SAM4MLLM: Enhance Multi-Modal Large Language Model for Referring Expression Segmentation

Yi-Chia Chen[1], Wei-Hua Li[1], Cheng Sun[2],
Yu-Chiang Frank Wang[1,2], and Chu-Song Chen[1]

[1] National Taiwan University, Taipei, Taiwan.
{r11922a04, d12922009, chusong}@csie.ntu.edu.tw
[2] NVIDIA, Taipei, Taiwan.
{chengs, frankwang}@nvidia.com

Abstract. We introduce SAM4MLLM, an innovative approach which integrates the Segment Anything Model (SAM) with Multi-Modal Large Language Models (MLLMs) for pixel-aware tasks. Our method enables MLLMs to learn pixel-level location information without requiring excessive modifications to the existing model architecture or adding specialized tokens. We introduce an inquiry-based approach that can effectively find prompt points for SAM to perform segmentation based on MLLM. It combines detailed visual information with the powerful expressive capabilities of large language models in a unified language-based manner without additional computational overhead in learning. Experimental results on pubic benchmarks demonstrate the effectiveness of our approach.

Keywords: LLM · MLLM · Referring Expression Segmentation.

1 Introduction

With the rapid growth of generative AI, large language models (LLM) [5,15,17, 46,47,49,60] become a focus of research and application due to their profound capabilities in understanding and generating text. They show innovative power in machine learning and marks the evolution of human-machine interaction.

Recently, the progress has been made from simple text processing to the complex multi-modal understanding. The advent of Multi-modal Large Language Models (MLLMs) [1,21,28,33,45,58,64] lies in incorporating image processing modules into LLMs. They successfully endow LLMs with the ability to process visual information, thereby bridging the significant gap between visual and linguistic tasks. Recent studies enabled MLLMs to engage in in-depth dialogues based on image content. Subsequent research enhanced MLLMs capabilities through data or structural modifications, leading to enhanced MLLMs that allow for the input and output of bounding boxes of objects to achieve fine-grained visual dialogues [6,10,11,31,37,59].

Referring Expression Segmentation (RES) [48,51,53,66] aims to label image pixels corresponding to specific objects or reuns mentioned in natural language expressions. It involves accurately identifying and segmenting the objects referred to by linguistic descriptions. In this paper, we focus on RES and use MLLM to solve this task. However, bounding boxes alone are insufficient for precisely indicating object locations within images. This has led to research efforts focused on improving annotation granularity to pixel-level for MLLM, enhancing image information encoding, and designing models capable of outputting detailed segmentation masks [27,39,42,50]. Despite significant progress, these advancements require substantial modifications to the original MLLMs architectures [42]. Some study introduces additional model structures to output segmentation masks [39]; some others leverage the use of special tokens different from those in original LLMs [27,50] or rely on the application of multiple losses for model optimization [27]. These adjustments introduce architectural complexity to MLLM and may complicate model extension for additional tasks.

In this paper, we propose a simple solution that can enhance MLLM's abilities to understand object localization in pixel level. Our approach is simple but effective, which upgrades MLLMs' visual capabilities to a new level for accurate understanding the referring expressions of pixel-level location in images.

Our method draws inspiration from the context provided below. Concurrently with the development of LLMs and MLLMs, the field of image segmentation has witnessed a significant breakthrough with the introduction of the Segment Anything Model (SAM) [25], a foundation model trained on the SA-1B [25] high-quality image segmentation dataset. SAM, a promptable segmentation model, can generate high-quality semantic-free segmentation masks in images based on prompts provided by the user, such as points or bounding boxes.

We observe that while MLLMs possess a profound understanding of image semantics, they struggle to articulate detailed pixel-level information. Conversely, SAM, although not semantically aware, can delineate intricate segmentation masks with minimal prompting. In light of this, we propose a novel methodology using SAM for MLLM (namely, SAM4MLLM) which seamlessly integrates MLLMs with SAM. Specifically, we employ a straightforward yet simple strategy, introducing pixel-level information into the training dataset without altering the original MLLM architecture. This enables MLLMs to grasp pixel-level information using the same text cross-entropy loss used by popular LLMs [6,17,23,47]. Considering the potential limitations of MLLMs in pixel expression due to input resolution constraints and a model architecture not explicitly designed for visual tasks, we further enhance the output with SAM, post-processing MLLM outputs to obtain higher precision segmentation masks in a relatively effortless manner. To establish a connection between SAM and MLLM, one straightforward approach is to enable MLLM to generate prompt points for SAM. However, effectively producing multiple points can be challenging. Therefore, we introduce a novel method that leverages the dialog capability of LLMs. Specifically, we proactively ask the MLLM to acquire effective prompt points for SAM. We tackle the problem of RES and demonstrate the effectiveness of our approach.

Main contributions of this work are as follows:

- We present SAM4MLLM, an approach allowing MLLMs to understand pixel-level details without altering the MLLM model architecture, introducing new tokens, or employing additional losses. It is simple yet effective for RES.
- To connect MLLM and SAM, we introduce a novel method of actively querying the language system to obtain prompt point cues.
- Through experiments on various RES benchmarks, including RES dataset, GRES, and ReasonSeg, we validate the effectiveness of SAM4MLLM and demonstrate its favorable performance in handling complex pixel-aware tasks.

2 Related Works

In this section, we review the related works of the topics: RES, image segmentation, MLLMs, and MLLMs toward segmentation.

Referring Expression Segmentation. Early researches in RES focus on integrating features from language and vision models to effectively merge these two types of information. Yu *et al.* [56] combine the language attention and subject, location, and relationship modules to localize the target region. In STEP [9], a DNN architecture is applied to iteratively refine the segmentation heatmap. Subsequently, Hui *et al.* [22] introduce Linguistic Structure guided Context Modeling (LSCM) to aggregate the multi-modal features. To understanding the language expression in different perspective, VLT [16] generates several sets of queries and introduces a query balance module to focus on the most suitable query. Zhu *et al.* [63] regard RES as a point prediction problem and design a simple transformer-based network to perform referring segmentation. Recently, the method in [51] leverages a novel adapter to facilitate cross-modal information. These studies lay the foundation for subsequent work on MLLMs in RES.

With the advancement of multi-modal model, they have been introduced into the RES field, enhancing the accuracy and efficiency of segmentation. Wang *et al.* [48] introduce the multi-modal model CLIP [40] to RES tasks. With the rise of MLLMs, research based on these models has emerged, leveraging their remarkable abilities in understanding text and images. [27, 39, 42, 50]. Furthermore, it has been pointed out in [32] that classical RES benchmarks are not sufficiently comprehensive in some cases, leading to the proposal of General Referring Expression Segmentation (GRES) [32] dataset to broaden its application scope. GRES allows for the reference to multiple objects simultaneously and can address the absence of objects in images, further enhancing the applicability in practice. In LISA [27], a more complex dataset, ReasonSeg, is proposed. It requires models to possess complex reasoning abilities and a basic understanding of the real world, addressing challenges that are closer to real-world scenarios.

Image Segmentation and Segment Anything. Image segmentation is a central task in computer vision, aiming to identify and label objects within images at the pixel level. Methods like Fully Convolutional Networks [35], Mask R-CNN [19] and Masf2Former [14] have greatly advanced the field. Recently, the Segment Anything Model (SAM) [25] is trained on the SA-1B [25] dataset with one billion high-quality segmentation annotations. SAM can segment high-quality object masks based on simple prompts. EfficientViT-SAM [8] further introduces multi-scale linear attention into the ViT backbone of SAM, increasing the speed of SAM by several times without compromising performance. Our SAM4MLLM employ MLLMs to guide SAM for precise object segmentation.

Multimodal Large Language Models (MLLMs). LLMs have proven their exceptional capabilities in the domains of language understanding and generation, with notable examples including GPT-3 [17], BLOOM [49], PaLM [15], OPT [60], LLaMA [46], LLaMA-2 [47], Mistral [23], Qwen [6], and others, significantly advancing the field of natural language processing. These models have not only demonstrated near-human levels of proficiency but have also spurred interest in the study of visual-language interaction, leading to the development of MLLMs. MLLMs are built upon LLMs by integrating innovative techniques that combine visual and linguistic modalities, such as the Perceiver Resampler introduced by Flamingo [2], the prompt tuning token by LLaMA-Adapter [58], the Q-Former by BLIP-2 [28], and the use of linear projection layers in LLaVA [33] to enable LLMs to interpret images.

MLLMs Toward Segmentation. In MLLMs, researchers focus not only on enhancing the model's understanding of multimodal data but also empowering MLLMs with the capability to process detailed information. For instance, Det-GPT [38] introduces a method that combines MLLMs with open-vocabulary object detectors. GPT4RoI [59] incorporates region-of-interest information into instructions. Kosmos-2 [37] constructs a large-scale grounding image-text pairs dataset, named GRIT, which assists MLLMs in understanding regional information within images. Shikra [11] encodes all regional information in a linguistic form, eliminating the need for introducing new vocabulary, position encoders, or decoders to MLLMs. Ferret [54] uses a hybrid regional representation method that combines discrete coordinates with continuous features to describe regions within images. However, the model outputs of these methods are limited to bounding boxes and have not yet achieved pixel-level precision operations.

Building on this foundation, Lai *et al.* [27] propose a method based on introducing `[SEG]` tokens and a SAM decoder, enabling MLLMs to perform reasoning segmentation tasks. In PerceptionGPT [39], a lightweight visual task encoder and decoder are adopted to handle segmentation masks, allowing MLLMs to input and output segmentation masks. Ren *et al.* encode masks within images into segmentation codebooks, coupled with a lightweight decoder for mask output. GSVA [50] extends upon LISA [27] by supporting multiple `[SEG]` tokens and introducing `[REJ]` tokens, applying MLLMs to GRES tasks. Rasheed *et al.* [42] propose a Grounding LMM model capable of generating natural language responses seamlessly integrated with corresponding object segmentation masks.

Although the aforementioned models can output masks, they require modifications to the original MLLM architecture or the addition of new model structures to output masks, or the introduction of special tokens not belonging to the original LLMs. They may need to utilize multiple loss functions for simultaneous model optimization, which increases the complexity of MLLM design and poses obstacles to the model's expansion to more tasks. Our SAM4MLLM does not have these burdens, merely integrating with the off-the-shelf SAM model to output high-quality segmentation masks, thereby providing a new solution path for complex pixel-level tasks.

3 Method

In this section, we present our SAM4MLLM method. We first introduce how to encode segmentation masks with SAM's prompt, and then our solutions for prompting SAM using MLLM.

3.1 Encode Segmentation Mask into SAM Prompt

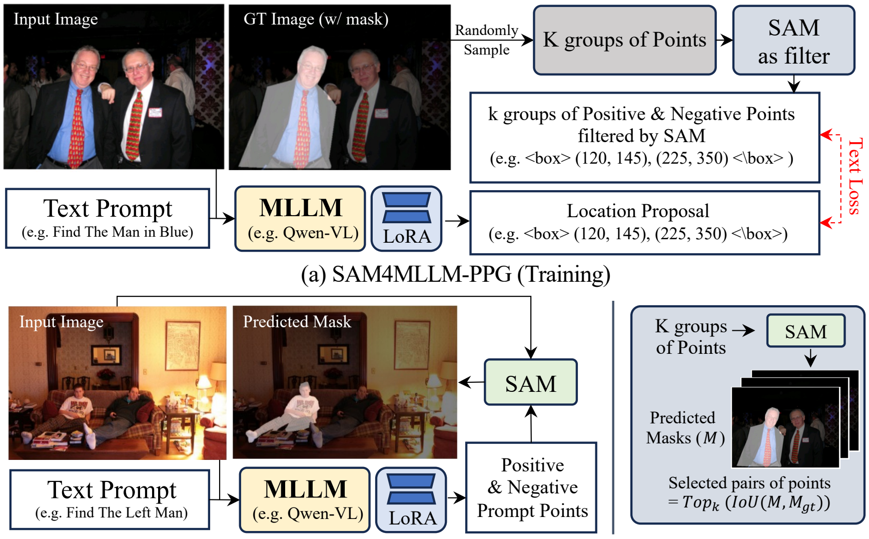
Existing MLLMs for segmentation (e.g., LISA [27], PerceptionGPT [39], GLaMM [42], GSVA [50]) rely on specialized design of model architectures, segmentation-specific tokens, and heterogeneous loss functions to predict object masks. E.g., LISA [27] introduces a special token `[SEG]` and the associated architecture. It uses dice and binary-cross-entropy losses for segmentation, combined with text loss for training. This increases the model complexity and optimization difficulty.

Our method leverages SAM's characteristic that it can convert few discrete text prompt tokens (*i.e.*, bounding box plus several points indicating whether they are inside or outside the object region) to high-quality continuous-boundary segmentation masks. Our **SAM4MLLM** uses the discretized image coordinate for points. We encode an arbitrary-shaped mask by using a bounding box and $\mathcal{K}$ points. The bounding box is expressed as $Prompt_B \in \mathbb{N}^4$; the prompt of $\mathcal{K}$ points, each of which contains three values, x coordinate, y coordinate, and whether the point is on the object mask, are encoded as $Prompt_P \in \mathbb{N}^{\mathcal{K}\times 3}$.

By encoding continuous segmentation masks into discrete SAM prompts, we avoid adding any tokens or altering the model structure, while maintaining training with only text auto-regression cross-entropy loss. This method is consistent with the original training mode of language models, enabling MLLMs to understand pixel-level information and facilitate easier future model expansion.

3.2 Prompting SAM Using MLLM

To incorporate SAM into MLLM in a unified way, a main issue lies in acquiring the prompt points for SAM, including the points that are positive (inside) and negative (outside) the object mask region. To do this, we introduce two solutions, *Prompt-Point Generation* (PPG) and *Proactive Query of Prompt-Points* (PQPP). The former directly generates the proposal points by using the MLLM model in the inference stage. The later, on the other hand, acquires the points in an indirect manner; it uniformly samples the points in the bounding box at first, and then for each point asks the MLLM model whether the point is inside the object region or not. We respectively introduce them in the following.



(a) SAM4MLLM-PPG (Training)

(b) SAM4MLLM-PPG (Inference)

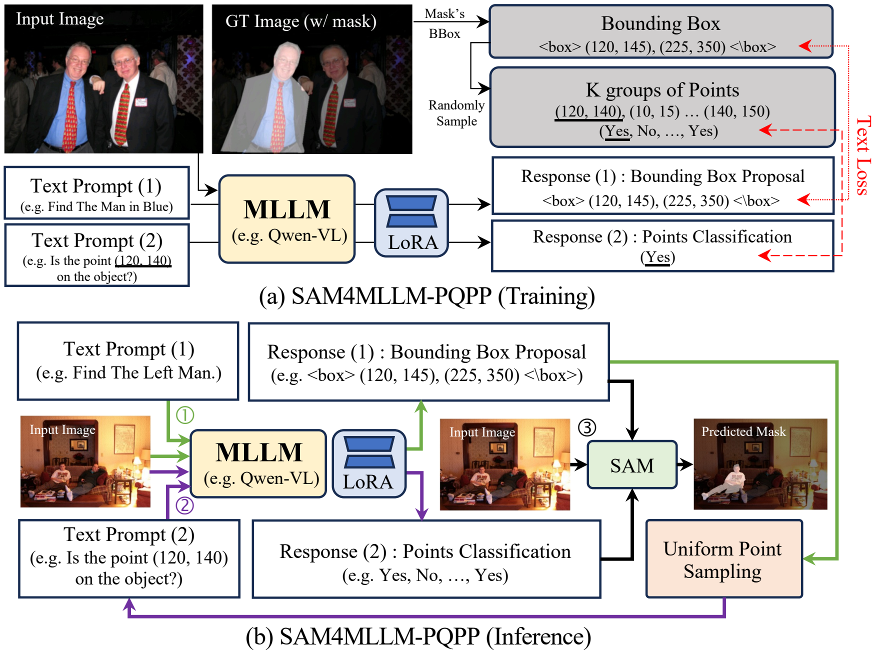
(c) SAM as filter

Fig. 1: Architecture diagram of SAM4MLLM-PPG. (a) The training process of PPG, (b) The inference process of PPG, (c) SAM as filter.

SAM4MLLM-PPG. In this method, an MLLM that can take both text-prompt and image inputs is adopted. To align the MLLM with segmentation utility, we use the parameter-efficient fine tuning technique, LoRA [20], to train the model based on some RES datasets with image-text pairs and ground-truth masks. LoRA outputs the location prompt including the bounding box $Prompt_B \in \mathbb{N}^4$ and k groups of positive and negative points $Prompt_P \in \mathbb{N}^{(n_1+n_2)k\times 3}$, as illustrated in Fig. 1(a), where a group contains n_1 positive and n_2 negative points ($n_1 = 2, n_2 = 1$ in our implementation).

To provide the location supervision to LoRA, we randomly sample K groups of points ($K > k$) in the training stage based on the object mask and then send these prompts to SAM. For every group, SAM delivers the segmentation result. We filter out the prompts with low IoUs compared to the ground-truth masks and only keep the top-k groups (Fig. 1(c)). In our implementation, only text loss (auto-regression cross-entropy loss) is required; K is typically 64 and $k = 1$. In the inference stage, LoRA directly delivers the points that are sent to SAM for segmentation, as shown in Fig. 1(b). More details an be found in Sec. 4.1.

SAM4MLLM-PQPP. In this method, instead of producing the prompts directly, we propose to leverage the power of MLLM's query-response capability. We sample the prompt points and proactively ask the MLLM if they are inside (or outside) the mask. In the training phase, a bounding box and K groups of points are randomly sampled based on the ground-truth mask, and a dialog containing two rounds are conducted. In the first round of the dialog, LoRA responses a bounding box. In the second round, for each of the $(n_1+n_2)K$ points, LoRA responses whether it is inside the mask (yes or no) during training.


Input Image
GT Image (w/ mask)
Mask's BBox
Randomly Sample
Bounding Box
<box> (120, 145), (225, 350) <\box>
K groups of Points
(120, 140), (10, 15) … (140, 150)
(Yes, No, …, Yes)
Text Loss
Text Prompt (1)
(e.g. Find The Man in Blue)
Text Prompt (2)
(e.g. Is the point (120, 140) on the object?)
MLLM
(e.g. Qwen-VL)
LoRA
Response (1) : Bounding Box Proposal
<box> (120, 145), (225, 350) <\box>
Response (2) : Points Classification
(Yes)
Text Prompt (1)
(e.g. Find The Left Man.)
Response (1) : Bounding Box Proposal
(e.g. <box> (120, 145), (225, 350) <\box>)
①
②
③
Input Image
MLLM
(e.g. Qwen-VL)
LoRA
Input Image
SAM
Predicted Mask
Text Prompt (2)
(e.g. Is the point (120, 140) on the object?)
Response (2) : Points Classification
(e.g. Yes, No, …, Yes)
Uniform Point Sampling

(a) SAM4MLLM-PQPP (Training)

(b) SAM4MLLM-PQPP (Inference)

Fig. 2: Architecture diagram of SAM4MLLM-PQPP. (a) The training process of PQPP, (b) The inference process of PQPP.

In the inference phase, LoRA outputs a bounding box in the first round for the input text query and image. We then uniformly sample the points in the bounding box. The sampled grid points are sent to MLLM-LoRA again in the second round and asked if they are positive (or negative); the answers are applied to SAM for segmentation. We typically set the grid size as 5×5. The training and inference phases of SAM4LLM-PQPP are illustrated in Fig. 2. To provide high-quality prompt points before sending to SAM, the low-confident points will be removed. More details of SAM4LLM-PQPP an be found in Sec. 4.1.

Characteristics of the two solutions are as follows. The PPG adopts directly the MLLM to produce prompt points in addition to the bounding box. Yet it would be challenging to learn for simultaneously producing multiple points. Hence, only few prompt points are used in PPG. The PQPP leverages MLLM's dialog capability. It can first inquiry a rough bounding box and then probe multiple points of interest in the bounding box through query-answering for prompting the SAM. We compare their performance in the experiments.

3.3 RES Training

To align the foundational MLLM to the RES task, we use the datasets containing the RES-relevant examples to guide the model toward the goal. Three datasets are used for training our SAM4MLLM to align with the RES task. Two of them (RES dataset and gRefCOCO dataset) contain RES data with ground-truth masks. The third (VQA) is a visual dialog dataset without masks, employed to enhance further the general capability of joint vision-language understanding. During training, to preserve the generalization ability of MLLM on images, we freeze most of the network parameters, and adjust only the visual resampler of MLLM together with the LoRA adapter. The datasets are briefed below.

Referring Expression Segmentation Datasets (RES dataset): Each sample in this dataset provides an image accompanied by a phrase denoting a specific object in the image. The phrase corresponds to only one object. This dataset includes publicly available subsets, refCOCO [57], refCOCO+ [57], and refCOCOg [36]. They are based on images from the MSCOCO [30] but compiled through different annotation processes. The primary difference between RefCOCO+ and RefCOCO is that the former prohibits the use of location-based descriptions (e.g., "the person on the right side of the picture"), thereby compelling annotators to focus on describing the appearance features of objects. RefCOCOg provides longer and more detailed descriptions that cover not only appearance information but may also include actions, locations, and details about relationships with other objects.

Generalized Referring Expression Segmentation (GRES) Dataset [32]: Similar to the RES dataset, each sample offers an image and a phrase describing the objects to segment. The difference is that the phrase may not be present in the image or may refer to multiple objects simultaneously. We use the publicly available gRefCOCO [32] dataset for this task. Our SAM4MLLM can naturally generate additional SAM prompts for segmenting multiple instances. In case where the queried objects are not present, we train our model to predict “object not in the image.”

Visual Question Answering (VQA): To maintain the visual dialogue capability of MLLM, we incorporated VQA data, specifically using the VQAv2 [4].

For all the datasets mentioned above, we do not use data augmentation during training because flipping and/or cropping may change the relative position or relationship of objects in the image.

4 Experiments

In this section, we outline the experimental setups for our SAM4MLLM method, covering the network architecture, implementation details, evaluation datasets, and analysis of experimental results.

4.1 Implementation Details

Network Architecture: We use Qwen-VL-7B-Chat [6] as our MLLM backbone architecture for its ability to output bounding boxes from the pre-training phase. Specifically, the LoRA adapter is configured as follows: LoRA rank is set to 256, LoRA alpha to 128, and LoRA dropout to 0.05. Regarding the SAM [25], we use EfficientViT-XL1-SAM [8] to accelerate the experiments, and we observe only minor accuracy loss in our pilot study.

Training Details: Our training is conducted on 8 NVIDIA 32G V100 GPUs, using float16 precision. We employ deepspeed [3] ZeRO2 for multi-GPU distributed training. We use Lion [12] as ours optimizer. The learning rate is set to 1e−5, with a weight decay of 0.1. We employ a CosineAnnealing learning rate scheduler with a warmup period covering 3% of the total steps. The loss function includes only the text cross entropy loss of LLM. The batch size per GPU is 2, with a gradient accumulation set to 8. We truncate the maximum text length to 2048 during training and only train the model for 3 epochs to prevent overfitting.

Fine-tune SAM light-weight decoder: To ensure fair comparison and optimal performance of our model on the COCO extended dataset, we fine-tune our SAM lightweight mask decoder specifically. Given the inherent bias in mask annotations within the COCO dataset [25], conducting inference directly without fine-tuning SAM's lightweight decoder would not offer a fair comparison against methods that have been trained on COCO with mask decoders. Therefore, we fine-tuned our SAM decoder on the COCO dataset for one epoch. Additionally, to prevent data leakage, we excluded from COCO the images present in the RefCOCO, RefCOCO+, RefCOCOg, and gRefCOCO validation and test sets.

PPG Pointing Strategy Detail: During the training data generation phase of PPG, we randomly sample 64 point groups within the ground truth bounding box. Each group consists of two positive points inside the ground truth mask and one negative point outside. We keep the 16 groups with the highest Intersection over Union (IoU) with the ground truth mask, and then randomly pick a single group from these. The chosen group of points are encoded into text using the proposed "mask as prompt" method to serve as the label for training. During testing, our parser retrieves two positive and one negative points as well.

PQPP Pointing Strategy Detail: To train the PQPP, we randomly sample 10 points from the ground truth bounding box. These points are labeled as positive if they fall inside the ground truth mask, and negative otherwise. During testing, we uniformly sample 5×5 grid of points from the bounding box produced by the MLLM. We then query the MLLM to determine whether each point lies inside the object. Based on MLLM's response (Yes or No), we filter the outcomes according to the response's confidence level associated with the output token (the probability of emitting that token). We only retain points with a confidence level greater than 0.9 and feed them into SAM to generate the mask.

4.2 Benchmarks

We use the datasets described in Sec. 3.3 for training. Their test splits are used for evaluation (**RES dataset**, **GRES**, **VQA**). In addition, we use **ReasonSeg** [27] as a zero-shot evaluation for segmentation from complex reasoning scenarios. This comprehensive evaluation assesses the versatility and effectiveness of our model across a diverse range of referring expression segmentation scenarios.

It is worth mentioning that, compared to other MLLM-based methods, our approach uses significantly less training data. A detailed comparison is presented in Tab. 1. For instance, GLaMM [42] is trained using the GranD [42] dataset,

Table 1: Comparison of the training data from different methods. SAM4MLLM uses less training data than other MLLM-based methods, especially in terms of the number of masks.

Method	**Train set w/ mask**	# img./mask	**Train set w/o mask**
LISA [27]	ADE20K [62], COCO-Stuff [7], PACO-LVIS [41], PartImageNet [18], PASCAL-Part [13], refCLEF [44], refCOCO [57], refCOCO+ [57], refCOCOg [36]	150K/1.2M	LLaVA-Instruct-150k [33]
PerceptionGPT [39]	refCOCO [57], refCOCO+ [57], refCOCOg [36], Visual Genome [26], Flicker30k [55]	150K/3M	MSCOCO-Caption [30], LLaVA-Instruct-150K [33]
GSVA [50]	ADE20K [62], COCO-Stuff [7], PACO-LVIS [41], PartImageNet [18], PASCAL-Part [13], refCLEF [44], refCOCO [57], refCOCO+ [57], refCOCOg [36], gRefCOCO [32]	150K/1.2M	LLaVA-Instruct-150k [33]
GLaMM [42]	GranD [42] (Automatically annotated for SA-1B) , GranD-f [42] (Based on Flickr-30K [55], RefCOCOg [36], and PSG [52])	11M/810M	-
SAM4MLLM (Ours)	refCOCO [57], refCOCO+ [57], refCOCOg [36], gRefCOCO [32]	100K/82K	VQAv2 [4]

which has 11M images and 810M object masks. Its annotations are collected through various vision and language models including GPT-4-based rewrites of existing open-source datasets. In contrast, our SAM4MLLM only uses a small amount of mask annotation data (100K images, 82K object masks) to enable MLLMs to learn general information, but can produce high-quality segmentation masks in conjunction with SAM.

4.3 Main Results

We compare the two variants of our SAM4MLLM, PPG and PQPP, with previous arts on various tasks. There have been numerous LLM-based methods emerging recently, but our comparisons primarily focus on their results using models of similar scales (7B).

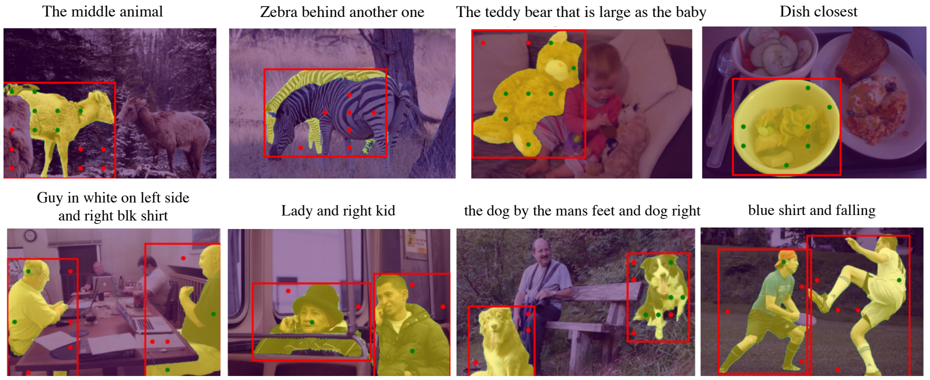
RES dataset: In Tab. 2, we present the performance of PPG and PQPP on the refCOCO datasets [57], where our approach outperforms most of the recent LLM-based methods and achieves comparable results to the most recent GLaMM [42]. Additionally, we observe distinct performance variances across datasets. Specifically, our method shows superior results to GLaMM on the RefCOCOg dataset with complex narrative queries, while we obtain inferior results on the RefCOCO and RefCOCO+ dataset with simple short text queries. This advantage likely arises from our model's streamlined architecture, which preserves the language model's comprehension and inference capabilities more effectively, leading to better results on complex queries.

Table 2: Comparison of methods on refCOCO, refCOCO+, and refCOCOg datasets.

Method	**refCOCO**			**refCOCO+**			**refCOCOg**	
	val	testA	testB	val	testA	testB	val(U)	test(U)
LLM based (13B)								
PerceptionGPT-13B [CVPR-24] [39]	75.3	79.1	72.1	68.9	74.0	61.9	70.7	71.9
GSVA-Llama2-13B [CVPR-24] [50]	79.2	81.7	77.1	70.3	73.8	63.6	75.7	77.0
traditional methods								
MAttNet (CVPR-18) [56]	56.51	62.37	51.70	46.67	52.39	40.08	47.64	48.61
STEP [ICCV-19] [9]	60.04	63.46	57.97	48.19	52.33	40.41	-	-
LSCM [ECCV-20] [22]	61.37	64.99	59.55	49.34	53.12	43.50	-	-
VLT [ICCV-21] [16]	65.65	68.29	62.73	55.50	59.20	49.36	52.99	56.65
SeqTR [ECCV-22] [63]	67.26	69.79	64.12	54.14	58.93	48.19	55.67	55.64
CRIS [CVPR-22] [48]	70.5	73.2	66.1	65.3	68.1	53.7	59.9	60.4
LAVT [CVPR-22] [53]	72.7	75.8	68.8	62.1	68.4	55.1	61.2	62.1
ReLA [CVPR-23] [32]	73.8	76.5	70.2	66.0	71.0	57.7	65.0	66.0
X-Decoder [CVPR-23] [65]	-	-	-	-	-	-	64.6	-
PolyFormer-L [CVPR-23] [34]	76.94	78.49	74.83	72.15	75.71	66.73	71.15	71.17
VPD [ICCV-2023] [61]	73.25	-	-	62.69	-	-	61.96	-
ETRIS [ICCV-2023] [51]	71.06	74.11	66.66	62.23	68.51	52.79	60.28	60.42
SEEM [NeurIPS-23] [66]	-	-	-	-	-	-	65.7	-
LLM based (7B)								
LISA-7B [CVPR-24] [27]	74.9	79.1	72.3	65.1	70.8	58.1	67.9	70.6
PixelLM-7B [CVPR-24] [43]	73.0	76.5	68.2	66.3	71.7	58.3	69.3	70.5
PerceptionGPT-7B [CVPR-24] [39]	75.1	78.6	71.7	68.5	73.9	61.3	70.3	71.7
GSVA-7B [CVPR-24] [50]	77.2	78.9	73.5	65.9	69.6	59.8	72.7	73.3
GLaMM-7B [CVPR-24] [42]	79.5	**83.2**	**76.9**	72.6	**78.7**	64.6	74.2	74.9
SAM4MLLM-7B-PPG	76.2	80.1	72.0	71.2	75.9	64.3	74.2	74.3
SAM4MLLM-7B-PQPP	77.1	80.9	72.5	71.5	76.8	64.7	**74.5**	75.2
SAM4MLLM-7B-PQPP-LLaVA1.6	**79.6**	82.8	76.1	**73.5**	77.8	**65.8**	**74.5**	**75.6**

Notably, our model is trained with nearly 100 times fewer images and 10000 times fewer masks compared to GLaMM, yet it still achieves comparable quality. The results show that by simply leveraging the connection between off-the-shelf MLLMs and SAM, our model can align MLLMs (already trained with large amounts of multimodal data) to the RES task using considerably less data. Since our model training is consistent with that of the original LLM, our method even performs better than all other methods in the case of longer and more complex sentence understanding (RefCOCOg).

AS for our PPG and PQPP approaches, SAM4MLLM-PQPP outperforms SAM4MLLM-PPG in all cases, as shown in Tab. 2. The results reveal that there is a trade-off between cost and quality when using PPG and PQPP. PPG predicts SAM prompts in once and needs only a single-turn conversation with MLLM. PQPP requires a two-turn conversation for bbox prediction and points classification but achieves better accuracy in most cases. We further use LLaVA1.6 as our MLLM backbone architecture instead of Qwen-VL for PQPP on these datasets. With a more powerful MLLM, the performance can be enhanced.



The middle animal

Zebra behind another one

The teddy bear that is large as the baby

Dish closest

Guy in white on left side and right blk shirt

Lady and right kid

the dog by the mans feet and dog right

blue shirt and falling

Fig. 3: Qualitative examples of SAM4MLLM on RES (top) and GRES (bottom) tasks. See Sec. 4.3 for detailed description.

GRES: We present the comparison on the gRefCOCO dataset [32] in Tab. 3. Unlike the RES dataset, this dataset contains the cases where multiple instances or no instances are referred. In this generalized RES task, our method sets the new state-of-the-art among the 7B models on most of the splits and metrics, except for "Test Set A", where we lag slightly behind the recent GSVA [50].

ReasonSeg: Our method also demonstrates superior results on the complex reasoning segmentation task, as shown in Tab. 4. It is worth noting that we evaluate on this dataset in a zero-shot manner, meaning our model was not trained on relevant tasks before.

VQA: This dataset is not desifned for RES but for visual question answering. We use it to verify that our model, although enhanced by image segmentation functionality, still maintains its original capabilities. The VQA scores in Tab. 5 demonstrate that our approach does not compromise the VQA abilities acquired during the pre-training phase of our MLLM backbone. In fact, the VQA performance is even boosted, perhaps due to our fine-tuning on more datasets.

PQPP and PPG: Our PQPP consistently outperforms PPG on most results. We discuss the effect of points prompting strategy further in the ablation studies.

Qualitative results: Fig. 3 presents qualitative examples of our SAM4MLLM approach on various referring expression segmentation datasets. We showcase our results on RES task in the upper row. The leftest image is from refCOCO, showing the successful segmentation of a specific zebra referred to as "behind another one." The middle-left image, sourced from refCOCO+, demonstrates the accurate identification of the "middle animal" among multiple instances. The middle-right image from refCOCOg illustrates the model's ability to handle more complex referring expressions, such as "The teddy bear that is as large as the baby." Finally, the rightest image, also from refCOCO+, showcases the model's understanding of relative positions, correctly segmenting the "dish closest" to the referred object. The bottom row demonstrates SAM4MLLM ability on generalized RES task, where our method accurately segments multiple instances as per

Table 3: Results on gRefCOCO(GRES).

Method	**Validation Set**			**Test Set A**			**Test Set B**		
	gIoU	cIoU	N-acc.	gIoU	cIoU	N-acc.	gIoU	cIoU	N-acc.
LLM-based (13B) model									
LISA-13B [CVPR-24] [27]	65.24	63.96	57.49	69.99	71.00	55.43	62.11	62.29	56.34
GSVA-13B [CVPR-24] [50]	70.04	66.38	66.02	73.29	72.79	64.72	65.45	63.20	62.47
traditional methods									
MAttNet [CVPR-18] [56]	48.24	47.51	41.15	59.30	58.66	44.04	46.14	45.33	41.32
LTS [CVPR-21] [24]	52.70	52.30	-	62.64	61.87	-	50.42	49.96	-
VLT [ICCV-21] [16]	52.00	52.51	47.17	63.20	62.19	48.74	50.88	50.52	47.82
CRIS [CVPR-22] [48]	56.27	55.34	-	63.42	63.82	-	51.79	51.04	-
LAVT [CVPR-22] [53]	58.40	57.64	49.32	65.90	65.32	49.25	55.83	55.04	48.46
ReLA [CVPR-23] [32]	63.60	62.42	56.37	70.03	69.26	59.02	61.02	59.88	58.40
LLM based (7B)									
LISA-7B [CVPR-24] [27]	61.63	61.76	54.67	66.27	68.50	50.01	58.84	60.63	51.91
GSVA-7B [CVPR-24] [50]	66.47	63.29	62.43	**71.08**	69.93	65.31	62.23	60.47	60.56
SAM4MLLM-7B-PPG	68.37	65.66	**63.71**	69.05	69.62	**65.96**	63.71	62.35	61.25
SAM4MLLM-7B-PQPP	**68.96**	**66.33**	62.96	70.54	**70.13**	65.82	**63.98**	**63.21**	**61.61**

the given text. These examples highlight SAM4MLLM's capability to accurately segment objects based on diverse referring expressions across different datasets.

5 Ablation Study

To gain a deeper understanding of the factors contribution, we conducted some ablation studies focusing on the best-performing variant, SAM4MLLM-PQPP. Our investigations centered around the following aspects: confidence threshold for filtering points in PQPP and sampling strategy for selecting points within the bounding box, providing insights into our method's robustness and adaptability.

Points Filtering Threshold First, we examined the impact of the confidence threshold used in PQPP to filter points based on the MLLM's responses. We experimented with threshold values ranging from 0.6 to 0.95 on the RefCOCOg validation set and evaluated their effect on the cIoU metric. Our results, presented in Table 6a, reveal that a threshold of 0.9 strikes the optimal balance, as further increasing or decreasing the threshold leads to a notable decline in cIoU. This finding highlights the importance of carefully tuning the confidence threshold to ensure the best possible segmentation quality.

Points Sampling Strategy Next, we explored the influence of the point sampling strategy within the bounding box on the overall performance. We compared two approaches: grid-based sampling and random sampling, while also varying the number of sampled points. As shown in Table 6b, a 5x5 grid sampling pattern consistently yields the highest accuracy. The result suggests that a uniform distribution of points across the bounding box provides the most informative cues for the MLLM to accurately determine the object's location and shape.

Table 4: Results on ReasonSeg(Zero-Shot).

Method	val	
	gIoU	cIoU
LISA-13B [CVPR-24] [27]	48.9	46.9
OVSeg [CVPR-23] [29]	28.5	18.6
GRES [CVPR-23] [32]	22.4	19.9
X-Decoder [CVPR-23] [65]	22.6	17.9
SEEM [NeurIPS-23] [66]	25.5	21.2
LISA-7B [CVPR-24] [27]	44.4	46.0
SAM4MLLM-7B-PPG	46.1	47.9
SAM4MLLM-7B-PQPP	**46.7**	**48.1**

Table 5: Result on Vision Question Answering

Method	VQAv2
Qwen-VL-Chat-7B [6]	78.2
SAM4MLLM-7B-PPG	**78.7**
SAM4MLLM-7B-PQPP	**78.7**

Table 6: Ablation studies of SAM4MLLM-PQPP.

(a) Effect of point filter threshold.

Method	RefCOCOg val
0.6	69.5
0.7	70.8
0.8	72.2
0.9	**74.5**
0.95	73.8

(b) Effect of point sampling strategy in inference.

Method	RefCOCOg val
5x5 grid (N=25)	**74.5**
6x6 grid (N=36)	74.4
random sample (N=25)	73.7
random sample (N=36)	74.2

6 Conclusion

In this paper, we introduced SAM4MLLM, a novel approach that integrates the Segment Anything Model (SAM) with Multi-Modal Large Language Models (MLLMs) to address the Referring Expression Segmentation (RES) task. By encoding object masks as discrete text prompts, our method enables MLLMs to understand and generate pixel-level object localization information without requiring complex architectural modifications or additional loss functions. Our method is simple but effective. Through experiments on various RES benchmarks, we demonstrate that SAM4MLLM achieves competitive performance while maintaining the simplicity and generalizability of the original language models. Our work explores a new direction for leveraging the capabilities of foundation models to tackle complex vision-language tasks in a more streamlined and unified manner. We hope that the insights gained from this research will inspire further investigations into effectively combining the strengths of different models to solve challenging multimodal problems. Future work could involve extending our approach to handle a broader range of visual reasoning tasks and conducting more in-depth analyses to better understand the interplay between language models and visual foundation models.

Acknowledgement This work was supported in part under grants NSTC 112-2634-F-002-005 and NSTC112-2634-F-006-002, NTU under grants 113L900902, and Nvidia-NTU collaboration project under grants 113HT912004.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. NeurIPS (2022)
3. Aminabadi, R.Y., Rajbhandari, S., Awan, A.A., Li, C., Li, D., Zheng, E., Ruwase, O., Smith, S., Zhang, M., Rasley, J., et al.: Deepspeed-inference: enabling efficient inference of transformer models at unprecedented scale. In: SC (2022)
4. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: ICCV (2015)
5. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
6. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966 (2023)
7. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: CVPR (2018)
8. Cai, H., Gan, C., Han, S.: Efficientvit: Enhanced linear attention for high-resolution low-computation visual recognition. ECCV (2023)
9. Chen, D.J., Jia, S., Lo, Y.C., Chen, H.T., Liu, T.L.: See-through-text grouping for referring image segmentation. In: ICCV (2019)
10. Chen, J., Zhu, D., Shen, X., Li, X., Liu, Z., Zhang, P., Krishnamoorthi, R., Chandra, V., Xiong, Y., Elhoseiny, M.: Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. arXiv preprint arXiv:2310.09478 (2023)
11. Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., Zhao, R.: Shikra: Unleashing multimodal llm's referential dialogue magic. arXiv preprint arXiv:2306.15195 (2023)
12. Chen, X., Liang, C., Huang, D., Real, E., Wang, K., Pham, H., Dong, X., Luong, T., Hsieh, C.J., Lu, Y., et al.: Symbolic discovery of optimization algorithms. NeurIPS (2024)
13. Chen, X., Mottaghi, R., Liu, X., Fidler, S., Urtasun, R., Yuille, A.: Detect what you can: Detecting and representing objects using holistic models and body parts. In: CVPR (2014)
14. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: CVPR (2022)
15. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al.: Palm: Scaling language modeling with pathways. Journal of Machine Learning Research (2023)
16. Ding, H., Liu, C., Wang, S., Jiang, X.: Vision-language transformer and query generation for referring segmentation. In: ICCV (2021)
17. Floridi, L., Chiriatti, M.: Gpt-3: Its nature, scope, limits, and consequences. Minds and Machines (2020)
18. He, J., Yang, S., Yang, S., Kortylewski, A., Yuan, X., Chen, J.N., Liu, S., Yang, C., Yu, Q., Yuille, A.: Partimagenet: A large, high-quality dataset of parts. In: ECCV (2022)
19. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV (2017)

20. Hu, E.J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. In: ICLR (2021)
21. Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O.K., Patra, B., et al.: Language is not all you need: Aligning perception with language models. NeurIPS (2023)
22. Hui, T., Liu, S., Huang, S., Li, G., Yu, S., Zhang, F., Han, J.: Linguistic structure guided context modeling for referring image segmentation. In: ECCV (2020)
23. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al.: Mistral 7b. arXiv preprint arXiv:2310.06825 (2023)
24. Jing, Y., Kong, T., Wang, W., Wang, L., Li, L., Tan, T.: Locate then segment: A strong pipeline for referring image segmentation. In: CVPR (2021)
25. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. ICCV (2023)
26. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. IJCV (2017)
27. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. CVPR (2024)
28. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. ICML (2023)
29. Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D.: Open-vocabulary semantic segmentation with mask-adapted clip. In: CVPR (2023)
30. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
31. Lin, Z., Liu, C., Zhang, R., Gao, P., Qiu, L., Xiao, H., Qiu, H., Lin, C., Shao, W., Chen, K., et al.: Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. arXiv preprint arXiv:2311.07575 (2023)
32. Liu, C., Ding, H., Jiang, X.: Gres: Generalized referring expression segmentation. In: CVPR (2023)
33. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. NeurIPS (2024)
34. Liu, J., Ding, H., Cai, Z., Zhang, Y., Satzoda, R.K., Mahadevan, V., Manmatha, R.: Polyformer: Referring image segmentation as sequential polygon generation. In: CVPR (2023)
35. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015)
36. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: CVPR (2016)
37. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding multimodal large language models to the world. ICLR (2024)
38. Pi, R., Gao, J., Diao, S., Pan, R., Dong, H., Zhang, J., Yao, L., Han, J., Xu, H., Zhang, L.K.T.: Detgpt: Detect what you need via reasoning. ACL (2023)
39. Pi, R., Yao, L., Gao, J., Zhang, J., Zhang, T.: Perceptiongpt: Effectively fusing visual perception into llm. CVPR (2024)
40. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
41. Ramanathan, V., Kalia, A., Petrovic, V., Wen, Y., Zheng, B., Guo, B., Wang, R., Marquez, A., Kovvuri, R., Kadian, A., et al.: Paco: Parts and attributes of common objects. In: CVPR (2023)

42. Rasheed, H., Maaz, M., Shaji, S., Shaker, A., Khan, S., Cholakkal, H., Anwer, R.M., Xing, E., Yang, M.H., Khan, F.S.: Glamm: Pixel grounding large multimodal model. CVPR (2024)
43. Ren, Z., Huang, Z., Wei, Y., Zhao, Y., Fu, D., Feng, J., Jin, X.: Pixellm: Pixel reasoning with large multimodal model. CVPR (2024)
44. Rohrbach, A., Rohrbach, M., Hu, R., Darrell, T., Schiele, B.: Grounding of textual phrases in images by reconstruction. In: ECCV (2016)
45. Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
46. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
47. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
48. Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M., Liu, T.: Cris: Clip-driven referring image segmentation. In: CVPR (2022)
49. Workshop, B., Scao, T.L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A.S., Yvon, F., et al.: Bloom: A 176b-parameter open-access multilingual language model. arXiv preprint arXiv:2211.05100 (2022)
50. Xia, Z., Han, D., Han, Y., Pan, X., Song, S., Huang, G.: Gsva: Generalized segmentation via multimodal large language models. CVPR (2024)
51. Xu, Z., Chen, Z., Zhang, Y., Song, Y., Wan, X., Li, G.: Bridging vision and language encoders: Parameter-efficient tuning for referring image segmentation. In: ICCV (2023)
52. Yang, J., Ang, Y.Z., Guo, Z., Zhou, K., Zhang, W., Liu, Z.: Panoptic scene graph generation. In: ECCV (2022)
53. Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., Torr, P.H.: Lavt: Language-aware vision transformer for referring image segmentation. In: CVPR (2022)
54. You, H., Zhang, H., Gan, Z., Du, X., Zhang, B., Wang, Z., Cao, L., Chang, S.F., Yang, Y.: Ferret: Refer and ground anything anywhere at any granularity. ICLR (2024)
55. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics (2014)
56. Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., Berg, T.L.: Mattnet: Modular attention network for referring expression comprehension. In: CVPR (2018)
57. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: ECCV (2016)
58. Zhang, R., Han, J., Zhou, A., Hu, X., Yan, S., Lu, P., Li, H., Gao, P., Qiao, Y.: Llama-adapter: Efficient fine-tuning of language models with zero-init attention. ICLR (2023)
59. Zhang, S., Sun, P., Chen, S., Xiao, M., Shao, W., Zhang, W., Chen, K., Luo, P.: Gpt4roi: Instruction tuning large language model on region-of-interest. arXiv preprint arXiv:2307.03601 (2023)
60. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., et al.: Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068 (2022)
61. Zhao, W., Rao, Y., Liu, Z., Liu, B., Zhou, J., Lu, J.: Unleashing text-to-image diffusion models for visual perception. ICCV (2023)

62. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. ICCV (2019)
63. Zhu, C., Zhou, Y., Shen, Y., Luo, G., Pan, X., Lin, M., Chen, C., Cao, L., Sun, X., Ji, R.: Seqtr: A simple yet universal network for visual grounding. In: ECCV (2022)
64. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)
65. Zou, X., Dou, Z.Y., Yang, J., Gan, Z., Li, L., Li, C., Dai, X., Behl, H., Wang, J., Yuan, L., et al.: Generalized decoding for pixel, image, and language. In: CVPR (2023)
66. Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Wang, J., Wang, L., Gao, J., Lee, Y.J.: Segment everything everywhere all at once. NeurIPS (2024)