arXiv:2403.14141v1 [cs.CV] 21 Mar 2024

Empowering Segmentation Ability to Multi-modal Large Language Models

Yuqi Yang[*] Peng-Tao Jiang[*] Jing Wang[*] Hao Zhang Kai Zhao Jinwei Chen Bo Li
vivo Mobile Communication Co., Ltd.

Abstract

Multi-modal large language models (MLLMs) can understand image-language prompts and demonstrate impressive reasoning ability. In this paper, we extend MLLMs' output by empowering MLLMs with the segmentation ability. The extended MLLMs can both output language responses to the image-language prompts and segment the regions that the complex question or query in the language prompts focuses on. To this end, the existing work, LISA, enlarges the original word embeddings with an additional segment token and fine-tunes dialogue generation and query-focused segmentation together, where the feature of the segment token is used to prompt the segment-anything model. Although they achieve superior segmentation performance, we observe that the dialogue ability decreases by a large margin compared to the original MLLMs. To maintain the original MLLMs' dialogue ability, we propose a novel MLLMs framework, coined as LLaVASeg, which leverages a chain-of-thought prompting strategy to instruct the MLLMs to segment the target region queried by the user. The MLLMs are first prompted to reason about the simple description of the target region from the complicated user query, then extract the visual attributes of the target region according to the understanding of MLLMs to the image. These visual attributes, such as color and relative locations, are utilized to prompt the downstream segmentation model. Experiments show that the proposed method keeps the original dialogue ability and equips the MLLMs' model with strong reasoning segmentation ability. The code is available at `https://github.com/YuqiYang213/LLaVASeg`.

1. Introduction

Large Language Models (LLMs) show robust dialogue and reasoning skills when scaling up the data and model size in Natural Language Processing (NLP). There emerge massive chatbots based on LLMs [7, 13, 35, 36], such as OpenAI's chatGPT [28], which take language prompts as input and output language responses. Since the human perception system can process multiple modal information, such as visual and audio information, LLMs hold the potential to be extended to multi-modal large language models with multi-modal input. Thus, prompted by the popular LLMs, Multi-modal Large Language Models (MLLMs) have also been widely investigated, which receive not only language but also images or audio as input. There also emerge many extraordinary MLLMs, such as GPT-4V [29], Flamingo [1], BLIP [18], LLaVA [4], and PaLM [7], etc.

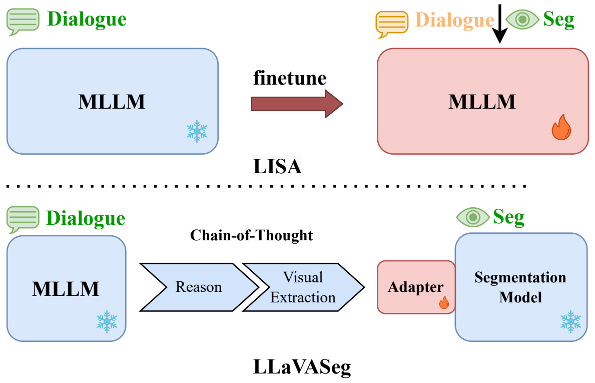


Figure 1. Comparison of previous fine-tuning method [17] and our method. Though previous works empowered the MLLMs with segmentation ability, they downgraded the MLLMs' original reasoning ability. On the contrary, our method not only equips the MLLMs with segmentation ability but maintains the MLLMs' original reasoning ability by freezing parameters.

MLLMs act like a person, able to read, see, listen, and then generate language responses. Associating with the human perception system, humans can not only understand multi-modal information but also locate the language-focused target objects or areas quickly in the physical world. Thus, some pioneer works [17, 31, 38] extend the capability of MLLMs to different visual tasks with the help of instruction fine-tuning. For instance, Shikra [3] encodes the coordinates of the target areas into language and uses them to construct instruction-following data with the target location in the output. Recently, LISA [17] enlarges MLLMs' word embeddings with a <SEG> token. The output feature of the

---

[*] The first three authors contributed equally to this paper. Peng-Tao Jiang is the corresponding author.

<SEG> token will prompt the decoder of the segment anything model (SAM) [15] to segment target areas. However, we observe that during fine-tuning MLLMs, the MLLM's dialogue ability is heavily affected compared with the original MLLMs. Although LISA introduced instructions from VQA datasets to maintain the dialogue ability, it cannot completely address this problem and the degradation is still significant. In addition, when prompting the LISA model to output both the segmentation mask and the explanations, the quality of the segmentation mask also decreased by a large margin.

In this paper, we aim to empower the MLLMs with segmentation abilities without damaging their original reasoning ability. To this end, one straightforward solution is to prompt the MLLMs to output the name of the target area. The segmentation model takes the output name and image as input and segments the target referred to by its name. However, since the open-world target name usually doesn't contain any visual information, it is challenging for the segmentation model to directly align the target name with a specific image region [48]. Furthermore, this solution does not fully utilize the rich knowledge of the MLLMs to facilitate the model locating the target more precisely.

To address this issue, we prompt the MLLMs to generate both the abstract target name and detailed image-specific visual attributes. The visual attributes include the shape, the color, and the relative location of the target. Since the visual attributes can provide further visual information than the target name, it can help the segmentation model target objects and refine the segmentation mask. To generate the visual attributes of the target object from the complicated user query, we propose a chain-of-thought prompting to explicitly generate the detailed visual attributes, which can be better understood by the segmentation model. The chain-of-thought prompting includes three steps. In the first step, we prompt the MLLMs to generate responses for understanding the user's query and finding the target from the image. In this step, the MLLMs tend to generate responses containing lengthy explanations of the reasoning step. In the second step, the MLLMs are prompted to extract the target name from lengthy explanations and describe it in the shortest possible way. In the third step, we prompt the MLLMs to generate visual attributes (color, shape, *etc*.) of the target object and its relative location in the image. These discriminative attributes are sent to the downstream segmentation model and instruct the segmentation model to generate the final segmentation mask. We propose multi-scale adapters in the segmentation model and fine-tune them to fuse the extracted attributes with the visual features.

Following the above chain-of-thought prompting paradigm, our method, termed LLaVASeg, can empower the off-the-shelf MLLMs, such as **LLaVA**, with **Seg**mentation ability while not damaging their original reasoning ability. We show the comparison of our LLaVASeg and fine-tuning-based method in Fig. 1. Experiments show that LLaVASeg achieves superior performance in both segmentation and dialogue, which proves the effectiveness of our method. Unlike LISA [17] utilize instruction fine-tuning paradigm, our LLaVASeg uses chain-of-thought prompting to prompt the off-the-shelf MLLMs and performs comparable segmentation performance. We hope our work can show more insights into the rich capabilities of MLLMs.

In summary, the contributions of this paper are two-fold:

- We propose a novel framework that not only equips the MLLMs with segmentation ability but also maintains the MLLMs' original powerful conversational ability. Our method is MLLMs-agnostic and the plug-and-play advantage supports its application to different MLLMs.
- We propose a novel chain-of-thought prompting strategy that iteratively prompts MLLMs to generate suitable responses for downstream segmentation models. The chain-of-thought prompting strategy bridges the gap between reasoning and segmentation without additional MLLMs fine-tuning.

2. Related Work

2.1. Multi-Modal Large Language Model

The large language models [6, 8, 13, 30, 35, 36] have exhibited impressive language understanding ability. To extend this ability to the domain of vision, many works [1, 4, 18, 29, 47] introduce auxiliary vision models to deal with additional visual input. For example, LLaVA [22], based on LLaMA [35], aligns the visual features and language features with instruction tuning. Recently, some work [3, 17, 24, 31] attempted to enable MLLMs with more vision tasks like detection and segmentation. Shikra [3] fine-tunes MLLMs on specific instruction-following conversations of which all the specific objects are equipped with the coordinates of the bounding boxes. LISA [17] broadens the word embedding with an additional <SEG> token and fine-tunes the multimodal large language models on the instruction-following segmentation data and LLaVA's instruction tuning data. Different from the above methods, our method does not fine-tune the MLLMs and utilizes the chain-of-thought prompting strategy to extract rich information about the query-focused region. The rich information is used to prompt the downstream segmentation models.

2.2. Chain-of-Thought

LLMs have shown impressive reasoning ability. Recently, some methods [16, 41] tend to improve the reasoning ability by prompting LLMs to think step by step. This prompting strategy, termed chain-of-thought prompting, further explores the potential of LLMs. Boosting with chain-of-thought prompting, LLMs can optimize the reasoning process by decomposing the hard task explicitly [14, 32, 46] or calibrating the results with different chain-of-thoughts [11, 39]. Some methods [34, 43, 44] extend the chain-of-thought prompting to the vision domain. Among them, CoTDeT [34] leverages the chain-of-thought prompting to explore the affordance required by specific tasks to benefit object detection. Different from these methods, our method proposes to leverage the chain-of-thought prompting to prompt the MLLMs to generate the visual attributes, such as the shape, the color and the relative location of the query-focused image region. The visual attributes can provide rich visual information about the target, facilitating the segmentation model to locate target regions and refine the segmentation mask.

2.3. Referring Segmentation

The referring segmentation aims to segment the objects or regions pointed out by the text description. A typical paradigm for referring segmentation methods [9, 10, 42] is combining the text features into the visual features to assist in segmenting the target objects. Specifically, VLT [9] and EFN [10] fuse visual and text features in the decoder stage, while LAVT [42] choose to combine the text and visual features at the encoder stage and obtain better performance. However, most existing referring segmentation datasets [12] rely on a simple description, resulting in many referring segmentation methods lacking reasoning and analyzing ability. Our method, with the help of MLLMs, performs a better understanding of complex queries that require common knowledge and understanding of the world.

3. Method

In this section, we will introduce our method in detail. The framework of the proposed LLaVASeg is shown in Fig. 2. First, we will introduce the problem definition and the pipeline of MLLMs in Sec. 3.1. Subsequently, we will dive into the details of our chain-of-thought prompting design in Sec. 3.2. Then, we will introduce the prompting segmentation network in Sec. 3.3. Finally, we discuss our data construction pipeline in Sec. 3.4.

3.1. Preliminaries

Reasoning Segmentation is first proposed in [17], which aims to parse the question-oriented or query-focused visual region in an image and output a segmentation mask to locate it. Although this task shares similarities with referring segmentation task [12], it requires more knowledge about the real world and common sense to reason about the target region from intricate expression.

The Pipeline of MLLMs The MLLMs take multi-modal information as input that includes an image $\mathbf{I} \in \mathcal{R}^{H\times W\times 3}$ and text $\mathcal{T}$ of random length. The image $\mathbf{I}$ is encoded with the multi-modal encoder v to share the same embedding space with the text embedding [4]. The text can be further divided into two parts. The first part is the user query $\mathcal{T}_q$, which is related to the requirements of the user. The second part is the pre-designed task prompt $\mathcal{T}_p$, which is hand-crafted or generated automatically. Although the prompt is not given by the user, a proper prompt can instruct the MLLMs to generate a response with higher quality [16]. Formally, the MLLMs f_l generate the answer $\mathcal{T}_a$ by

$$\mathcal{T}_a = f_l(v(\mathbf{I}), \{\mathcal{T}_p, \mathcal{T}_q\}), \tag{1}$$

where $\{\cdot\}$ indicates the concatenation of text input. The MLLMs also support multi-turn conversation by concatenating the former input and response with another query and prompt, which can be formulated as:

$$\mathcal{T}_a' = f_l(v(\mathbf{I}), \{\{\mathcal{T}_{p_1}, \mathcal{T}_{q_1}, \mathcal{T}_a\}, \mathcal{T}_{p_2}, \mathcal{T}_{q_2}\}), \tag{2}$$

where $\mathcal{T}_{p_i}$ and $\mathcal{T}_{q_i}$ indicate the task prompt and user query in the i-th turn.

Recently, LISA [17] adapted the MLLMs to the segmentation task by enlarging the word embeddings with a <SEG> token. The features of the <SEG> token are used to prompt the downstream segmentation model. However, the instruction tuning for the segmentation task heavily affects the conversational and reasoning ability of the original MLLMs. In this paper, instead of fine-tuning the MLLMs, we attempt to change the prompt $\mathcal{T}_p$ to generate responses containing the name and the visual attributes of the target. The visual attributes include the shape, color and relative location of the target, which provides additional image-specific information for the target. The prompt-based method can maintain the conversation and reasoning ability of MLLMs.

3.2. Chain-of-Thought Prompting Design

To generate the visual attributes, knowledge from both the real world and the given image is needed to link the complicated user query to the target area in the image. When naively prompting the MLLMs to reason from the user query, the MLLMs may only generate the rationales or explanations in response to the query, not the needed visual attributes. To address this problem, we propose to leverage chain-of-thought prompting, which explicitly prompts MLLMs to generate the visual attributes of query-focused regions step by step. Our chain-of-thought prompting includes three steps: reason prompting step, target prompting step and attribute prompting step. In the first step, we prompt the MLLMs to use common knowledge to find the target area. Since the response in the first step usually includes many unrelated explanations, we then prompt the MLLMs to find the exact target name and get rid of unrelated information in the second step. In the third step, we prompt the MLLMs to extract the visual attributes of the target area from the image. We will dive into the details of these steps in the rest of this section.

Reason Prompting Step: The first step is to prompt MLLMs to reason about the target region in the image from

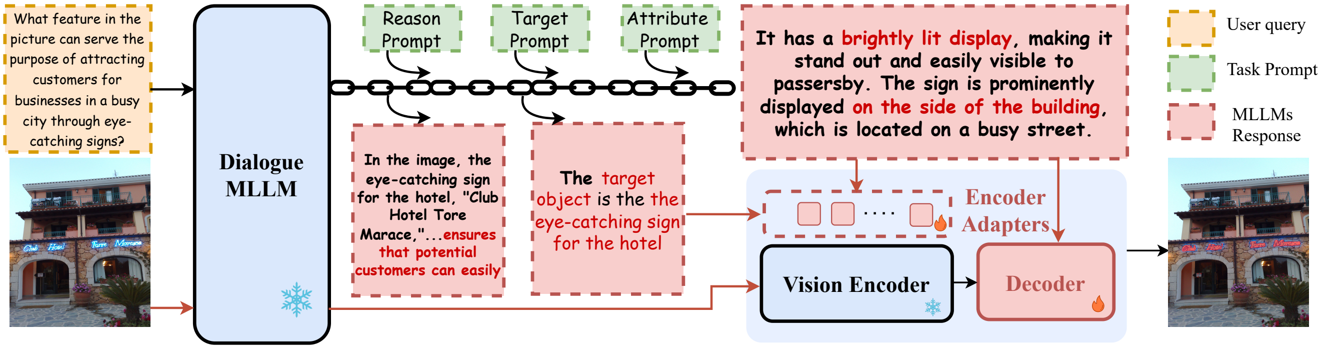

What feature in the picture can serve the purpose of attracting customers for businesses in a busy city through eye-catching signs?
Dialogue MLLM
Reason Prompt
Target Prompt
Attribute Prompt
In the image, the eye-catching sign for the hotel, "Club Hotel Tore Marace,"...ensures that potential customers can easily
The target object is the the eye-catching sign for the hotel
It has a brightly lit display, making it stand out and easily visible to passersby. The sign is prominently displayed on the side of the building, which is located on a busy street.
Encoder Adapters
Vision Encoder
Decoder
User query
Task Prompt
MLLMs Response

Figure 2. Overall pipeline of our LLaVASeg. Given the user query and image, the MLLMs generate the visual attributes through chain-of-thought prompting. Then, we input these visual attributes into the segmentation model to perform multi-scale prompting and generate the segmentation mask. During training, the MLLMs and the segmentation model are frozen, and only the lightweight adapters are trainable to guarantee efficiency.

the user query. The reasoning step can be formulated as a VQA task, so we rephrase the user query into a question for a specific object or area in this image. Concretely, we design the following prompt for this step:

Prompt: What is the object or part that is [USER QUERY] in this image?
Output: It is fire in the fire pit. The fire is hot and ...

[USER QUERY] is filled with the user query. When the user query is a question, we directly use it as the input and do not add other prompts. Notice that we do not instruct the MLLMs to output the rationale explicitly. We find that the MLLMs tend to answer the question followed by the explanation, although no explicit prompt is given. Additionally, forcing the model to output the rationale while not providing any knowledge for it may result in severe hallucination [43]. Given these concerns above, we leave the reason prompt as a question and do not prompt the MLLMs to give a detailed explanation explicitly.

Target Prompting Step: The MLLMs output both the name of the target region and the explanation including many unrelated descriptions (e.g. hot, light...) in the first step. These descriptions may disturb the MLLMs to generate exact visual attributions in the attribute prompting step, and it is not trivial to find the exact target from the response. In the second step, we input the user query, response, and image to the MLLMs to provide sufficient information so that the MLLMs can analyze the target. Specifically, the target prompt is formulated as follows:

Prompt: Here is the conversation:
The question is: [QUESTION]
The answer is: [ANSWER]
Please analyze the conversation and identify the distinct physical objects or areas that the user wants from the image.
Output: The user wants the fire from the image.

[QUESTION] and [ANSWER] is the question and answer in the first step, respectively.

Attribute Prompting Step: In this step, we prompt the MLLMs to provide visual information towards the target. Although it seems straightforward to prompt the MLLMs to output the location of the target object in the image, previous work [3] shows that the MLLMs cannot output accurate spatial positions in zero-shot settings. To address this issue, we choose to extract the visual attributes rather than the specific location to aid the downstream segmentation model. The visual attributes include color, shape, and the relative position to the other objects in the image. We prompt the MLLMs to extract the visual attributes as follows:

Prompt: Here is the target:
The target is: [TARGET]
Briefly describe the target entity or part's visual attributes that can discriminate them from the image. Each visual attribute can be color, shape, and relative position to other objects in the image.
Output: The fire can be discriminated from the image by its bright orange color and the fact that it is emitting heat and light. The fire is surrounded by ...

[TARGET] is the target analyzed in the second step. For simplicity, we present the most critical parts of the prompts. Note that we prompt the MLLMs to generate the discriminative visual attributes which are more helpful in segmenting the target region from the image.

In the implementation, the target prompting step and the attribute prompting step are performed in the same round of conversation for efficiency. Specifically, we prompt the MLLMs to first target the region and then extract its visual attributes explicitly as follows: *Follow these guidelines strictly: (1) Please analyze the conversation... (2) Briefly describe the target entity or...*. We empirically find that this can save inference time without an apparent performance drop.

3.3. Multi-Scale Prompting Segmentation

Since the extracted visual attributes can serve in both low-level (e.g. color) and high-level (e.g. name, relative location), we propose a multi-scale promptable segmentation pipeline to align the visual attributes and their corresponding area. In this section, we introduce the architecture and the training objectives of our segmentation pipeline.

Segmentation Pipeline: We input the target name with the visual attributes $\mathbf{E}$ and the image $\mathbf{I}$ into the segmentation model f_s to generate the final mask $\mathbf{M}$, which can be formulated as:

$$\mathbf{M} = f_s(\mathbf{I}, \mathbf{E}). \tag{3}$$

In our method, we directly use MLLMs' token embeddings corresponding to the target name and visual attributes and don't utilize extra text encoders like BERT [13]. We add an MLP to project these textual embeddings to the visual feature space. Given the projected textual embeddings $\mathbf{E}$, to perform multi-scale interaction between $\mathbf{E}$ and the visual features in the segmentation model, we introduce several language-aware adapters in different layers. Motivated by [42], we design our language-aware module based on the attention mechanism. Specifically, we leverage the cross-attention between the textual embedding and the visual features from the segmentation model to inject the visual attributes into the segmentation model.

Formally, we pick up several layers in the segmentation backbone to set the language-aware modules. When the language-aware module is set in p-th layer, the visual feature $\mathbf{V}_p$ will be fed into the language-aware module together with language embedding $\mathbf{E}$ to generate the fused feature $\mathbf{F}_p$, which can be formulated as:

$$\mathbf{F}_p = \mathcal{LAM}_p(\mathbf{V}_p, \mathbf{E}). \tag{4}$$

Each $\mathcal{LAM}_p$ mainly consists of a cross-attention module and a feed-forward layer. It takes $\mathbf{V}_p$ as query and $\mathbf{E}$ as key and value. After the interaction, the fused feature will be added by the original features $\mathbf{V}_p$ through a residual path.

For further interaction between visual features and textual features $\mathbf{E}$ in the decoding phrase, we take inspiration from SAM [15] and leverage the prompt-based decoder in our method. The decoder uses cross-attention in both directions, which is prompt-to-vision embedding and vice-versa, as its main components.

Training Objectives: During training, we freeze all the parameters of the segmentation model except for the adapter and decoder. Different from LISA [17], our model is only trained with the segmentation loss $\mathcal{L}_{mask}$. Specifically, our loss $\mathcal{L}_{mask}$ is a combination of per-pixel binary cross-entropy (BCE) loss and DICE loss, which can be formulated as

$$\mathcal{L}_{mask} = \lambda_{bce}\mathbf{BCE}(\hat{\mathbf{M}}, \mathbf{M}) + \lambda_{dice}\mathbf{DICE}(\hat{\mathbf{M}}, \mathbf{M}), \tag{5}$$

where $\hat{\mathbf{M}}$ is the ground-truth and $\mathbf{M}$ is the prediction. λ_{bce} and λ_{dice} denote the loss weights.

3.4. Training Data Construction

To train LLaVASeg, we collect two kinds of training data. The first one is the segmentation dataset with complicated user queries, including ReasonSeg [17]. We also use its validation set to evaluate the performance of our method. However, ReasonSeg only contains 1218 image-instruction pairs. The limited data is not enough to align the feature from MLLMs and the segmentation model, and will heavily harm the generalization ability of our LLaVASeg. To address this problem, we collect the segmentation dataset without complicated queries as our second kind of data. It includes referred segmentation datasets [12, 27] and semantic segmentation datasets [2, 45]. Further introduction for these datasets is in Sec. 4. In these datasets, only a brief description of the target is provided. This makes it difficult to construct the full chain-of-thought prompting when training with these datasets. As a result, we use different Q-A templates to simulate the first step of chain-of-thought prompting. Since the parameters of MLLMs are frozen throughout the training, this will not affect the quality of reasoning. A large number of training samples can effectively promote our LLaVASeg with strong segmentation and generalization ability.

4. Experiment

4.1. Experimental Setting

Datasets: Following LISA [17], our method adopts a composed dataset containing multiple datasets from different tasks. These datasets are collected from three categories as mentioned in Sec. 3.4. The first category is semantic segmentation datasets, including ADE20k [45], COCO-Stuff [2], and part semantic segmentation dataset PACO-LVIS [33], and PASCAL-Part [5]. The second is referring segmentation datasets, including RefCOCO [12], RefCOCO+ [12], RefCOCOg [27], Refclef [12]. The datasets mentioned above provide a large number of visual samples with different granularity to promote our LLaVASeg with strong segmentation ability and generalization abilities. The third category is the reasoning segmentation dataset ReasonSeg [17]. We evaluate the performance of the conversational ability based on the reasoning segmentation dataset by evaluating the MLLMs' response in the reason prompting step. Furthermore, we exclude the COCO samples in the refCOCO(+/g) validation set to avoid data leakage.

Implementation Details: For all the experiments, we use LLaVA [23] as the architecture of MLLMs if not specified. The backbone of LLaVA is set as Llama2 [36]. For the segmentation model, we choose the backbone of ViT-H SAM [15] as the vision backbone. For model training, AdamW [25] is used as our optimizer. The overall learning

Table 1. Reasoning segmentation results of LLaVASeg (ours) and previous related works. 'explain' denotes that we use the LISA model that can output both the segmentation mask and its explanation. ↑ denotes higher is better.

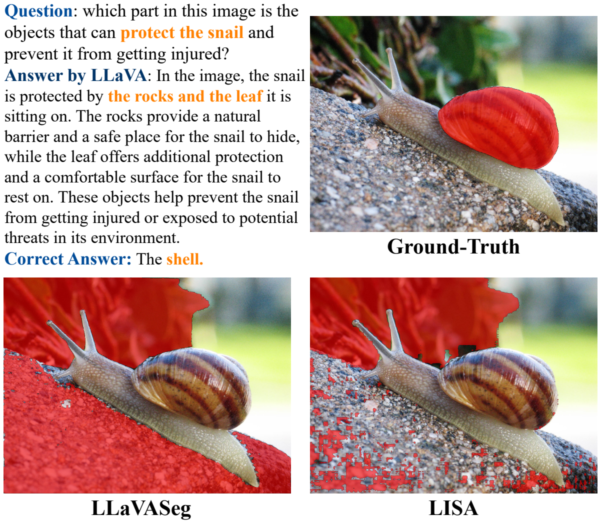
Method	ReasonSeg val		ReasonSeg val	
	gIoU ↑	cIoU ↑	ROUGE-L ↑	CIDEr ↑
OVSeg [19]	28.5	18.6	-	-
GRES [21]	22.4	19.9	-	-
X-Decoder [49]	22.6	17.9	-	-
SEEM [50]	25.5	21.2	-	-
LISA-7B	44.4	46.0	-	-
LISA-13B (explain)	57.3	**60.7**	0.290	0.107
LLaVASeg-7B	54.8	49.9	-	-
LLaVASeg-13B	**59.1**	52.8	**0.393**	**0.796**

rate is set as 0.0001, and the weight decay is set to 0.0001. λ_{bce} and λ_{dice} are set to 1 and 0.5, respectively. The overall batch size is set to 160, where each sample contains a pair of text and a referred mask. We train our model for 12000 iterations in total.

Evaluation Metrics: Similar to LISA [17], we adopt gIoU (the average of all per-image intersection-over-unions) and cIoU (the cumulative intersection over the cumulative union) as evaluation metrics on reasoning segmentation. We evaluate our framework's dialogue quality based on the CIDEr metric [37] and ROUGE-L metric [20]. ROUGE-L is a metric commonly used in natural language processing for evaluating the quality of summaries by calculating the length of the longest common subsequences between the generated summary and the reference summary. CIDEr [37] is a metric specifically designed for evaluating the quality of image captions. It takes into account consensus and diversity among reference captions, providing a more comprehensive evaluation of the generated image captions. We adopt these two evaluation metrics for a more complete and comprehensive evaluation of the generated response from our framework and LISA.

4.2. Reasoning Segmentation

Segmentation: We evaluate the segmentation performance of our LLaVASeg and compare it with the previous methods on the ReasonSeg dataset. For a fair comparison with LISA, we utilize the 'LISA-13B-explain' model which can output both the explanation and the segmentation mask. The result is shown in Tab. 1. It can be seen that our LLaVASeg achieves superior performance on segmentation, which is 1.8% higher than LISA in terms of the gIoU metric. Furthermore, we should note that our LLaVASeg does not fine-tune LLaVA on the Reasonseg dataset, and inherently makes mistakes in reasoning and results in lower cIoU than LISA. We present a failure case study as shown in Fig. 3. In this case, when the LLaVA model fails to reason about the correct answer, both LLaVASeg and LISA predict the wrong area. However, due to LLaVASeg's superior segmentation ability, it predicts the area referred by the LLaVA response more completely than LISA. Consequently, such samples result in higher cumulative union and lower cIoU. Besides, we also observe that gIoU is much more stable than cIoU since cIoU is largely biased towards large-area target [17]. This experiment shows that our model can achieve competitive segmentation performance without fine-tuning MLLMs.



Question: which part in this image is the objects that can **protect the snail** and prevent it from getting injured?

Answer by LLaVA: In the image, the snail is protected by **the rocks and the leaf** it is sitting on. The rocks provide a natural barrier and a safe place for the snail to hide, while the leaf offers additional protection and a comfortable surface for the snail to rest on. These objects help prevent the snail from getting injured or exposed to potential threats in its environment.

Correct Answer: The **shell.**

Figure 3. A failure case study for LLaVASeg and LISA. The upper left corner shows that the LLaVA fails to reason the correct answer from the user query.

Explanation Generation: To investigate the impact of fine-tuning MLLMs for segmentation tasks on their original reasoning ability and dialogue ability, we evaluate the explanation performance. Specifically, we prompt both the LISA and the off-the-shelf LLaVA used in our LLaVASeg, to produce the answer and the explanation to the user query. The experiment's results, presented in Tab. 1, reveal a notable decline in the dialogue performance when comparing LISA with LLaVA. These findings support our motivation and prove the significance of our work that empowers the MLLMs with segmentation ability while preserving their reasoning ability. We also present visualization results in Fig. 4 for an intuitive understanding of our method.

Vanilla Referring Segmentation: We evaluate our method on the referring segmentation datasets to show that the proposed LLaVASeg framework can also handle the vanilla referring segmentation task. As shown in Tab. 2, our method achieves competitive performance compared with previous works, which outperforms the previous works on most of the dataset splits. This demonstrates the superior segmentation

Table 2. Results on referring segmentation of LLaVASeg and other existing methods. All the results are evaluated with the cIoU metric.

Method	refCOCO			refCOCO+			refCOCOg	
	val	testA	testB	val	testA	testB	val(U)	test(U)
MCN [26]	62.4	64.2	59.7	50.6	55.0	44.7	49.2	49.4
VLT [9]	67.5	70.5	65.2	56.3	61.0	50.1	55.0	57.7
CRIS [40]	70.5	73.2	66.1	62.3	68.1	53.7	59.9	60.4
LAVT [42]	72.7	75.8	68.8	62.1	68.4	55.1	61.2	62.1
ReLA [21]	73.8	76.5	70.2	**66.0**	71.0	57.7	65.0	66.0
X-Decoder [49]	-	-	-	-	-	-	64.6	-
SEEM [50]	-	-	-	-	-	-	65.7	-
LISA-7B	74.9	79.1	72.3	65.1	70.8	**58.1**	67.9	**70.6**
LLaVASeg-7B	**76.2**	**79.1**	**72.9**	65.7	**71.4**	57.7	**69.8**	70.1

Table 3. **Ablation study.** We perform experiments on ReasonSeg with LLaVSeg. All the experiments are done in LLaVASeg-13B except for (a).

(a) Ablation on different prompts. The step refers to that described in Sec.3.2. Specifically, Step 1 refers to the reason prompting step. Step 2 refers to the target prompting step. Step 3 refers to the attribute prompting step.

Step 1	Step 2	Step 3	gIoU	cIoU
✔			36.7	31.4
✔	✔		50.2	43.8
✔	✔	✔	54.8	49.9

(b) Ablation the number of scales when prompting the segmentation model.

Settings	Backbone	gIoU	cIoU
Single-scale	LLaVA-13B	55.4	52.5
Multi-scale	LLaVA-13B	59.1	52.8

(c) Segmentation performance when replacing chain-of-thought with LISA.

Method	Backbone	gIoU	cIoU
Replaced with LISA	LLaVA-13B	54.3	50.9
LLaVASeg	LLaVA-13B	59.1	52.8

(d) Ablation on generalization ability. 'nors' denotes that our model has not been trained with the ReasonSeg dataset.

Method	Backbone	gIoU	cIoU
LLaVASeg (nors)	LLaVA-13B	52.8	45.3
LLaVASeg	LLaVA-13B	59.1	52.8

ability of our LLaVASeg.

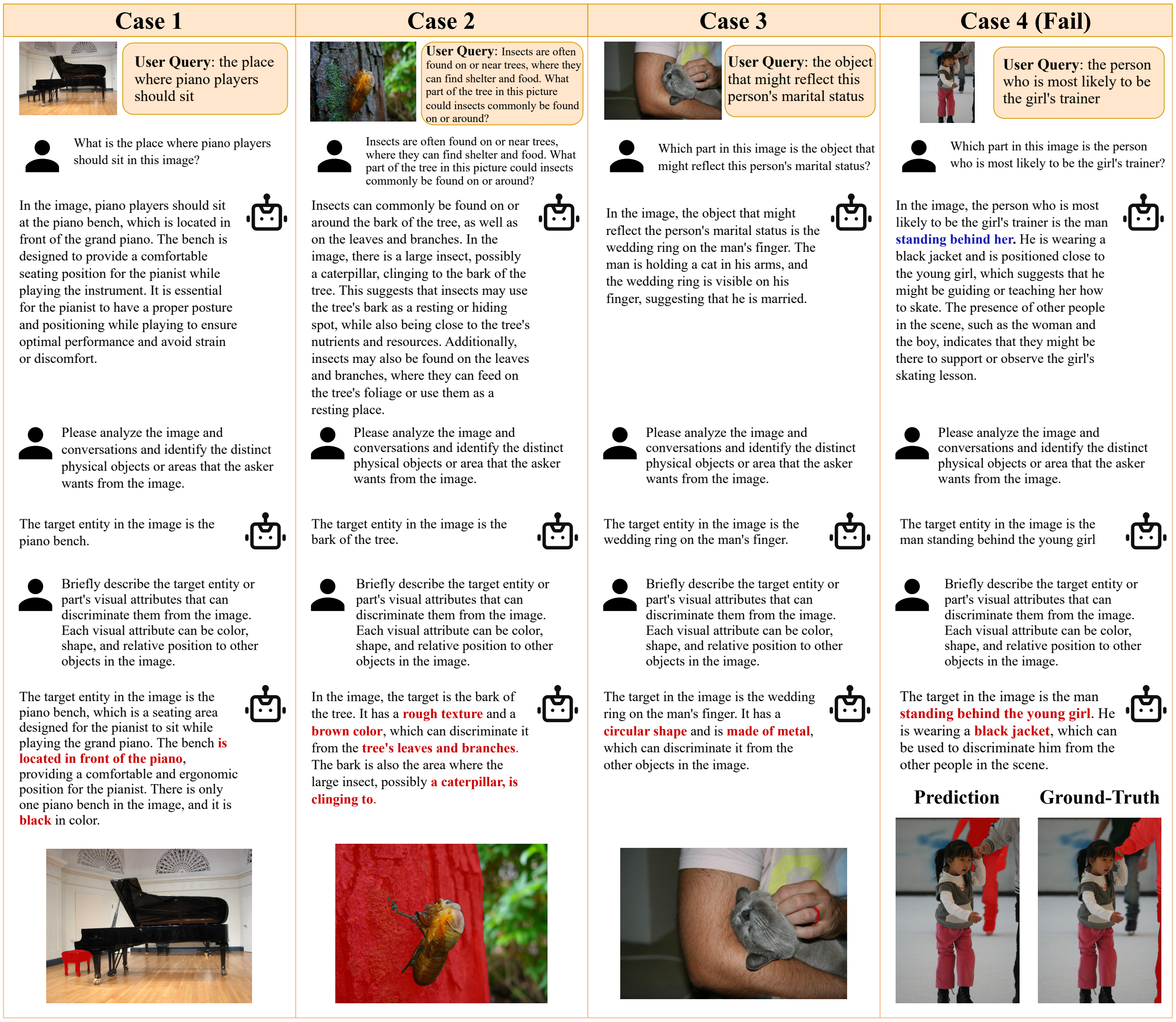
4.3. Ablation Study

Different Prompts: To show the effectiveness of the proposed chain-of-thought prompting, we ablate on the performance of our LLaVASeg when using different prompts. Specifically, we use only one or two steps of our chain-of-thought prompting in this ablation for comparison. The results are shown in Tab. 3a. It can be seen from the results that when only the first step is leveraged, the performance drops significantly. When the targeting prompting step is added, the performance increases. This indicates that the unrelated rationale and explanation hinder the segmentation severely. When we further add the attribute prompting step, the performance also increases by a large margin, which demonstrates the effectiveness of the visual attributes. This underscores the effectiveness of our chain-of-thought prompting in maximizing the reasoning segmentation accuracy.

Effectiveness of Multi-Scale Prompting: We ablate on the impact of the multi-scale features on the reasoning segmentation performance. The results are shown in Tab. 3b. It can be seen that the setting using the multi-scale features achieves superior performance than using single-scale features. the performance of both gIoU and cIoU increases accordingly. These results demonstrate that the extracted vision attributes including both low-level and high-level information can benefit from the multi-scale features.

Replacing Chain-of-Thought with LISA: In this experiment, we show the effectiveness of the explicitly generated visual attributes. We employed LISA, which extracts the visual attributes implicitly in the <SEG> token, to undertake the second and third steps in the chain-of-thought prompting as the comparison. Specifically, we give the MLLMs response from the first step to the LISA as the instruction for segmentation. The results are shown in Tab. 3c. We can see that LISA achieves a gIoU of 54.3% and a cIoU of 50.9%, and LLaVASeg outperforms it with a gIoU of 59.1% and a cIoU of 52.8%. We argue that the significant improvements mainly come from the explicit visual attributes extraction in our chain-of-thought prompting when comparing with LISA. With the same reasoning response as input, our LLaVASeg leverages an explicitly generated visual attribute to aid the segmentation model, while LISA does this implicitly in the <SEG> token.

Ablation on Generalization Ability: The experiments are evaluated on the ReasonSeg validation set. To test the gener-


Case 1

User Query: the place where piano players should sit

What is the place where piano players should sit in this image?

In the image, piano players should sit at the piano bench, which is located in front of the grand piano. The bench is designed to provide a comfortable seating position for the pianist while playing the instrument. It is essential for the pianist to have a proper posture and positioning while playing to ensure optimal performance and avoid strain or discomfort.

Please analyze the image and conversations and identify the distinct physical objects or areas that the asker wants from the image.

The target entity in the image is the piano bench.

Briefly describe the target entity or part's visual attributes that can discriminate them from the image. Each visual attribute can be color, shape, and relative position to other objects in the image.

The target entity in the image is the piano bench, which is a seating area designed for the pianist to sit while playing the grand piano. The bench **is located in front of the piano**, providing a comfortable and ergonomic position for the pianist. There is only one piano bench in the image, and it is **black** in color.

Case 2

User Query: Insects are often found on or near trees, where they can find shelter and food. What part of the tree in this picture could insects commonly be found on or around?

Insects are often found on or near trees, where they can find shelter and food. What part of the tree in this picture could insects commonly be found on or around?

Insects can commonly be found on or around the bark of the tree, as well as on the leaves and branches. In the image, there is a large insect, possibly a caterpillar, clinging to the bark of the tree. This suggests that insects may use the tree's bark as a resting or hiding spot, while also being close to the tree's nutrients and resources. Additionally, insects may also be found on the leaves and branches, where they can feed on the tree's foliage or use them as a resting place.

Please analyze the image and conversations and identify the distinct physical objects or area that the asker wants from the image.

The target entity in the image is the bark of the tree.

Briefly describe the target entity or part's visual attributes that can discriminate them from the image. Each visual attribute can be color, shape, and relative position to other objects in the image.

In the image, the target is the bark of the tree. It has a **rough texture** and a **brown color**, which can discriminate it from the **tree's leaves and branches**. The bark is also the area where the large insect, possibly **a caterpillar, is clinging to**.

Case 3

User Query: the object that might reflect this person's marital status

Which part in this image is the object that might reflect this person's marital status?

In the image, the object that might reflect the person's marital status is the wedding ring on the man's finger. The man is holding a cat in his arms, and the wedding ring is visible on his finger, suggesting that he is married.

Please analyze the image and conversations and identify the distinct physical objects or area that the asker wants from the image.

The target entity in the image is the wedding ring on the man's finger.

Briefly describe the target entity or part's visual attributes that can discriminate them from the image. Each visual attribute can be color, shape, and relative position to other objects in the image.

The target in the image is the wedding ring on the man's finger. It has a **circular shape** and is **made of metal**, which can discriminate it from the other objects in the image.

Case 4 (Fail)

User Query: the person who is most likely to be the girl's trainer

Which part in this image is the person who is most likely to be the girl's trainer?

In the image, the person who is most likely to be the girl's trainer is the man **standing behind her.** He is wearing a black jacket and is positioned close to the young girl, which suggests that he might be guiding or teaching her how to skate. The presence of other people in the scene, such as the woman and the boy, indicates that they might be there to support or observe the girl's skating lesson.

Please analyze the image and conversations and identify the distinct physical objects or area that the asker wants from the image.

The target entity in the image is the man standing behind the young girl

Briefly describe the target entity or part's visual attributes that can discriminate them from the image. Each visual attribute can be color, shape, and relative position to other objects in the image.

The target in the image is the man **standing behind the young girl**. He is wearing a **black jacket**, which can be used to discriminate him from the other people in the scene.

Figure 4. Visualization of the results of our pipeline. We also show one failure case caused by a mistake in the reasoning step for a deeper understanding to our method.

alization ability of LLaVASeg, we train LLaVASeg on the dataset excluding the ReasonSeg training set. The results in Tab. 3d show that there is a performance drop when the ReasonSeg training set is excluded. Although the performance degrades, our LLaVASeg without training on the ReasonSeg dataset still achieves competitive performance, which further supports its generalization ability.

5. Limitations and Future Works

The main limitations of our methods are two-fold. Firstly, the current LLaVASeg only supports one query in one round of interaction. A prompt design supporting multiple queries in one round awaits to be studied. Secondly, our LLaVASeg leverages off-the-shelf MLLMs. However, its performance can be further improved by instruction tuning with high-quality chain-of-thought instruction pairs. Given that the chain-of-thought instruction pairs are fully text rather than the visual prompt <SEG>, it can be expected that there is less impact on its conversational ability. To sum up, we aim to develop a more comprehensive prompting design and dive into the potential of instruction tuning in our future research.

6. Conclusion

In this work, we present a novel reasoning segmentation framework LLaVASeg. LLaVASeg equips the MLLMs with segmentation ability while maintaining their conversational and reasoning ability. To this end, we propose a novel chain-of-thought prompting strategy that iteratively prompts

MLLMs to generate image-specific textual attributes for prompt the segmentation model. To better use these attributes, we leverage a multi-scale prompting segmentation pipeline and train lightweight language-aware adapters for the interaction between textual attributes and visual features. Our method performs competitive segmentation ability by using the off-the-shelf MLLMs. We hope our work will provide valuable inspiration for the marrying of MLLMs and different vision tasks.

References

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

[2] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018.

[3] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

[4] Wei-Ge Chen, Irina Spiridonova, Jianwei Yang, Jianfeng Gao, and Chunyuan Li. Llava-interactive: An all-in-one demo for image chat, segmentation, generation and editing. *arXiv preprint arXiv:2311.00571*, 2023.

[5] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1971–1978, 2014.

[6] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023.

[7] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

[8] Shizhe Diao, Rui Pan, Hanze Dong, Ka Shun Shum, Jipeng Zhang, Wei Xiong, and Tong Zhang. Lmflow: An extensible toolkit for finetuning and inference of large foundation models. *arXiv preprint arXiv:2306.12420*, 2023.

[9] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16321–16330, 2021.

[10] Guang Feng, Zhiwei Hu, Lihe Zhang, and Huchuan Lu. Encoder fusion network with co-attention embedding for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15506–15515, 2021.

[11] Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. Complexity-based prompting for multi-step reasoning. *arXiv preprint arXiv:2210.00720*, 2022.

[12] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014.

[13] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[14] Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. Decomposed prompting: A modular approach for solving complex tasks. *arXiv preprint arXiv:2210.02406*, 2022.

[15] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[16] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[17] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023.

[18] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[19] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *CVPR*, 2023.

[20] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[21] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *CVPR*, 2023.

[22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[23] Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, et al. Llava-plus: Learning to use tools for creating multimodal agents. *arXiv preprint arXiv:2311.05437*, 2023.

[24] Zhaoyang Liu, Yinan He, Wenhai Wang, Weiyun Wang, Yi Wang, Shoufa Chen, Qinglong Zhang, Zeqiang Lai, Yang Yang, Qingyun Li, et al. Interngpt: Solving vision-centric tasks by interacting with chatgpt beyond language. *arXiv preprint arXiv:2305.05662*, 3, 2023.

[25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[26] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and

segmentation. In *CVPR*, 2020.

[27] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016.

[28] OpenAI. Chatgpt. `https://openai.com/blog/chatgpt/`. Accessed: 2023-09-27.

[29] OpenAI. Gpt-4v(ision) system card. `https://cdn.openai.com/papers/GPTV_System_Card.pdf`. Accessed: 2023-10-09.

[30] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Adv. Neural Inform. Process. Syst.*, 35:27730–27744, 2022.

[31] Renjie Pi, Jiahui Gao, Shizhe Diao, Rui Pan, Hanze Dong, Jipeng Zhang, Lewei Yao, Jianhua Han, Hang Xu, and Lingpeng Kong Tong Zhang. Detgpt: Detect what you need via reasoning. *arXiv preprint arXiv:2305.14167*, 2023.

[32] Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*, 2022.

[33] Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. Paco: Parts and attributes of common objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7141–7151, 2023.

[34] Jiajin Tang, Ge Zheng, Jingyi Yu, and Sibei Yang. Cotdet: Affordance knowledge prompting for task driven object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3068–3078, 2023.

[35] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[36] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[37] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[38] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv preprint arXiv:2305.11175*, 2023.

[39] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

[40] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *CVPR*, 2022.

[41] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[42] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18155–18165, 2022.

[43] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

[44] Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *arXiv preprint arXiv:2310.16436*, 2023.

[45] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[46] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2022.

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

[48] Lanyun Zhu, Tianrun Chen, Deyi Ji, Jieping Ye, and Jun Liu. Llafs: When large-language models meet few-shot segmentation. *arXiv preprint arXiv:2311.16926*, 2023.

[49] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *CVPR*, 2023.

[50] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *arXiv:2304.06718*, 2023.