regressive}}, & \text{conversation} \\ \mathcal{L}_{\text{regressive}} + \mathcal{L}_{\text{seg}}, & \text{segmentation} \end{cases} \quad (3)$$

Experiments

Experiment Settings

Datasets and Tasks. For segmentor fine-tuning, we train on the COCO-Panoptic (Kirillov et al. 2019) dataset. For alignment pre-training, we utilize the LLaVA-558K (Liu et al. 2023b) dataset. For end-to-end mixed fine-tuning, we incorporate one image conversation dataset and five types of image segmentation datasets into the training process. To balance the training data across these diverse datasets, we set the training epoch to 1 and adjust the resampling rates of different datasets using dataset balance resampling. After training, X-SAM is capable of performing a variety of tasks, including Image Conversation, Generic, Referring, Reasoning, GCG, Interactive, and VGD Segmentation. Additionally, X-SAM supports Open-Vocabulary (OV) (OV-semantic, OV-instance, OV-panoptic) segmentation, enabling it to segment all objects defined by the input prompt, even those never seen before. Note that COCO-VGD is our proposed VGD segmentation dataset, which is built on the COCO2017 dataset. Details of the datasets are presented in (Wang et al. 2025).

Evaluation Metrics. We conduct extensive experiments to evaluate the performance of X-SAM. For generic segmentation and open-vocabulary segmentation, we use PQ, mIoU, and mAP as the main metrics for panoptic, semantic, and instance segmentation, respectively. For referring segmentation and reasoning segmentation, we adopt cIoU and gIoU as metrics, following (Zhang et al. 2024c). For GCG segmentation, we use M, C, AP50, and mIoU as metrics, following (Rasheed et al. 2024). For interactive segmentation, we use mIoU and cIoU, also following (Zhang et al. 2024c). For VGD segmentation, we use AP and AP50. For image conversation, we adopt scores from common MLLM benchmarks as the main metrics, following (Liu et al. 2023b).

Implementation Details. We adopt the XTuner (Contributors 2023) codebase for training and evaluation. During segmentor fine-tuning, we train all parameters, set the batch size to 64, and use a learning rate of 1e-5 for the SAM encoder and 1e-4 for the other parameters. The number of training epochs is set to 36. For alignment pre-training, we train only the dual projector parameters, with a batch size of 256, a learning rate of 1e-3, and one training epoch. For end-to-end mixed fine-tuning, we train all parameters, set the batch size to 64, and use a learning rate of 4e-6 for the dual encoders and 4e-5 for the other parameters, with one training epoch. All training is conducted on 16 A100 GPUs. For image conversation evaluation, we use the VLMEvalKit (Duan et al. 2024) codebase to evaluate performance on MLLM benchmarks. For segmentation task evaluation, we follow the settings described in the corresponding papers and repositories. More implementation details are provided in (Wang et al. 2025).

Main Results

We conduct extensive evaluation on seven segmentation tasks, including Generic, Open-Vocabulary, Referring, Reasoning, GCG, Interactive, and VGD Segmentation.

Overall. In Table 2, we compare X-SAM with current segmentation-specific models and MLLMs. X-SAM demonstrates the most comprehensive capabilities. It achieves performance comparable to state-of-the-art in generic segmentation, and achieves the best performance on other benchmarks, with a single model. X-SAM sets a new state-of-the-art record for image segmentation benchmarks. Detailed results for each task are discussed below.

Referring Segmentation. We evaluate X-SAM on RefCOCO, RefCOCO+, and RefCOCOg, with the results

Methods	Val				Test			
	METEOR	CIDEr	AP50	mIoU	METEOR	CIDEr	AP50	mIoU
Kosmos-2(Peng et al. 2023)	**16.1**	27.6	17.1	55.6	**15.8**	27.2	17.2	56.8
LISA-7B(Lai et al. 2024)	13.0	33.9	25.2	62.0	12.9	32.2	24.8	61.7
GLaMM-7B†(Rasheed et al. 2024)	15.2	43.1	28.9	65.8	14.6	37.9	27.2	64.6
OMG-LLaVA-7B(Zhang et al. 2024b)	14.9	41.2	29.9	65.5	14.5	38.5	28.6	64.7
X-SAM (Ours)	15.4	**46.3**	**33.2**	**69.4**	15.1	**42.7**	**32.9**	**69.0**

Table 4: Comparison of GCG Segmentation. † indicates pretraining with the GranD dataset (Rasheed et al. 2024).

Method	Point		Scribble		Box		Mask	
	AP	AP50	AP	AP50	AP	AP50	AP	AP50
PSALM† (Zhang et al. 2024c)	2.0	3.3	2.8	4.4	3.7	5.8	2.3	3.3
SAM† (Kirillov et al. 2023b)	12.8	22.8	-	-	31.7	50.1	-	-
X-SAM (Ours)	**47.9**	**72.5**	**48.7**	**73.4**	**49.5**	**74.7**	**49.7**	**74.9**

Table 5: Comparison of VGD Segmentation. † indicates evaluation results following X-SAM setting.

FT	COCO-Pan PQ	A150-OV PQ	RefCOCO cIoU	Reason-Val gIoU
Specific	55.3	16.4	81.0	48.2
Mixed	54.5(↓ 0.8)	22.4(↑ 6.0)	85.4(↑ 4.4)	57.1(↑ 8.9)

Table 6: Ablation on Fine-Tuning(FT).

Encoder		COCO-Pan	A150-OV	GCG-Val	COCO-VGD
Img.	Seg.	PQ	PQ	mIoU	AP
ViT	-	54.5	16.4	64.8	40.7
ViT	Swin†	56.2(↑ 1.7)	18.6(↑ 2.2)	62.5(↓ 2.3)	48.6(↑ 7.9)
ViT	SAM	54.7(↑ 0.2)	20.9(↑ 4.5)	69.4(↑ 4.6)	47.9(↑ 7.2)

Table 7: Ablation on Dual Encoders. Swin† is initialized from Mask2Former (M2F) (Cheng et al. 2022a).

shown in Table 3. X-SAM outperforms PSALM (Zhang et al. 2024c) by 1.5% cIoU, 5.1% cIoU, and 10.0% cIoU on the validation sets of RefCOCO, RefCOCO+, and RefCOCOg, respectively. Compared to Sa2VA-8B (Yuan et al. 2025), X-SAM achieves better results with a smaller model size. It shows performance improvements of 3.5% cIoU, 1.8% cIoU, and 5.1% cIoU on RefCOCO, RefCOCO+, and RefCOCOg, respectively.

GCG Segmentation. Grounded conversation generation demands detailed image and pixel-level understanding, requiring MLLMs to link captioned objects to their segmentation masks. As shown in Table 4, X-SAM achieves a significant performance improvement compared to previous methods and obtains the best results on both the `Val` and `Test` sets. In terms of image-level understanding, X-SAM outperforms GLaMM (Rasheed et al. 2024) by 0.2% METEOR and 3.2% CIDEr on the `Val` set, and by 0.5% METEOR and 4.8% CIDEr on the `Test` set. In terms of pixel-level understanding, X-SAM outperforms OMG-LLaVA (Zhang et al. 2024b) by 3.3% AP and 3.9% mIoU on the `Val` set, and by 4.3% AP and 4.3% mIoU on the `Test` set.

VGD Segmentation. Visual grounded segmentation demands vision query understanding, requiring MLLMs to comprehend the visual modality and segment all related instances. Table 5 presents the VGD segmentation results. As VGD segmentation is our newly proposed task, we evaluate PSALM (Zhang et al. 2024c) following X-SAM's settings. X-SAM outperforms PSALM by 45.9% AP, 45.9% AP, 45.8% AP, and 47.4% AP on Point, Scribble, Box, and Mask visual prompts, respectively.

More results and discussions for other segmentation and conversation benchmarks are provided in (Wang et al. 2025).

Abaltions

We conduct ablation studies on mixed fine-tuning, dual encoders, multi-stage training, and segmentor architecture, presenting selected benchmark results due to space limitations.

Mixed Fine-tuning. We ablate the impact of mixed fine-tuning on X-SAM's performance. As shown in Table 6, mixed fine-tuning improves performance on out-of-domain COCO benchmarks, demonstrating X-SAM's robust segmentation capabilities—for example, a 6.0% PQ increase on A150-OV and an 8.9% gIoU increase on Reason-Val. However, it results in a 0.8% PQ decrease on COCO-Pan due to the challenge of balancing performance in multi-sources training.

Dual Encoders. We ablate the design of the dual encoders in X-SAM. As shown in Table 7, dual encoders with either a SAM or Swin encoder benefit VGD segmentation, achieving 7.2% AP and 7.9% AP on COCO-VGD, respectively. Moreover, dual encoders with a SAM encoder consistently improve performance on GCG-Val and A150-OV, while the Swin encoder, which lacks robust segmentation capabilities, provides only a small improvement on A150-OV and even has a negative impact on GCG-Val.

M-Stage	COCO-Pan PQ	A150-OV PQ	GCG-Val mIoU	Conv.-MMB Acc.
S3	45.2	19.4	60.6	67.2
S1, S3	54.5(↑ 9.3)	20.9(↑ 1.5)	65.4(↑ 4.8)	67.4(↑ 0.2)
S1, S2, S3	54.7(↑ 9.5)	20.9(↑ 1.5)	69.4(↑ 8.8)	69.3(↑ 2.1)

Table 8: Ablation on Multi-Stage(M-Stage) Training. S1: Stage 1, S2: Stage 2, S3: Stage 3, Conv.: conversation.

Conn.	Decoder	M-Scale	PQ	AP	mIoU
-	SAM	✗	40.9	26.3	49.5
MLP	M2F	✗	50.1(↑ 9.2)	38.9(↑ 12.6)	60.2(↑ 10.7)
Con.	M2F	✗	50.3(↑ 9.4)	39.1(↑ 12.8)	60.6(↑ 11.1)
Con.	M2F	✓	51.6(↑ 10.7)	41.5(↑ 15.2)	61.6(↑ 12.1)

Table 9: Ablation on Segmentor Architecture. Conn.: connector, M-Scale: multi-scale, Con.: convolution, M2F: Mask2Former.

Multi-stage Training. We ablate the impact of the multi-stage training strategy. As shown in Table 8, the S1 segmentor fine-tuning phase boosts the segmentation capability, producing a notable improvement of 9.3% PQ in COCO-Pan and 1.5% PQ in the A150-OV datasets. Meanwhile, the S2 alignment pre-training phase enhances image understanding capabilities, contributing an additional 2.1% Accuracy on Conv.-MMB. By integrating these stages, X-SAM demonstrates robust advances in image segmentation and comprehension, establishing its effectiveness in addressing complex visual tasks.

Segmentor Architecture. We ablate the impact of segmentor architecture by performing segmentor fine-tuning for 12 epochs. As shown in Table 9, M2F decoder brings a large improvement with 9.2% PQ as the effective design of M2F. The convolution connector performs better than the MLP connector, as the convolution spatial-awareness benefits segmentation, and multi-scale further improves the performance(10.7% PQ) with more diverse scale features.

More ablation results can be found in (Wang et al. 2025).

Conclusion

In this work, we propose X-SAM, a unified segmentation MLLM that extends the segmentation paradigm from *segment anything* to *any segmentation*, integrating all image segmentation tasks into a single model. Our method can process various multimodal inputs in MLLMs, including both text and visual queries. Moreover, to equip MLLMs with visual grounded perception capabilities, we introduce a new segmentation task, Visual GrounDed (VGD) segmentation, further extending the capabilities of the unified segmentation model. We conduct extensive experiments across all image segmentation tasks, and X-SAM achieves state-of-the-art performance on each task with a single model.

Acknowledgments

This work is supported by National Key Research and Development Program of China (2024YFE0203100), Scientific Research Innovation Capability Support Project for Young Faculty (No.ZYGXQNJSKYCXNLZCXM-I28), National Natural Science Foundation of China (NSFC) under Grants No.62476293, General Embodied AI Center of Sun Yat-sen University, National Natural Science Foundation of China (62402252) and (62536003) Guangdong High-Level Talent Programme (2024TQ08X283).

References

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433.

Athar, A.; Hermans, A.; Luiten, J.; Ramanan, D.; and Leibe, B. 2023. Tarvis: A unified approach for target-based video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18738–18748.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chen, J.; Shen, Y.; Gao, J.; Liu, J.; and Liu, X. 2018. Language-based image editing with recurrent attentive models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8721–8729.

Chen, Z.; Wang, W.; Tian, H.; Ye, S.; Gao, Z.; Cui, E.; Tong, W.; Hu, K.; Luo, J.; Ma, Z.; et al. 2024. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12): 220101.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022a. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1290–1299.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022b. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1290–1299.

Contributors, X. 2023. XTuner: A Toolkit for Efficiently Fine-tuning LLM. https://github.com/InternLM/xtuner.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Duan, H.; Yang, J.; Qiao, Y.; Fang, X.; Chen, L.; Liu, Y.; Dong, X.; Zang, Y.; Zhang, P.; Wang, J.; et al. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 11198–11201.

Fang, Z.; Li, X.; Li, X.; Buhmann, J. M.; Loy, C. C.; and Liu, M. 2023. Explore in-context learning for 3d point cloud understanding. *Advances in Neural Information Processing Systems*, 36: 42382–42395.

Hudson, D. A.; and Manning, C. D. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6700–6709.

Jain, J.; Li, J.; Chiu, M. T.; Hassani, A.; Orlov, N.; and Shi, H. 2023. Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2989–2998.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, 4904–4916. PMLR.

Ke, L.; Ye, M.; Danelljan, M.; Tai, Y.-W.; Tang, C.-K.; Yu, F.; et al. 2023. Segment anything in high quality. *Advances in Neural Information Processing Systems*, 36: 29914–29934.

Kirillov, A.; He, K.; Girshick, R.; Rother, C.; and Dollár, P. 2019. Panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9404–9413.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023a. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023b. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2024. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9579–9589.

Li, B.; Weinberger, K. Q.; Belongie, S.; Koltun, V.; and Ranftl, R. 2022a. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022b. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Li, S.; Ke, L.; Danelljan, M.; Piccinelli, L.; Segu, M.; Van Gool, L.; and Yu, F. 2024a. Matching anything by segmenting anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18963–18973.

Li, X.; Yuan, H.; Li, W.; Ding, H.; Wu, S.; Zhang, W.; Li, Y.; Chen, K.; and Loy, C. C. 2024b. Omg-seg: Is one model good enough for all segmentation? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 27948–27959.

Lin, B.; Ye, Y.; Zhu, B.; Cui, J.; Ning, M.; Jin, P.; and Yuan, L. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023a. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023b. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; et al. 2024b. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, 216–233. Springer.

Pan, T.; Tang, L.; Wang, X.; and Shan, S. 2024. Tokenize anything via prompting. In *European Conference on Computer Vision*, 330–348. Springer.

Peng, Z.; Wang, W.; Dong, L.; Hao, Y.; Huang, S.; Ma, S.; and Wei, F. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

Qi, L.; Kuen, J.; Guo, W.; Shen, T.; Gu, J.; Jia, J.; Lin, Z.; and Yang, M.-H. 2022a. High-quality entity segmentation. *arXiv preprint arXiv:2211.05776*.

Qi, L.; Kuen, J.; Wang, Y.; Gu, J.; Zhao, H.; Torr, P.; Lin, Z.; and Jia, J. 2022b. Open world entity segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7): 8743–8756.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Rajič, F.; Ke, L.; Tai, Y.-W.; Tang, C.-K.; Danelljan, M.; and Yu, F. 2025. Segment anything meets point tracking. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 9302–9311. IEEE.

Rasheed, H.; Maaz, M.; Shaji, S.; Shaker, A.; Khan, S.; Cholakkal, H.; Anwer, R. M.; Xing, E.; Yang, M.-H.; and Khan, F. S. 2024. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13009–13018.

Ren, Z.; Huang, Z.; Wei, Y.; Zhao, Y.; Fu, D.; Feng, J.; and Jin, X. 2024. Pixellm: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26374–26383.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Tschannen, M.; Gritsenko, A.; Wang, X.; Naeem, M. F.; Alabdulmohsin, I.; Parthasarathy, N.; Evans, T.; Beyer, L.; Xia, Y.; Mustafa, B.; et al. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.

VS, V.; Borse, S.; Park, H.; Das, D.; Patel, V.; Hayat, M.; and Porikli, F. 2024. PosSAM: Panoptic Open-vocabulary Segment Anything. arXiv:2403.09620.

Wang, H.; Qiao, L.; Jie, Z.; Huang, Z.; Feng, C.; Zheng, Q.; Ma, L.; Lan, X.; and Liang, X. 2025. X-SAM: From Segment Anything to Any Segmentation. *arXiv preprint arXiv:2508.04655*.

Wang, H.; Wang, W.; and Liu, J. 2021. Temporal memory attention for video semantic segmentation. In *2021 IEEE International Conference on Image Processing (ICIP)*, 2254–2258. IEEE.

Wang, X.; Zhang, X.; Cao, Y.; Wang, W.; Shen, C.; and Huang, T. 2023. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*.

Wei, C.; Zhong, Y.; Tan, H.; Liu, Y.; Zhao, Z.; Hu, J.; and Yang, Y. 2024. HyperSeg: Towards Universal Visual Segmentation with Large Language Model. *arXiv preprint arXiv:2411.17606*.

Xu, J.; Liu, S.; Vahdat, A.; Byeon, W.; Wang, X.; and De Mello, S. 2023. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2955–2966.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, 2048–2057. PMLR.

Xu, S.; Yuan, H.; Shi, Q.; Qi, L.; Wang, J.; Yang, Y.; Li, Y.; Chen, K.; Tong, Y.; Ghanem, B.; et al. 2024. Rapsam: Towards real-time all-purpose segment anything. *arXiv preprint arXiv:2401.10228*.

Yan, B.; Jiang, Y.; Wu, J.; Wang, D.; Luo, P.; Yuan, Z.; and Lu, H. 2023. Universal instance perception as object discovery and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15325–15336.

Yuan, H.; Li, X.; Zhang, T.; Huang, Z.; Xu, S.; Ji, S.; Tong, Y.; Qi, L.; Feng, J.; and Yang, M.-H. 2025. Sa2VA: Marrying SAM2 with LLaVA for Dense Grounded Understanding of Images and Videos. *arXiv preprint arXiv:2501.04001*.

Yuan, H.; Li, X.; Zhou, C.; Li, Y.; Chen, K.; and Loy, C. C. 2024a. Open-vocabulary SAM: Segment and recognize twenty-thousand classes interactively. In *European Conference on Computer Vision*, 419–437. Springer.

Yuan, Y.; Li, W.; Liu, J.; Tang, D.; Luo, X.; Qin, C.; Zhang, L.; and Zhu, J. 2024b. Osprey: Pixel understanding with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 28202–28211.

Zhang, T.; Li, X.; Fei, H.; Yuan, H.; Wu, S.; Ji, S.; Loy, C. C.; and Yan, S. 2024a. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. *Advances in Neural Information Processing Systems*, 37: 71737–71767.

Zhang, T.; Li, X.; Fei, H.; Yuan, H.; Wu, S.; Ji, S.; Loy, C. C.; and Yan, S. 2024b. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. *Advances in Neural Information Processing Systems*, 37: 71737–71767.

Zhang, Z.; Ma, Y.; Zhang, E.; and Bai, X. 2024c. Psalm: Pixelwise segmentation with large multi-modal model. In *European Conference on Computer Vision*, 74–91. Springer.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.

Zhou, Q.; Li, X.; He, L.; Yang, Y.; Cheng, G.; Tong, Y.; Ma, L.; and Tao, D. 2022b. TransVOD: End-to-end video object detection with spatial-temporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6): 7853–7869.

Zou, X.; Yang, J.; Zhang, H.; Li, F.; Li, L.; Wang, J.; Wang, L.; Gao, J.; and Lee, Y. J. 2023. Segment everything everywhere all at once. *Advances in neural information processing systems*, 36: 19769–19782.