NOVO: Bridging LLaVA and SAM with Visual-only Prompts for Reasoning Segmentation

Kyung-Yoon Yoon and Yeong-Jun Cho[*]
Department of Artificial Intelligence Convergence
Chonnam National University, Gwangju 61186, South Korea
{kyungyoon201, yj.cho}@jnu.ac.kr

Abstract

In this study, we propose NOVO (NO text, Visual-Only prompts), a novel framework that bridges vision-language models (VLMs) and segmentation models through visual-only prompts. Unlike prior approaches that feed text-derived <SEG> token embeddings into segmentation models, NOVO instead generates a coarse mask and point prompts from the VLM output. These visual prompts are compatible with the Segment Anything Model (SAM), preserving alignment with its pretrained capabilities. To further enhance boundary quality and enable instance-level segmentation, we introduce a training-free refinement module that reduces visual artifacts and improves the quality of segmentation masks. We also present RISeg, a new benchmark comprising 918 images, 2,533 instance-level masks, and diverse reasoning queries to evaluate this task. Experiments demonstrate that NOVO achieves state-of-the-art performance across multiple metrics and model sizes, demonstrating its effectiveness and scalability in reasoning segmentation.

Code and Dataset: *https://WILL_BE_SOON*

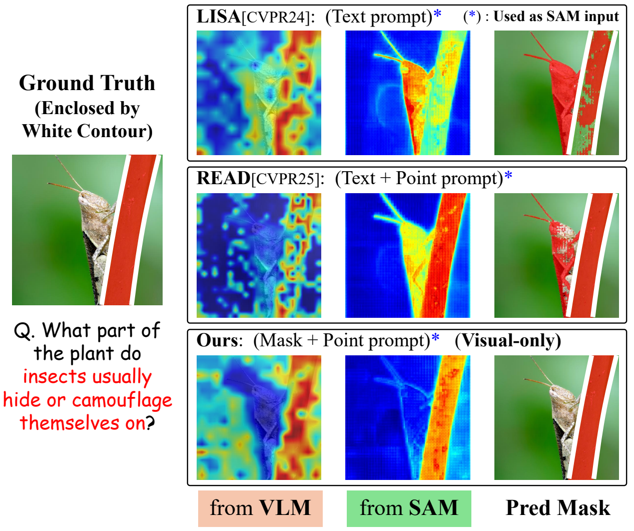
LISA[CVPR24]: (Text prompt)*
(*) : Used as SAM input
Ground Truth
(Enclosed by
White Contour)
READ[CVPR25]: (Text + Point prompt)*
Q. What part of the plant do insects usually hide or camouflage themselves on?
Ours: (Mask + Point prompt)* (Visual-only)
from VLM
from SAM
Pred Mask

Figure 1. Comparison of VLM activation maps, SAM mask logits, and predicted masks from different models. Our method, which uses visual prompts only, produces masks that better align with the semantics of the input query. In contrast, LISA and READ often show a mismatch between VLM activations and SAM mask logits. Best viewed in color.

1. Introduction

In real-world scenarios, segmentation is not always about detecting clearly labeled objects. It often requires understanding visual context and reasoning over indirect descriptions. For example, Fig.1 shows a case where the goal is to segment the region described by the prompt, “find the part of the plant where insects usually hide or camouflage themselves,” rather than directly segmenting the leaves themselves. This challenge is known as Reasoning Segmentation, which deals with complex, query-based segmentation. It takes an implicit text query as input and predicts a binary mask for the relevant region. Unlike traditional Referring Segmentation [7, 19], which targets explicitly mentioned objects, this task requires deeper language understanding and visual reasoning.

A common strategy in reasoning segmentation is to extract the <SEG> token embedding from a vision-language model (VLM) and use it as a prompt for segmentation. Methods like LISA [9] and READ [20] follow this approach using SAM [8] for segmentation. However, SAM is trained to respond to visual prompts–such as points, boxes, and masks–rather than text embeddings. Therefore, directly using a text-based <SEG> token embedding may not align with this training, potentially leading to inaccurate predictions. To examine this issue, we compare VLM activation maps and SAM’s predicted masks using existing models, as shown in Fig. 1. While the VLM activation in existing methods (e.g., LISA, READ) often points to the correct region, the corresponding SAM output masks frequently fail to align with it. This motivates a new approach that transforms the output of VLMs (i.e., <SEG> token embedding) into visual prompts, allowing SAM to accurately segment regions based on the input query.

To address these limitations, we propose a new framework called NOVO (NO text, Visual-Only prompts) for reasoning segmentation as described in Sec. 4. Instead of directly feeding text-based embeddings into the segmentation model, NOVO performs reasoning segmentation using only visual prompts. Specifically, NOVO generates two types of visual prompts: a coarse mask prompt derived from the VLM activation map, and point prompts sampled from regions with high semantic relevance. These visual prompts are then used to guide SAM, without relying on any text-based inputs.

To further boost performance, we propose a refinement module (Sec. 4.5) that selects the most relevant masks from SAM's outputs. This design also allows NOVO to fully leverage SAM's segmentation capabilities, resulting in accurate and semantically aligned outputs. Interestingly, our refinement process naturally extends to instance-level reasoning segmentation, without requiring additional model modifications. To support evaluation of instance-level reasoning segmentation, we construct a new benchmark dataset, RISeg (Sec. 5). It includes 918 images from COCO, 2,533 instance masks, and diverse reasoning queries automatically generated using GPT-4o and verified by human experts.

NOVO outperforms previous state-of-the-art methods and achieves new best results on the ReasonSeg benchmark. Specifically, NOVO achieves 62.4% gIoU and 65.7% cIoU on the test set when using a 7B vision-language backbone, outperforming the previous best by +3.9% and +6.2%, respectively. With a stronger 13B backbone, NOVO further improves to 65.3% gIoU and 66.0% cIoU, establishing a new state-of-the-art. In addition, our NOVO Refinement significantly improves boundary quality by mitigating visual artifacts such as holes and imprecise contours. It also enables instance-level segmentation without additional training and consistently outperforms prior methods on the RISeg benchmark.

Our main contributions are summarized as follows:

- We propose a novel reasoning segmentation framework, NOVO, that uses only visual prompts to better align the VLM's understanding with the segmentation output.
- We introduce a refinement module that enhances mask quality and enables instance-level segmentation without requiring additional training.
- We construct a new benchmark dataset, RISeg, designed to evaluate instance-level reasoning segmentation.
- Extensive experiments show that NOVO consistently outperforms previous methods and achieves new state-of-the-art results on reasoning segmentation benchmarks.

2. Related Works

Prompt-based Segmentation. Recent advances in segmentation have shifted from fixed class-based pixel prediction to prompt-based methods that adapt flexibly to user inputs. The Segment Anything Model (SAM) [8] is a foundation model that enables high-quality segmentation using various user prompts (e.g., text, points, boxes, masks). Building on this, recent research has explored the use of natural language prompts in addition to visual ones. Approaches such as OVSeg [12], GRES [14], X-Decoder [32], and SEEM [33] introduce unified frameworks that leverage CLIP [21] or other vision-language models to address diverse segmentation tasks. These models aim to improve zero-shot inference while supporting both open-vocabulary and referring segmentation scenarios. However, they still struggle with queries requiring contextual or implicit reasoning, leading to extensions toward models with stronger reasoning ability.

Reasoning Segmentation. To address queries requiring contextual or implicit understanding, the task of "*Reasoning Segmentation*" has been introduced. This task aims to segment objects based on language queries that require high-level reasoning, such as indirect or context-dependent descriptions. LISA [9] first introduced the concept of reasoning segmentation by using the <SEG> token embedding from a vision-language model (VLM) as a prompt to guide a segmentation model. SESAME [25] extended this approach to handle queries involving non-existent objects (i.e., false premises). READ [20] further analyzed the semantic function of the <SEG> token and incorporated highly activated points derived from similarity maps. LISA++ [27] built upon this framework by introducing multiple <SEG> tokens to support instance-level reasoning and interactive segmentation. These methods generally use SAM as the segmentation backbone, which is mainly trained with visual prompts. As a result, using language-based embeddings with SAM may cause a mismatch with its training objectives and lead to reduced segmentation performance. Subsequent works such as GSVA [26], SAM4MLLM [4], GLaMM [23], POPEN [31], and RSVP [17] further expand reasoning segmentation by handling absent objects, MLLM-guided prompts, adopting end-to-end decoding, applying preference optimization, or leveraging chain-of-thought.

Vision-Language Models (VLMs). Recent advances in vision-language models (VLMs) have significantly influenced the development of text-based reasoning segmentation. Representative multimodal models include Flamingo [1], BLIP-2 [11], mPLUG-Owl [28], OTTER [10], LLaVA [15], and MiniGPT-4 [30]. These models have shown strong performance on complex tasks such as instruction-based question answering, explanation generation, and visual reasoning. These models integrate visual and language features, enabling contextual understanding and user intent recognition, which are essential capabilities for reasoning segmentation.

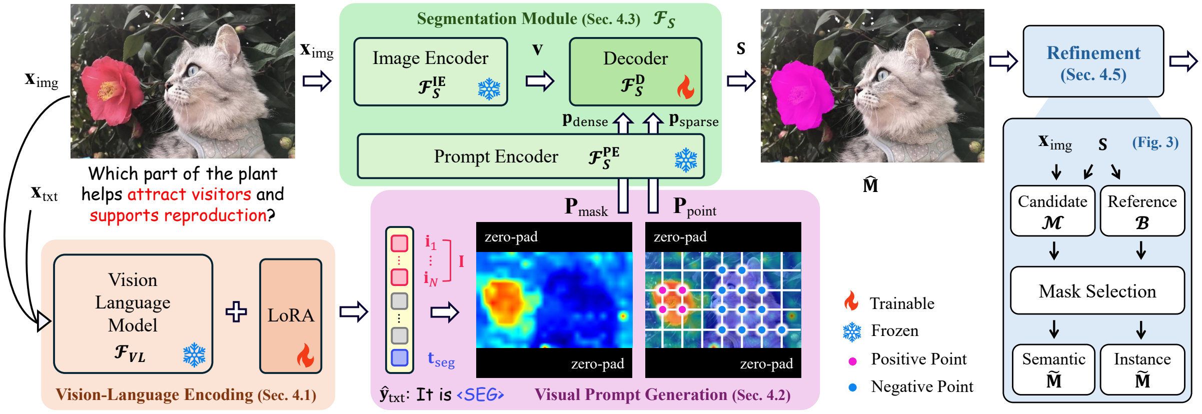
$\mathbf{x}_{\text{img}}$
$\mathbf{x}_{\text{txt}}$
Which part of the plant helps attract visitors and supports reproduction?
$\mathbf{x}_{\text{img}}$
Segmentation Module (Sec. 4.3) $\mathcal{F}_S$
Image Encoder $\mathcal{F}_S^{\text{IE}}$
$\mathbf{V}$
Decoder $\mathcal{F}_S^{\text{D}}$
$\mathbf{p}_{\text{dense}}$
$\mathbf{p}_{\text{sparse}}$
Prompt Encoder $\mathcal{F}_S^{\text{PE}}$
$\mathbf{S}$
$\hat{\mathbf{M}}$
Refinement (Sec. 4.5)
$\tilde{\mathbf{M}}$
$\mathbf{P}_{\text{mask}}$
$\mathbf{P}_{\text{point}}$
Vision Language Model $\mathcal{F}_{VL}$
LoRA
Vision-Language Encoding (Sec. 4.1)
$\mathbf{i}_1$
$\mathbf{i}_N$
$\mathbf{I}$
$\mathbf{t}_{\text{seg}}$
zero-pad
zero-pad
zero-pad
zero-pad
$\hat{\mathbf{y}}_{\text{txt}}$: It is <SEG>
Visual Prompt Generation (Sec. 4.2)
Trainable
Frozen
Positive Point
Negative Point
$\mathbf{x}_{\text{img}}$ $\mathbf{S}$ (Fig. 3)
Candidate $\mathcal{M}$
Reference $\mathcal{B}$
Mask Selection
Semantic $\tilde{\mathbf{M}}$
Instance $\tilde{\mathbf{M}}$

Figure 2. Overview of the proposed NOVO. It encodes an input image $\mathbf{x}_{\text{img}}$ and a reasoning text query $\mathbf{x}_{\text{txt}}$ via a VLM to extract the <SEG> token embedding $\mathbf{t}_{\text{seg}}$. Together with image patch embeddings $\mathbf{I}$, it generates the mask prompt $\mathbf{P}_{\text{mask}}$ and point prompt $\mathbf{P}_{\text{point}}$, which are passed as the sole inputs to the NOVO's segmentation module to produce a segmentation mask $\hat{\mathbf{M}}$. The predicted mask can be refined using our proposed method in Sec. 4.5, which not only improves segmentation quality but also enables instance-level mask generation.

3. Motivation and Main Idea

Reasoning segmentation aims to generate a binary mask $\hat{\mathbf{M}} \in \{0,1\}^{h \times w}$ from a text query $\mathbf{x}_{\text{txt}}$ and an input image $\mathbf{x}_{\text{img}}$, defined as $\hat{\mathbf{M}} = \mathcal{F}_\theta(\mathbf{x}_{\text{img}}, \mathbf{x}_{\text{txt}})$. The model $\mathcal{F}_\theta$, composed of a vision-language model (VLM) $\mathcal{F}_{\text{VL}}$ and a segmentation model $\mathcal{F}_{\text{S}}$, is trained to optimize the parameters θ such that the predicted mask $\hat{\mathbf{M}}$ closely matches the ground-truth mask. For the segmentation model, previous studies [9, 20] adopted Segment Anything Model (SAM) [8]. The VLM infers the location of the object described by the text query, while the SAM generates a precise segmentation mask based on that location.

As shown in Fig. 1, we investigated the causes of performance degradation in reasoning segmentation. While the VLM often accurately infers the target region from the text query, existing methods such as LISA [9] and READ [20] often fail to leverage this information. Moreover, the inconsistency between the outputs of the VLM and SAM leads to inaccurate reasoning masks. We note that SAM has been extensively pretrained on large-scale data using visual prompts such as points and boxes rather than text-based prompts. However, relying on text-based prompts in previous methods (e.g., LISA and READ) may not sufficiently activate the pretrained segmentation capabilities of SAM.

Rather than enhancing SAM's understanding of text prompts through additional training, we propose a novel approach called NOVO (NO text, Visual-Only prompts), which transforms VLM outputs into visual prompts that align with SAM's pretraining format. This design more effectively leverages SAM's pretrained capabilities for reasoning segmentation. To further enhance segmentation quality, we also propose NOVO Refinement, a training-free method that converts segmentation logits into point prompts and uses SAM to generate precise instance-level masks.

4. Proposed NOVO

We propose NOVO (NO text, Visual-Only prompts) for reasoning segmentation as illustrated in Fig. 2. It consists of four main stages: (1) Vision-Language Encoding, (2) Visual Prompt Generation, (3) Segmentation Module, and (4) NOVO Refinement.

4.1. Vision-Language Encoding

In NOVO, we utilize LLaVA [15] as the vision-language model (VLM), which takes an image $\mathbf{x}_{\text{img}}$ and a reasoning-based text query $\mathbf{x}_{\text{txt}}$ as input. The model encodes the target region for segmentation using the <SEG> token along with image embeddings and the corresponding token embedding. LLaVA's image encoder requires a fixed square aspect ratio, and while some prior approaches adopt a center-cropping strategy, we instead apply zero-padding based on the longer side to preserve the original visual information. For consistency, we apply the same padding strategy to the segmentation model described in Sec. 4.3.

An input image $\mathbf{x}_{\text{img}}$ is transformed into N patch embeddings via CLIP's image encoder in LLaVA and projected to match the dimensionality of the LLM's hidden states. These image embeddings are then concatenated with the token embeddings of the input query $\mathbf{x}_{\text{txt}}$, forming a unified multimodal sequence that is processed by the LLM. As a result, the model generates an output text response $\hat{\mathbf{y}}_{\text{txt}}$, which includes a <SEG> token. Although $\hat{\mathbf{y}}_{\text{txt}}$ is the text output of the VLM, we additionally extract two components from its final hidden states: the image patch embeddings $\mathbf{I} = [\mathbf{i}_1, \ldots, \mathbf{i}_N] \in \mathbb{R}^{N \times d}$ and the <SEG> token embedding $\mathbf{t}_{\text{seg}} \in \mathbb{R}^d$, where d is the embedding dimension. The model is trained to ensure that $\mathbf{t}_{\text{seg}}$ effectively encodes the semantics of the region to be segmented.

4.2. Visual Prompt Generation

We explain how to transform the VLM's output embeddings into visual prompts suitable for SAM input. Inspired by READ [20], we use the image embeddings $\mathbf{I}$ and the `<SEG>` token embedding $\mathbf{t}_{\text{seg}}$ to generate the mask prompt $\mathbf{P}_{\text{mask}}$ and the point prompt $\mathbf{P}_{\text{point}}$.

First, the mask prompt is computed by measuring the cosine similarity between $\mathbf{t}_{\text{seg}}$ and each image patch embedding $\mathbf{i}_n$, resulting in a coarse activation map that highlights the region indicated by the text query $\mathbf{x}_{\text{txt}}$ as follows:

$$\mathbf{P}_{\text{mask}} = [\text{sim}(\mathbf{t}_{\text{seg}}, \mathbf{i}_1), \text{sim}(\mathbf{t}_{\text{seg}}, \mathbf{i}_2), \ldots, \text{sim}(\mathbf{t}_{\text{seg}}, \mathbf{i}_N)], \quad (1)$$

where $\text{sim}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}$ denotes the cosine similarity function. Initially, $\mathbf{P}_{\text{mask}}$ is an N-dimensional vector, which is then reshaped and upsampled via bilinear interpolation to form a 256×256 visual prompt map.

Additionally, based on the high and low activation values in $\mathbf{P}_{\text{mask}}$, we extract the coordinates of the top and bottom-ranked positions to construct the point prompt $\mathbf{P}_{\text{point}}$, consisting of positive and negative points. This helps emphasize semantically relevant and irrelevant regions, respectively. Unlike previous reasoning segmentation methods such as LISA [9], which use only text embeddings, and READ [20], which combines text embeddings with point prompts, our approach relies solely on visual inputs. This simple yet focused design improves reasoning segmentation performance by leveraging only visual prompts, which align more directly with SAM's capabilities.

4.3. Segmentation Module

To generate accurate segmentation masks, we adopt the Segment Anything Model (SAM), a general-purpose segmentation model capable of handling various visual prompts. SAM consists of three major components: the image encoder $\mathcal{F}_S^{\text{IE}}$, the prompt encoder $\mathcal{F}_S^{\text{PE}}$, and the mask decoder $\mathcal{F}_S^{\text{D}}$.

First, the image encoder $\mathcal{F}_S^{\text{IE}}$ transforms the input image $\mathbf{x}_{\text{img}}$ into a high-dimensional image embedding, denoted as $\mathbf{v} = \mathcal{F}_S^{\text{IE}}(\mathbf{x}_{\text{img}})$. The prompt encoder $\mathcal{F}_S^{\text{PE}}$ converts the visual prompts $\mathbf{P}_{\text{mask}}, \mathbf{P}_{\text{point}}$ into dense and sparse prompt embeddings respectively, as follows:

$$\mathbf{p}_{\text{dense}} = \mathcal{F}_S^{\text{PE}}(\mathbf{P}_{\text{mask}}), \quad \mathbf{p}_{\text{sparse}} = \mathcal{F}_S^{\text{PE}}(\mathbf{P}_{\text{point}}), \quad (2)$$

where $\mathcal{F}_S^{\text{PE}}$ is designed as a unified module that processes different prompt types, such as masks and points, by internally branching based on the input type. Finally, the mask decoder $\mathcal{F}_S^{\text{D}}$ takes $\mathbf{v}$, $\mathbf{p}_{\text{dense}}$, and $\mathbf{p}_{\text{sparse}}$ as input, and outputs segmentation logits $\mathbf{S}$, which are then binarized to produce the binary segmentation mask:

$$\hat{\mathbf{M}} = \mathcal{F}_S^{\text{D}}(\mathbf{v},\ \mathbf{p}_{\text{dense}},\ \mathbf{p}_{\text{sparse}}). \quad (3)$$

This architecture allows the segmentation model in NOVO to fully exploit the cues provided by visual prompts.

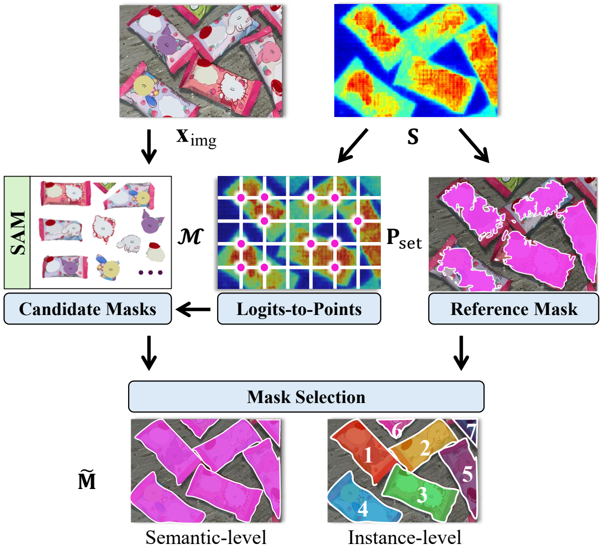

Figure 3. Overview of the NOVO Refinement. Without any additional training, our refinement method effectively combines the initial mask with SAM's segmentation capability, not only enhancing the overall segmentation quality but also enabling instance-level reasoning segmentation.

4.4. Loss Functions

To train networks, we adopt a multi-objective training scheme that jointly optimizes the text generation and mask prediction losses to improve both language reasoning and visual segmentation capabilities. The text generation loss $\mathcal{L}_{\text{txt}}$ encourages the VLM to generate natural and accurate responses to the given text query. It is formulated as a weighted cross-entropy loss between the generated response $\hat{\mathbf{y}}_{\text{txt}}$ and the ground-truth response $\mathbf{y}_{\text{txt}}$:

$$\mathcal{L}_{\text{txt}} = \lambda_{\text{CE}} \cdot \mathcal{L}_{\text{CE}}(\hat{\mathbf{y}}_{\text{txt}}, \mathbf{y}_{\text{txt}}). \quad (4)$$

The mask prediction loss $\mathcal{L}_{\text{mask}}$ encourages the predicted mask $\hat{\mathbf{M}}$, generated based on visual prompts, to align with the ground-truth mask $\mathbf{M}$. We adopt a weighted combination of Binary Cross Entropy (BCE) and Dice loss, defined as:

$$\mathcal{L}_{\text{mask}} = \lambda_{\text{BCE}} \cdot \mathcal{L}_{\text{BCE}}(\hat{\mathbf{M}}, \mathbf{M}) + \lambda_{\text{Dice}} \cdot \mathcal{L}_{\text{Dice}}(\hat{\mathbf{M}}, \mathbf{M}), \quad (5)$$

where $\lambda_{\text{CE}}, \lambda_{\text{BCE}}, \lambda_{\text{Dice}}$ are empirically determined weights for each term. Finally, the model is trained in an end-to-end manner using a total loss that combines both objectives as follows:

$$\mathcal{L} = \mathcal{L}_{\text{txt}} + \mathcal{L}_{\text{mask}}. \quad (6)$$

4.5. NOVO Refinement for High-Quality Instance-level Segmentation

While the proposed NOVO achieves high accuracy in reasoning segmentation, its predicted masks $\hat{\mathbf{M}}$ often exhibit visual artifacts such as holes and inaccurate boundaries. These artifacts result from directly binarizing the segmentation logits $\mathbf{S}$ using a fixed threshold, a process which fails to capture the fine-grained structure of the target region.

To address this, we propose a refinement strategy that leverages the high-quality mask generation capability of the Segment Anything Model (SAM). Rather than producing the final mask from binarized logits, we use $\mathbf{S}$ to generate point prompts $\mathbf{P}_{\text{set}}$, which guide SAM to produce precise instance-level mask candidates, resulting in a refined output mask $\widetilde{\mathbf{M}}$. This strategy not only improves the visual quality of segmentation but also enables instance reasoning segmentation without additional training. The proposed refinement has three steps: (1) logits-to-point sampling, (2) reference mask construction, and (3) mask selection, as shown in Fig. 3.

First, to construct the point prompts $\mathbf{P}_{\text{set}}$, we sample locations from the segmentation logits $\mathbf{S} \in \mathbb{R}^{h \times w}$. We overlay a fixed-interval grid on $\mathbf{S}$ and select grid points whose values exceed 0. The sampled points $\mathbf{P}_{\text{set}}$ are used as visual prompts to SAM, along with the input image $\mathbf{x}_{\text{img}}$. SAM then generates a set of J instance-level candidate masks as follows:

$$\mathcal{M} = \{M_1, M_2, \ldots, M_J\}, \quad M_j \in \{0,1\}^{h \times w}. \tag{7}$$

Each point in $\mathbf{P}_{\text{set}}$ acts as a prompt, leading to one or more candidate masks being generated in its vicinity.

Second, to guide the selection of masks from the candidate set $\mathcal{M}$, we define a binary reference mask $\mathcal{B}$ based on the distribution of the logits $\mathbf{S}$. We apply a fixed threshold $\tau = 0$ to binarize the logits by $\mathcal{B} = \mathbb{1}[\mathbf{S} \geq \tau]$. The reference mask $\mathcal{B}$ serves as the criterion for evaluating and selecting high-quality masks from $\mathcal{M}$. We then compute an overlap score between each candidate mask $M_j \in \mathcal{M}$ and the reference mask $\mathcal{B}$. The score measures how well each candidate M_j aligns with the reference mask, and is defined as:

$$\text{overlap}(M_j, \mathcal{B}) = \frac{|M_j \cap \mathcal{B}|}{|M_j|}. \tag{8}$$

Finally, we select masks with scores exceeding a predefined threshold δ (empirically set to 0.7) to construct the refined output $\widetilde{\mathbf{M}}$. Spatially overlapping masks are merged into a single output mask via union. Importantly, our refinement strategy naturally supports both semantic and instance-level segmentation without additional model training:

- **Semantic-level output**: All selected masks are aggregated to form a unified segmentation map with improved object completeness and boundaries.
- **Instance-level output**: Each selected mask without spatial overlap is treated independently for instance-wise output, enabling instance segmentation.

5. `RISeg`: Dataset for Instance Reasoning Segmentation

To evaluate the performance of instance-level reasoning segmentation, we introduce a new benchmark dataset, `RISeg`. Built upon the COCO2017 [13] validation set, `RISeg` includes 918 images, each paired with a reasoning-focused text query and annotated with one or more ground-truth instance masks–totaling 2,533 instances. Unlike `RefCOCO/+/g` [7, 18], which uses explicit referring expressions, `RISeg` introduces implicit reasoning-based queries requiring deeper context. These instances are distributed across 6 supercategories–person, animal, food & kitchenware, object & furniture, transport & outdoor, and others–and 74 fine-grained object classes.

An overview of `RISeg` is shown in Fig. 4. The construction of `RISeg` follows a four-step pipeline: (1) Image selection: Images containing multiple instances of the same class are selected from the COCO validation set. (2) Query generation: Reasoning queries are generated using a large language model (e.g., GPT-4), guided by the image, its caption, and object labels. The prompts are designed to reflect object attributes, spatial relations, and contextual reasoning. (3) Human validation: Each generated query is reviewed by human annotators to ensure semantic clarity and logical correctness. (4) Mask alignment: Finally, validated queries are paired with their corresponding ground-truth instance masks. Further details are provided in the supplementary materials.

6. Experimental Results

6.1. Settings

Implementation Details. In this study, we adopt LLaVA 1.5-7B and LLaVA 1.5-13B [15] as the Vision-Language Model (VLM) $\mathcal{F}_{\text{VL}}$, and employ the Segment Anything Model (SAM) [8] with a ViT-H backbone as the segmentation module $\mathcal{F}_{\text{S}}$. All experiments were conducted using the DeepSpeed framework on four NVIDIA A100 GPUs (40GB each), with a total training time of approximately 26 hours. Input images were padded to square shapes before being fed into the network, and outputs were restored to their original aspect ratios for evaluation. We used the AdamW optimizer with a learning rate of 0.0005 and a learning rate scheduler with 100 warm-up steps. The loss weights were set to $\lambda_{\text{CE}} = 1.0$ for text generation, and $\lambda_{\text{BCE}} = 2.0$ and $\lambda_{\text{Dice}} = 0.5$ for mask prediction. In NOVO, we apply LoRA [6] to the VLM for parameter-efficient tuning, and jointly train the mask decoder in SAM's segmentation module. All other components, including the image encoder and prompt encoder in SAM, as well as the remaining parts of the VLM, are frozen.

Datasets. We use a range of datasets relevant to reasoning segmentation for training and evaluation: (1) Se-

Methods	val (overall)		test (short query)		test (long query)		test (overall)	
	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU
OVSeg [12]	28.5	18.6	18.0	15.5	28.7	22.5	26.1	20.8
GRES [14]	22.4	19.9	17.6	15.0	22.6	23.8	21.3	22.0
X-Decoder [32]	22.6	17.9	20.4	11.6	22.2	17.5	21.7	16.3
SEEM [33]	25.5	21.2	20.1	11.5	25.6	20.8	24.3	18.7
Grounded-SAM [16]	26.0	14.5	17.8	10.8	22.4	18.6	21.3	16.4
SESAME [25]	34.8	39.1	28.3	27.6	31.6	32.7	30.5	30.4
LLaVA1.5-7B + OVSeg [12, 15]	38.2	23.5	24.2	18.7	44.6	37.1	39.7	31.8
LISA-7B (ft) [9]	52.9	54.0	40.6	40.6	49.4	51.0	47.3	48.4
LISA-7B-LLaVA1.5 (ft) [9]	61.3	62.9	48.3	46.3	57.9	59.7	55.6	56.9
LISA++-7B-LLaVA1.5 (ft) [27]	64.2	68.1	49.6	**51.1**	59.3	61.7	57.0	59.5
READ-7B-LLaVA1.5 (ft) [20]	59.8	67.6	52.6	49.5	60.4	61.0	58.5	58.6
RSVP-GPT-4o [17]	64.7	63.1	**55.4**	50.4	61.9	62.5	60.3	60.0
Ours-7B-LLaVA1.5 (ft)	**66.3**	**69.0**	53.1	49.5	**65.4**	**70.3**	**62.4**	**65.7**
LLaVA1.5-13B + OVSeg [12, 15]	37.9	26.4	27.1	19.4	46.1	40.6	41.5	34.1
LISA-13B-LLaVA1.5 (ft) [9]	65.0	**72.9**	55.4	50.6	63.2	65.3	61.3	62.2
READ-13B-LLaVA1.5 (ft) [20]	-	-	55.4	53.7	64.4	65.1	62.2	62.8
Ours-13B-LLaVA1.5 (ft)	**66.7**	71.9	**57.5**	**57.0**	**67.8**	**68.1**	**65.3**	**66.0**

Table 1. Reasoning segmentation performance on the ReasonSeg dataset. (ft) denotes a model fine-tuned on 239 samples from the ReasonSeg training set. Bold numbers indicate the best performance among 7B-based and 13B-based models, respectively.

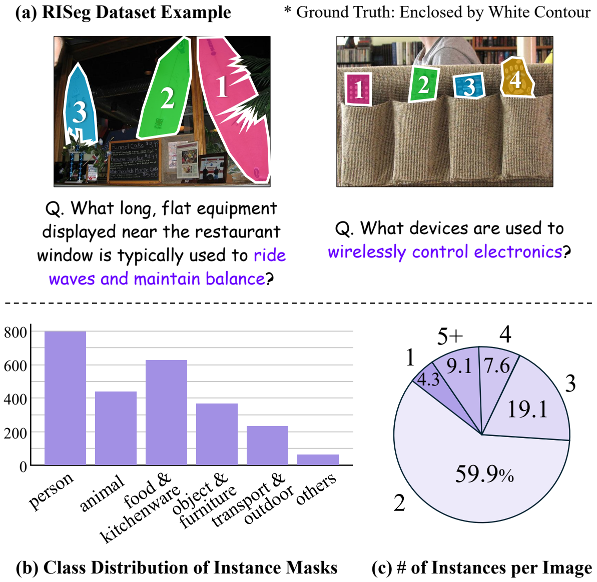

Figure 4. Overview of the RISeg Dataset. (a) Examples from the RISeg dataset, where each image is paired with a reasoning-based text query and multiple ground-truth instance masks (shown with white contours). (b) Distribution of instance mask classes grouped by categories. (c) The number of annotated instances per image.

Methods	val (overall)			test (overall)		
	gIoU	B-IoU	B-F1	gIoU	B-IoU	B-F1
READ-7B[†]	59.8	29.1	38.9	56.3	30.4	39.8
READ-7B (+R)[†]	**61.3**	**36.1**	**46.1**	**56.4**	**34.6**	**44.3**
Ours-7B	66.3	33.9	44.0	62.4	33.6	42.8
Ours-7B (+R)	**67.1**	**40.0**	**50.3**	**62.7**	**39.1**	**48.6**
Ours-13B	66.7	35.5	45.1	**65.3**	36.5	46.0
Ours-13B (+R)	**67.2**	**40.9**	**51.3**	65.2	**40.6**	**50.2**

Table 2. Semantic segmentation results on ReasonSeg with and without NOVO Refinement. (+R) denotes our refinement. [†] indicates values reproduced in our experimental setting.

Methods	AP50	AP75	mAP	AP-S	AP-M	AP-L
LISA++-7B	14.6	5.0	6.1	1.1	2.5	10.6
READ-7B (+R)	43.2	**30.1**	27.2	7.8	18.7	**38.6**
Ours-7B (+R)	**44.7**	30.0	**27.2**	**9.8**	**19.9**	37.0

Table 3. Instance-level reasoning segmentation results on RISeg dataset. (+R) denotes our Refinement method.

mantic segmentation: ADE20K [29], COCO-Stuff [2], PACO-LVIS [22], and PASCAL-Part [3]; (2) Referring segmentation: refCLEF, RefCOCO, RefCOCO+ [7], RefCOCOg [18], and ReasonSeg [9]; (3) Visual question answering: LLaVA Instruct 150K [15]. To further enhance the model's ability to handle false premises and reasoning mismatches, we additionally use FP-RefCOCO(+/g) [25] and R-RefCOCO [24].

Evaluation Metrics. We evaluate segmentation performance based on two main criteria: semantic-level and instance-level accuracy. For semantic-level evaluation, we follow prior works [9, 20] and evaluate both global IoU (gIoU) and cumulative IoU (cIoU). The gIoU is computed as the average IoU across individual images, while the cIoU is calculated as the intersection over union aggregated over the entire dataset. Since cIoU is more sensitive to object size and

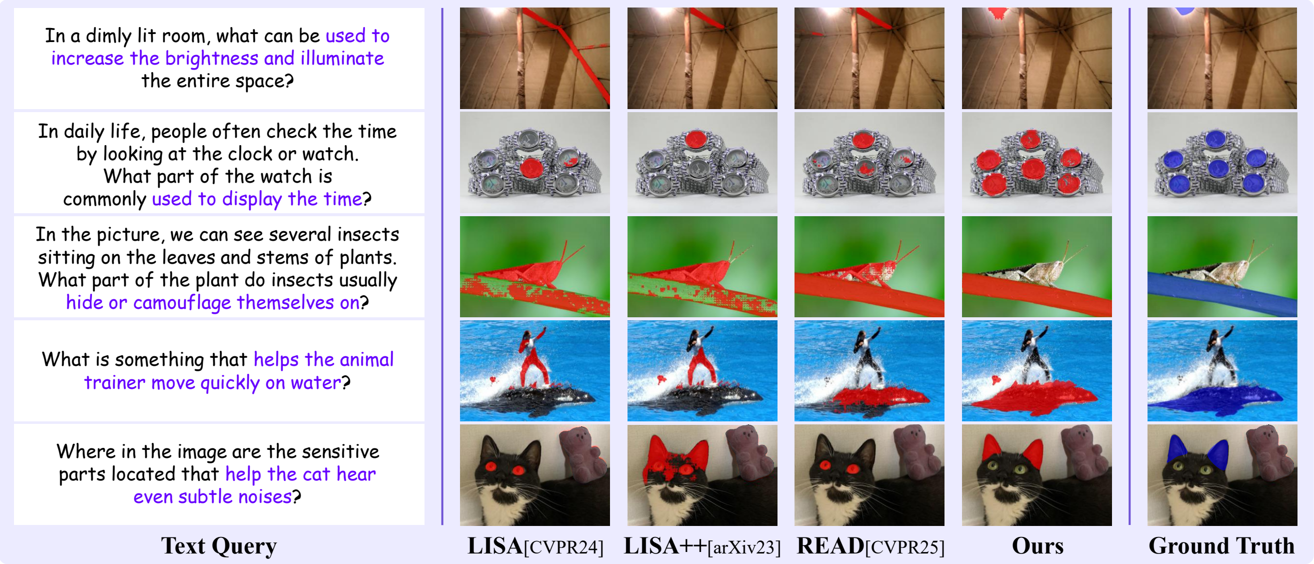
In a dimly lit room, what can be used to increase the brightness and illuminate the entire space?
In daily life, people often check the time by looking at the clock or watch. What part of the watch is commonly used to display the time?
In the picture, we can see several insects sitting on the leaves and stems of plants. What part of the plant do insects usually hide or camouflage themselves on?
What is something that helps the animal trainer move quickly on water?
Where in the image are the sensitive parts located that help the cat hear even subtle noises?
Text Query
LISA[CVPR24]
LISA++[arXiv23]
READ[CVPR25]
Ours
Ground Truth

Figure 5. Qualitative comparison of NOVO and existing methods on reasoning segmentation tasks. NOVO produces accurate and coherent masks, even in challenging reasoning cases such as ambiguous boundaries and multi-instance scenarios. More qualitative examples are provided in the supplementary materials.

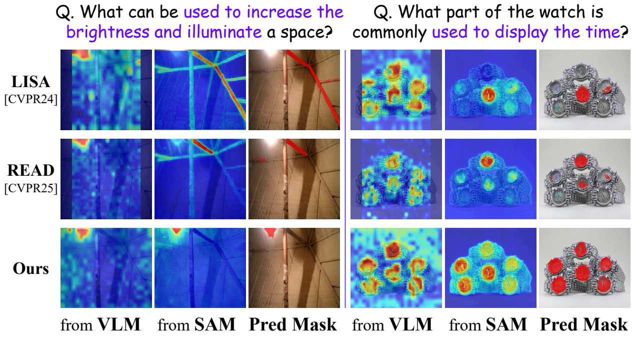
Q. What can be used to increase the brightness and illuminate a space?
Q. What part of the watch is commonly used to display the time?
LISA [CVPR24]
READ [CVPR25]
Ours
from VLM
from SAM
Pred Mask
from VLM
from SAM
Pred Mask

Figure 6. Comparison of VLM activation maps, SAM's mask logits, and predicted masks across different models.

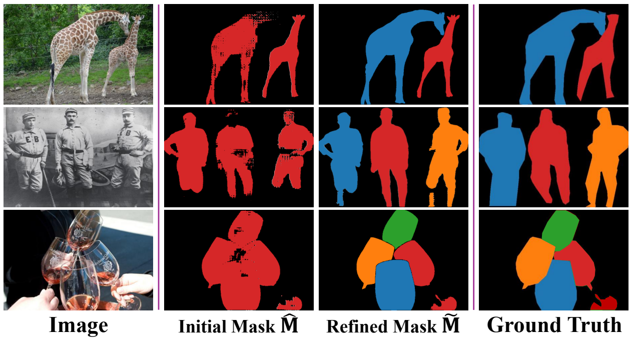
Image
Initial Mask $\hat{\mathbf{M}}$
Refined Mask $\tilde{\mathbf{M}}$
Ground Truth

Figure 7. Qualitative comparison of reasoning segmentation results. Different colors indicate different object instances.

may introduce bias, we focus on gIoU as it provides a more balanced evaluation across different object sizes. In addition, to assess the boundary quality of the refined masks introduced in Sec.4.5, we employ Boundary IoU and Boundary F1 metrics [5]. For instance-level evaluation, we measure the mean Average Precision (mAP), following [27].

6.2. Performance Comparison

Table 1 compares the reasoning segmentation performance of our proposed NOVO model with existing state-of-the-art methods. 7B and 13B refer to the VLM baseline models LLaVA 1.5-7B and LLaVA 1.5-13B, respectively. Ours-7B achieved 66.3% gIoU and 69.0% cIoU on the validation set, outperforming READ-7B by +6.5% and +1.4%, respectively. On the test set, it recorded 62.4% gIoU and 65.7% cIoU, showing improvements of +3.9% and +7.1% over READ-7B, while delivering performance comparable to larger 13B-scale models. Furthermore, Ours-13B achieved the highest performance among all methods, with 66.7% gIoU and 71.9% cIoU on the validation set, and 65.3% gIoU and 66.0% cIoU on the test set. These results highlight the effectiveness of our approach. Specifically, our method converts VLM outputs into visual prompts, allowing the segmentation model to generate more accurate and robust masks.

Figure 5 provides qualitative results of the NOVO model on complex reasoning segmentation tasks. Our model performs well in challenging cases, such as detecting objects with ambiguous boundaries (1st row) and identifying multiple instances simultaneously (2nd row). These results validate the effectiveness of NOVO in consistently generating query-aligned masks across diverse and complex reasoning scenarios. Additional qualitative examples are provided in the supplementary materials. Figure 6 presents a visual comparison of three stages: the coarse activation maps from the VLM, the mask logits from SAM, and the final predicted masks. We show results for LISA [9], READ [20], and our proposed method at each stage. While LISA and READ reasonably highlight relevant regions in the VLM activation maps, they often produce inaccurate masks at the SAM stage. In contrast, our method maintains consistency between the VLM's coarse activations and SAM's outputs, leading to more precise and semantically coherent masks. This demonstrates the effectiveness of our design in leveraging SAM's pretrained segmentation capabilities.

6.3. Effectiveness of NOVO Refinement

In this section, we assess the effectiveness of NOVO Refinement through both semantic-level and instance-level segmentation evaluations.

Table 2 compares semantic-level segmentation performance on the `ReasonSeg` dataset. Boundary-related metrics such as Boundary-IoU and Boundary-F1 were computed with a 3-pixel tolerance. On both the validation and test sets, our refined models (+R) achieved consistent improvements in gIoU, B-IoU, and B-F1. These results demonstrate that our refinement module effectively improves mask quality, particularly around object boundaries. Our NOVO Refinement also enables instance-level reasoning segmentation. To evaluate this capability, we conducted experiments on the `RISeg` dataset, as summarized in Tab. 3. We compared our method with LISA++ [27], a prior approach capable of instance-level reasoning. Regardless of the underlying baseline, our refined models (+R) consistently outperformed LISA++ across all AP metrics and object sizes.

Figure 7 presents a qualitative comparison of instance-level reasoning segmentation results before and after NOVO refinement. The initial mask $\hat{\mathbf{M}}$, obtained by simple thresholding the segmentation logits, often contained visual artifacts such as holes and inaccurate boundaries. In contrast, the refined masks $\widetilde{\mathbf{M}}$ effectively mitigate these artifacts through adaptive mask generation, resulting in more precise segmentation. Moreover, the proposed NOVO refinement leverages SAM's segmentation outputs, enabling instance-level segmentation as a natural outcome of the refinement process.

6.4. Ablation Study on Prompt Types

In this section, we investigate the effect of different prompt types through an ablation study on the `ReasonSeg` validation set, using gIoU and cIoU as evaluation metrics. All experiments adopt the pretrained LLaVA-based READ model [20] as the vision-language backbone and are trained under the same hyperparameters as our baseline.

Exp. ID	$\mathbf{P}_{\text{mask}}$	$\mathbf{P}_{\text{point}}$	gIoU	cIoU
1	✓		59.9	66.1
2		✓	55.8	52.3
3	✓	✓	**66.3**	**69.0**

Table 4. Ablation study on the use of mask and point prompts. Combined prompts yield the best scores in both gIoU and cIoU.

The results are summarized in Table 4. When a single type of prompt is used, mask prompts outperform point prompts (59.9 vs. 55.8 gIoU), suggesting that mask-level cues provide stronger spatial guidance to the segmentation model. Nevertheless, point prompts remain effective in highlighting key regions. The combination of mask and point prompts yields the best performance overall (66.3 gIoU and 69.0 cIoU), representing relative improvements of +10.5 and +16.7 points compared to the point-only setting. This demonstrates that mask prompts deliver fine-grained localization, while point prompts emphasize salient regions. Together, they play complementary roles in enhancing reasoning segmentation.

7. Conclusions and Future Work

In this study, we proposed NOVO, a new framework for reasoning segmentation that focuses solely on visual prompts, effectively bridging the gap between vision-language understanding and the visual prompt–based segmentation model SAM. NOVO generates a coarse mask and point prompts compatible with SAM's input format, avoiding semantic misalignment and enabling accurate segmentation. A refinement module further improves mask quality and naturally extends NOVO to instance-level segmentation, all without additional training. NOVO achieves new state-of-the-art results on both the `ReasonSeg` and `RISeg` benchmarks, demonstrating the effectiveness of visual prompt transformation for reasoning segmentation.

While NOVO demonstrates strong performance, it also has several limitations. First, NOVO Refinement operates with fixed thresholds and heuristics. Although effective in removing boundary noise and improving mask sharpness, it may struggle with complex errors. A learning-based refinement could offer a more adaptive solution. Second, our benchmark dataset `RISeg` is relatively small (918 images, 2,533 instance masks). While suitable for initial evaluation, its limited coverage of categories and reasoning types may restrict generalization. Future work could expand the dataset across more domains and reasoning dimensions.

As future work, we aim to extend NOVO to multi-label reasoning segmentation, where multiple relevant objects must be segmented and ranked according to a given query. For example, the query "Find the fruits with the highest vitamin content in order" requires both object identification and semantic ranking. This extension could further enhance the reasoning capability and applicability of NOVO.

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

[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. 6

[3] 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

[4] Yi-Chia Chen, Wei-Hua Li, Cheng Sun, Yu-Chiang Frank Wang, and Chu-Song Chen. Sam4mllm: Enhance multi-modal large language model for referring expression segmentation. In *European Conference on Computer Vision*, pages 323–340. Springer, 2024. 2

[5] Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C Berg, and Alexander Kirillov. Boundary iou: Improving object-centric image segmentation evaluation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15334–15342, 2021. 7

[6] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 5

[7] 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. 1, 5, 6

[8] 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. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 1, 2, 3, 5

[9] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024. 1, 2, 3, 4, 6, 8

[10] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Joshua Adrian Cahyono, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2

[11] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2

[12] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7061–7070, 2023. 2, 6

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[14] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23592–23601, 2023. 2, 6

[15] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 2, 3, 5, 6

[16] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024. 6

[17] Yi Lu, Jiawang Cao, Yongliang Wu, Bozheng Li, Licheng Tang, Yangguang Ji, Chong Wu, Jay Wu, and Wenbo Zhu. Rsvp: Reasoning segmentation via visual prompting and multi-modal chain-of-thought. *arXiv preprint arXiv:2506.04277*, 2025. 2, 6

[18] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 5, 6

[19] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 792–807. Springer, 2016. 1

[20] Rui Qian, Xin Yin, and Dejing Dou. Reasoning to attend: Try to understand how seg token works. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24722–24731, 2025. 1, 2, 3, 4, 6, 8

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[22] 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. 6

[23] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018, 2024. 2

[24] Jianzong Wu, Xiangtai Li, Xia Li, Henghui Ding, Yunhai Tong, and Dacheng Tao. Toward robust referring image segmentation. *IEEE Transactions on Image Processing*, 33: 1782–1794, 2024. 6

[25] Tsung-Han Wu, Giscard Biamby, David Chan, Lisa Dunlap, Ritwik Gupta, Xudong Wang, Joseph E Gonzalez, and Trevor Darrell. See say and segment: Teaching lmms to overcome false premises. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13459–13469, 2024. 2, 6

[26] Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3858–3869, 2024. 2

[27] Senqiao Yang, Tianyuan Qu, Xin Lai, Zhuotao Tian, Bohao Peng, Shu Liu, and Jiaya Jia. Lisa++: An improved baseline for reasoning segmentation with large language model. *arXiv preprint arXiv:2312.17240*, 2023. 2, 6, 7, 8

[28] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 2

[29] 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*, pages 633–641, 2017. 6

[30] 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. 2

[31] Lanyun Zhu, Tianrun Chen, Qianxiong Xu, Xuanyi Liu, Deyi Ji, Haiyang Wu, De Wen Soh, and Jun Liu. Popen: Preference-based optimization and ensemble for lvlm-based reasoning segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30231–30240, 2025. 2

[32] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15116–15127, 2023. 2, 6

[33] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in neural information processing systems*, 36:19769–19782, 2023. 2, 6