RSVP: Reasoning Segmentation via Visual Prompting and Multi-modal Chain-of-Thought

Yi Lu[1,2]*, Jiawang Cao[1]*, Yongliang Wu[1,3]*, Bozheng Li[1,4],
Licheng Tang[1], Yangguang Ji[1], Chong Wu[5], Jay Wu[1], Wenbo Zhu[1]

[1]Opus AI Research   [2]University of Toronto   [3]Southeast University
[4]Brown University   [5]City University of Hong Kong
tomlu@cs.toronto.edu   wenbo_zhu@berkeley.edu

Abstract

Multi-modal Large Language Models (MLLMs) have demonstrated remarkable reasoning capability while lacking explicit mechanisms for visual grounding and segmentation, creating a gap between cognitive reasoning and visual perception. To bridge this gap, we introduce **R**easoning **S**egmentation via **V**isual **P**rompting (RSVP), a novel framework that unifies multi-step multimodal reasoning with grounded visual understanding. RSVP is a two-stage structuralized framework that integrates reasoning-driven localization with segmentation refinement. In the reasoning stage, RSVP employs multimodal chain-of-thought visual prompts to help MLLMs understand queries and infer targets, generating interpretable region proposals that enhance visual grounding. In the segmentation stage, RSVP refines these proposals with a Vision-Language Segmentation Module (VLSM), which seamlessly integrates textual and visual cues to produce precise segmentation masks. By explicitly modeling the interaction between multimodal reasoning and segmentation, RSVP introduces a new paradigm for interpretable reasoning segmentation. It exploits MLLMs' inherent localization capabilities, enabling the models to not only reason about objects but also generate structured visual representations. Our extensive experiments demonstrate that RSVP achieves state-of-the-art performance, surpasses state-of-the-art methods by up to +6.5 gIoU and +9.2 cIoU on ReasonSeg, and achieves 49.7 mAP on SegInW under zero-shot settings. These results validate RSVP as an effective and scalable framework for integrating cognitive reasoning with structured visual understanding.
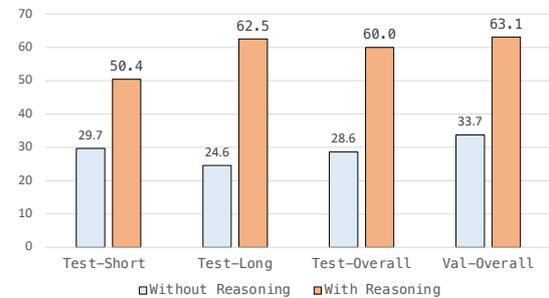
1 Introduction

Recent advances in multi-modal learning have enhanced MLLMs' ability to reason about visual content (Cao et al., 2024; Peng et al., 2025). However,
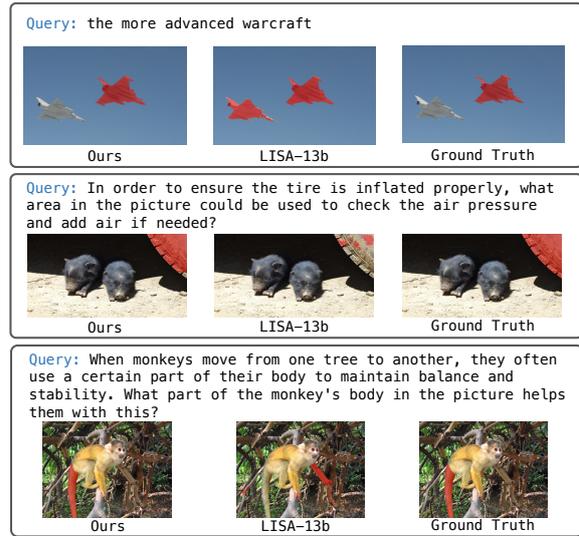
(a) Highlight of the comparison among the result of RSVP-GPT, the result of LISA-13B, and the ground truth mask.

(b) In reasoning segmentation tasks (cIoU result on ReasonSeg), while the segmentation model remains unchanged, enhancing reasoning results in significant improvements.

Figure 1: (a) and (b) depict different aspects of our segmentation pipeline performance. More demo results are available in Appendix C.

a key challenge remains unresolved: bridging the gap between cognitive reasoning and visual segmentation.

Reasoning Segmentation has emerged as a crucial task in multi-modal grounding, requiring models to produce segmentation masks from complex, implicit textual queries (Lai et al., 2024). Unlike traditional referring segmentation, which relies on explicit descriptions, reasoning segmentation demands models to infer whether a target object ex-

---

*Equal Contributions.

ists and where it is located using common sense knowledge and multi-step reasoning based on both textual and visual information. This makes it a significant step toward more interpretable and intelligent vision-language systems.

While Large Language Models (LLMs) excel at logical reasoning and contextual inference, they lack visual processing capabilities. MLLMs combine textual and visual modalities but remain incapable of generating precise segmentation masks (Das et al., 2024; Hu et al., 2024; Yang et al., 2023c). Conversely, referring segmentation models can identify object boundaries but struggle with high-level inference and reasoning, preventing them from effectively tackling reasoning segmentation. Existing solutions attempt to close this gap through fine-tuning large language-segmentation models on large-scale datasets, which is costly and impractical, or through heavy supervised training, which lacks scalability. While parameter-efficient tuning methods like LoRA (Hu et al., 2021) reduce computational costs, they still require substantial effort. Moreover, these models lack modularity—new architectures require retraining on large-scale data to benefit from performance improvements.

To address these challenges, we propose **RSVP**: **R**easoning **S**egmentation via **V**isual **P**rompting, a reasoning-driven multi-stage framework that unifies multi-modal chain-of-thought prompting with visual segmentation. Unlike prior methods that treat reasoning and segmentation as separate components, RSVP explicitly models their interaction, enabling MLLMs to generate interpretable, stepwise region proposals that bridge reasoning and segmentation within a modular framework following a two-stage pipeline:
Reasoning Stage. By introducing Multi-modal Chain-of-Thought Visual Prompting, MLLMs are guided to understand queries, infer object properties, reason about existence, and generate region proposals, enabling explicit visual grounding.
Segmentation Stage. These coarse proposals are refined using a Vision-Language Segmentation Module (VLSM), which integrates textual and visual cues to produce precise segmentation masks.

By integrating structured reasoning with segmentation, RSVP enables MLLMs to reason about objects while producing explainable visual representations. Experiments on ReasonSeg (Lai et al., 2024) and SegInW (Zou et al., 2023) show state-of-the-art zero-shot performance, surpassing zero-shot

baselines by +6.5 gIoU, +9.2 cIoU, and achieving 49.7 mAP. Notably, RSVP consistently outperforms baselines on both open-source and closed-source LLM foundations, demonstrating strong generalization.

Our contributions are summarized as follows:

- We propose a reasoning-driven multi-stage framework that leverages MLLMs' reasoning capabilities for explicit region proposal generation, significantly reducing training costs while improving interpretability.

- We develop a multi-modal chain-of-thought prompting paradigm that bridges the gap between reasoning and segmentation by producing object properties, explainable rationales, and structured region proposals.

- A joint text-image segmentation model is developed following our design, that achieves state-of-the-art results on ReasonSeg (gIOU=60.3, cIOU=60.0) and SegInW (mAP=49.7) under zero-shot settings, demonstrating its effectiveness in both reasoning and open-world segmentation.

2 Related Work

2.1 Reasoning Segmentation

First introduced by LISA (Lai et al., 2024), Reasoning Segmentation extends referring segmentation by requiring models to reason over implicit queries. Unlike traditional referring segmentation which directly identifies objects based on simple descriptions (e.g., *"The brown dog at the front"*), reasoning segmentation involves abstract contextual inference (e.g., *"What area in the picture could lead to other parts of the garden?"*). This demands models to integrate visual understanding, object properties, common sense and world knowledge.

2.2 Multi-modal Large Language Models

MLLMs combine vision encoders with language models to process multi-modal inputs, bridging textual and visual tasks (Zhang et al., 2024a; Wu et al., 2024c; Zhang et al., 2025). Open-source models such as LLaVA (Liu et al., 2023b) and Mini-GPT4 (Zhu et al., 2023) demonstrate remarkable generalization on downstream tasks, while proprietary systems like GPT-4o (OpenAI et al., 2024) and Gemini (Team et al., 2024) push the limits of multi-modal cognition (Li et al., 2025).
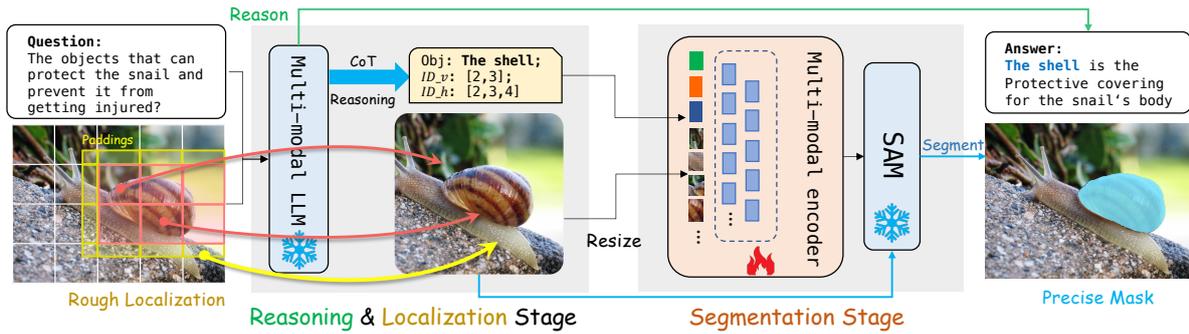
Figure 2: Overview of the proposed model. An input image is divided into horizontal and vertical regions to assist localization. In the reasoning stage, an MLLM receives a query about the object's protective features and identifies "the shell" as the protective object, generating region proposal using region IDs (id_v and id_h). Red boxes indicate the regions of interest determined by the MLLM, yellow box denotes the padding p for complete visual content. The CoT process enhances reasoning accuracy. In the segmentation stage, a multi-modal encoder integrates textual and visual information, resizing the image for detailed feature extraction. Finally, SAM refines the segmentation by highlighting the shell that acts as a protective covering for the snail.

MLLMs have been applied in vision-language reasoning tasks, including LISA (Lai et al., 2024), Kosmos-2 (Peng et al., 2024), and Groundhog (Zhang et al., 2024b), where models generate region proposals or segmentation masks. However, these methods either require extensive training or rely on complex architectures with large parameter sizes. Our approach differs by introducing a lightweight reasoning-driven framework that exploits MLLMs' innate reasoning and localization capabilities without additional training.

2.3 Visual Prompting

Inspired by prompt engineering in Natural Language Processing (NLP), visual prompting (Wu et al., 2024a; Yang et al., 2023d) modifies the input space using human-perceivable markers such as bounding boxes (Lin et al., 2024), numbers (Yang et al., 2023a), or shapes (Shtedritski et al., 2023). These cues help MLLMs focus on key image regions without altering model parameters, mitigating issues like visual hallucination (Bai et al., 2024) and language bias (Qu et al., 2024).

Visual prompting has shown effectiveness in fine-grained visual attention for tasks such as referring expressions (Shtedritski et al., 2023), visual question-answering (VQA) (Zhou et al., 2024) and video localization (Wu et al., 2024b). Det-ToolChain (Wu et al., 2024a) integrates CoT reasoning (Wei et al., 2022) with visual prompts to enhance object detection using MLLMs like GPT-4V and Gemini.

To our knowledge, no previous work has explored visual prompting for reasoning segmentation. Unlike prior approaches that focus on basic spatial attention, we integrate multi-modal chain-of-thought visual prompting to provide structured reasoning for segmentation, allowing models to generate interpretable region proposals.

2.4 Multi-modal Chain of Thought

Motivated by recent advances in LLMs, Multi-modal Chain-of-Thought (CoT) has become a prominent approach for enhancing MLLMs' reasoning capabilities. A line of work (Zhang et al., 2023) extends CoT reasoning to Vision-Language tasks by introducing a two-stage framework that separates reasoning chain generation and answer inference. M3-CoT, introduced by Chen et al. (2024) as a comprehensive benchmark dataset, provides rich multi-step, multi-modal samples of mathematical and scientific problems. Additionally, Shao et al. (2024) proposes VisCoT, a multi-round interactive image understanding pipeline that mimics human localized focus to extract key information, improving performance on VQA and document comprehension tasks. Literature Chen et al. (2025) Comprehensively surveys long CoT applications of MLLMs in areas such as mathematics, science and commonsense puzzles. However, application of Multi-modal CoTs of reasoning segmentation tasks remain underexplored.

2.5 Text-Prompted Segmentation

Text-prompted segmentation (referring segmentation) involves extracting object segmentation masks based on natural language queries. Transformer-based models like SAM (Kirillov et al., 2023) leverage CLIP (Radford et al., 2021) embeddings for segmentation, while SAM-CLIP (Wang et al., 2023) fuses the visual backbones of SAM and CLIP. Grounded-SAM (Ren

et al., 2024) employs Grounding-DINO (Liu et al., 2023c) to detect objects before refining segmentation with SAM.

However, these models lack a structured knowledge summarization and reasoning process between text embedding and segmentation. Moreover, they are trained on short, explicit queries, limiting their ability to handle abstract and implicit reasoning. In contrast, we incorporate multi-modal reasoning within segmentation models, enabling superior performance on complex, implicit queries. As shown in Figure 1, by integrating MLLMs for reasoning and rough localization region proposals while retaining similar segmentation models, we achieve superior performance compared to non-reasoning text-prompted segmentation.

3 Method

3.1 Overview

We aim to develop an efficient, modular reasoning segmentation framework that integrates MLLM's inherent cognitive reasoning capabilities with structured visual segmentation while minimizing the need for fine-tuning. To achieve this goal, we introduce RSVP, a two-stage framework consisting:

1) Multi-modal Chain-of-Thought (CoT) Visual Prompting, which enables an MLLM to reason about object attributes and generate interpretable region proposals.

2) Vision-Language Segmentation Module (VLSM), which refines these proposals into precise segmentation masks.

Unlike existing approaches that rely on extensive fine-tuning (Lai et al., 2024), RSVP exploits MLLMs' intrinsic reasoning and localization capabilities through structured CoT visual prompting. This enables zero-shot segmentation while improving interpretability.

3.2 Multi-modal Chain-of-Thought Visual Prompting

Region Proposal Initialization. To generate structured region proposals, we simplify reasoning segmentation into a text-guided localization task, where an MLLM identifies the object of interest and its approximate location. Unlike prior works that require model fine-tuning, our method directly leverages MLLMs' reasoning and rough localization abilities via visual prompting.

Inspired by Set-of-Mark (SoM) prompting (Yang et al., 2023a), we introduce a region-aware visual
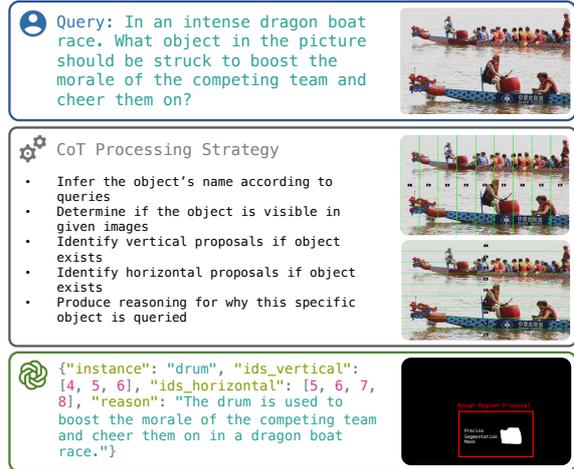

Figure 3: Illustration of the CoT processing strategy in action for a query about a dragon boat race.

prompt to explicitly structure spatial queries. As illustrated in Figure 3, we preprocess the input image I by dividing it into M horizontal and N vertical sections, assigning unique region IDs. The reasoning process explains why "drum" is the object that matches the query: "boosts team morale". Region IDs for localization are determined as (ids_v: [4, 5, 6], ids_h: [5, 6, 7, 8]). The rough region proposal highlights the drum's location in the image as a crucial target. The determined region and inferred object name are passed to the second-stage segmentation model to produce the precise mask, detailed prompt inputs are shown at Appendix D. The structured prompt consists of:

(1) A reasoning-based textual query guiding the MLLM to infer object attributes.

(2) A region-aware localization task, where the MLLM assigns region IDs to detected objects.

(3) A CoT-based reasoning step, which provides explicit rationale behind the object's presence and location.

The MLLM outputs structured proposals containing: (1) Object name and attributes: A textual description (e.g., *"a red drum"*). (2) Region localization: A set of horizontal and vertical region IDs. (3) Reasoning rationale: Justification for why the object belongs to the detected region.

Region Proposal Formulation. Given the MLLM-predicted horizontal (id_h) and vertical (id_v) region indices, we construct the bounding horizontal / vertical regions, R_h and R_v:

$$R_h = [\max(\min(\text{id}_h), 1) - p_y, \min(\max(\text{id}_h), M) + p_y]$$
$$R_v = [\max(\min(\text{id}_v), 1) - p_x, \min(\max(\text{id}_v), N) + p_x]$$

where p_x, p_y are padding terms ensuring the bounding box does not truncate object edges.

3.3 Multi-modal Chain-of-Thought Reasoning

While MLLMs excel at multimodal reasoning, they lack structured segmentation capabilities. To address this gap, we design a CoT-based visual prompt that guides MLLMs through a step-wise progressive reasoning process, ensuring accurate object identification and localization.

CoT Reasoning Process. Given an input image I and textual query T_q, the MLLM generates a structured reasoning output:

$$T_a = \mathrm{MLLM}(I, T_q)$$

Unlike single-step predictions, we structure the query into explicit step-wise CoT prompts, which require the MLLM to:

(1) Infer the most probable object class (e.g., *"drum"*). (2) Identify key attributes (e.g., *"red, cylindrical"*). (3) Locate objects within image regions using visual referencing with region IDs. (4) Provide rationale explaining how and why the inference was made.

This explicit reasoning improves interpretability while ensuring robust region proposals. If the model found the object of interest is absent from the image, it returns an empty region list as well as its rationale.

3.4 Vision-Language Segmentation Module

Once the object description and approximate location are obtained, the task reduces to referring segmentation, where we generate a fine-grained mask. To achieve this, we design a Vision-Language Segmentation Module (VLSM) that integrates textual descriptions and localized image crops.

Region Cropping. We extract a cropped region I' centered around the bounding box:

$$H' = \frac{H}{M} \times (\mathrm{end}_v - \mathrm{start}_v + 1) + 2p_y$$

$$W' = \frac{W}{N} \times (\mathrm{end}_h - \mathrm{start}_h + 1) + 2p_x$$

Where $\frac{H}{M}$ / $\frac{W}{N}$ is the height / width of each vertical region. $(\mathrm{end}_v - \mathrm{start}_v + 1)$ and $(\mathrm{end}_h - \mathrm{start}_h + 1)$ calculate the number of units included in the cropped sections, p_x and p_y refers to horizontal and vertical padding.

Multi-modal Feature Encoding. We build upon the SAM framework (Kirillov et al., 2023) by creating a module for integrated visual-language feature extraction. Since Segment Anything Model

(SAM) (Kirillov et al., 2023) lacks native text support, we incorporate BEiT-3 (Wang et al., 2022) as a joint vision-language encoder.

First, the cropped image I' is resized to $I_r \in \mathbb{R}^{224 \times 224 \times 3}$. The target description T_a, tokenized with XLMRobertaTokenizer (Rachmadi et al., 2023), is integrated into the BEiT-3 model. Image and text features are processed independently, producing $\mathbf{F}_{\mathrm{image}} \in \mathbb{R}^{(N+1) \times D}$ and $\mathbf{F}_{\mathrm{text}} \in \mathbb{R}^{L \times D}$. These embeddings are combined via cross-attention in Transformer layers, yielding a unified token for cross-modal representation. It is then processed through a projector with two linear layers and ReLU, mapping it to a 256-dimensional space, which is sent to SAM's prompt encoder and segmentation decoder, which predicts the pixel-wise segmentation mask.

3.5 Inference Pipeline and Computational Efficiency

During inference, RSVP operates in two steps. MLLM first generates structured region proposals (CoT-based Region Proposal Generation). Identified regions are then processed by VLSM to generate the final mask, corresponding to the Vision-Language Segmentation step. We compared the inference latency on the ReasonSeg-Test split (comprising 754 images) using our pipeline (7B models) and LISA. The detailed results are as follows:

Model	Stage	Time (s)	Model Size
RSVP-LLaVA	First-stage	9.20	7B
	Second-stage	0.77	
	Total	9.97	
RSVP-Qwen	First-stage	9.35	7B
	Second-stage	0.72	
	Total	10.07	
RSVP-Qwen	First-stage	7.36	2B
	Second-stage	0.73	
	Total	8.09	
LISA-13B	Total	15.88	13B
LISA-7B	Total	9.41	7B

Table 1: Average inference time (seconds per image) on the ReasonSeg-Test split. RSVP's latency is broken down into two stages. LISA is a single-stage model, only total time is reported.

Inference time was measured after loading each model onto a single A100 40GB GPU. RSVP-7B exhibited a worst-case latency of 10.07 seconds per image, compared to 9.41 seconds for LISA-7B and 15.88 seconds for LISA-13B. This represents only a $\sim 7\%$ increase over LISA-7B, a marginal overhead considering RSVP's added capabilities.

The primary computational cost originates from the firsr-stage MLLM inference. In real-world applications, this overhead can be further reduced through optimization techniques such as quantization or deployment frameworks such as vLLM (Kwon et al., 2023). Importantly, unlike end-to-end fine-tuning methods, RSVP requires no additional training, making it computationally efficient for reasoning-based segmentation while maintaining inference speed comparable to existing approaches.

4 Experiments

We evaluate RSVP through quantitative and qualitative analysis to assess its performance in reasoning segmentation, open vocabulary segmentation, and referring segmentation. The section is organized as follows. Experiment Setup is described in Sec. 4.1, Details of the implementation are explained in Sec. 4.2, Sec. 4.3 presents evaluation results. In addition, the ablation study is conducted concerning various design choices in Sec. 4.4.

4.1 Evaluation Setup

We apply the following datasets and evaluation metrics for the following experiments:

Datasets. For reasoning segmentation, we conduct our experiments on the ReasonSeg validation and testing dataset, which is proposed by LISA (Lai et al., 2024). Specifically, the ReasonSeg-Val dataset consists of 200 samples, while between different splits of the ReasonSeg-Test dataset (around 770 samples), "Short Query" refers to reasoning segmentation annotations with a single short sentence, while in the "Long Query" split, each referring annotation contains multiple long sentences. To demonstrate RSVP's generality in the zero-shot open-world segmentation task, we evaluate our model's performance on the Segmentation in the Wild (SegInW) zero-shot benchmark (Zou et al., 2022), which comprises 25 zero-shot in-the-wild segmentation datasets. refCOCOg (Mao et al., 2015) is utilized for testing the referring expression segmentation task on our second-stage model: Visual-Language Segmentation Module (VLSM) to demonstrate its capability in this task.

Metrics. Following previous works (Lai et al., 2024; Yang et al., 2023b), we apply Generalized IoU (gIoU) and Cumulative IoU (cIoU) as the performance evaluation metric in referring segmentation, as well as the reasoning segmentation task. Generalized IoU (gIoU) is computed as the average IoU over all images in the test set:

$$\text{gIoU} = \frac{1}{N} \sum_{i=1}^{N} \text{IoU}^{(i)} = \frac{1}{N} \sum_{i=1}^{N} \frac{I^{(i)}}{U^{(i)}}$$

while Cumulative IoU (cIoU) is defined as the ratio of the cumulative intersection over the cumulative union across all images:

$$\text{cIoU} = \frac{\sum_{i=1}^{N} I^{(i)}}{\sum_{i=1}^{N} U^{(i)}}$$

Where: N is the number of images in the test set, $B_{\text{pred}}^{(i)}$ and $B_{\text{gt}}^{(i)}$ are the predicted and ground truth bounding boxes for the i-th image, $I^{(i)}$ is their intersection area, and $U^{(i)}$ is their union area.

For open-world segmentation, we applied mean average precision (mAP) as the evaluation metric following the convention of the SegInW benchmark and previous works.

4.2 Implementation Details

Our two-stage model combines a zero-shot prompted MLLM for reasoning and a referring segmentation model for mask generation.

First-stage MLLMs. For evaluating commercial MLLMs, we employ GPT-4o and Gemini-Flash[1]. For open-source MLLMs, we apply LLaVA-NeXT and Qwen2-VL[2].

Second-stage Segmentation Model. The segmentation model is trained on refCOCO, initialized with BEiT-3 and SAM weights. The model training uses LoRA and DeepSpeed (Rasley et al., 2020), with AdamW optimizer (Loshchilov and Hutter, 2017) with a learning rate of 1e-4 alongside Dice Loss and Binary Cross-Entropy (BCE) Loss. Training involves around 16,000 epochs with a total batch size of 256.

4.3 Main Results

Reasoning Segmentation. RSVP is extensively tested on ReasonSeg's validation and test set to demonstrate its effectiveness. Table 2 compares RSVP with conventional zero-shot methods, LISA models which are further fine-tuned on ReasonSeg's training split, and LISA models which are not further trained on any reasoning segmentation dataset.

---

[1]GPT-4o version: gpt-4o-2024-08-06, Gemini-Flash version: gemini-1.5-flash-002.

[2]Obtained 7B version model weights from their Hugging-Face pages.

| Method | MLLM | Val | | Test | | | | | |
| | | Overall | | Short Query | | Long Query | | Overall | |
		gIOU	cIOU	gIOU	cIOU	gIOU	cIOU	gIOU	cIOU
OVSeg(Wang et al., 2024)	-	28.5	18.6	18.0	15.5	28.7	22.5	26.1	20.8
SEEM(Zou et al., 2024)	-	25.5	21.2	20.1	11.5	25.6	20.8	24.3	18.7
Grounded-SAM(Ren et al., 2024)	-	26.0	14.5	17.8	10.8	22.4	18.6	21.3	16.4
LLaVA-7B + OVSeg	LLaVA-7B	38.2	23.5	24.2	18.7	44.6	37.1	39.7	31.8
LLaVA-7B + CoT + OVSeg	LLaVA-7B	42.1	26.6	27.5	19.6	44.6	40.2	40.4	34.0
LISA(Lai et al., 2024)-7B	LLaVA-7B	53.6	52.3	47.1	48.5	49.2	48.9	48.7	48.8
LISA(Lai et al., 2024)-13B	LLaVA-13B	57.7	60.3	50.8	50.0	54.7	50.9	53.8	50.8
LISA(Lai et al., 2024)-7B (ft)	LLaVA-7B	61.3	62.9	48.3	46.3	57.9	59.7	55.6	56.9
LISA++(Yang et al., 2023b)-7B (ft)	LLaVA-7B	64.2	68.1	49.6	51.1	59.3	61.7	57.0	59.5
LISA(Lai et al., 2024)-13B (ft)	LLaVA-13B	65.0	72.9	55.4	50.6	63.2	65.3	61.3	62.2
RSVP-LLaVA	LLaVA-7B	59.2	56.7	47.9	42.0	58.4	53.0	55.9	50.7
RSVP-Qwen	Qwen2-VL-7B	58.6	48.5	48.5	44.3	57.1	53.8	56.6	51.6
RSVP-Gemini	Gemini-Flash	56.9	49.2	47.3	40.2	60.2	**65.6**	57.1	59.2
RSVP-GPT	GPT-4o	**64.7**	**63.1**	**55.4**	**50.4**	61.9	62.5	60.3	**60.0**

Table 2: Reasoning segmentation results of our model and previous related works on ReasonSeg(Lai et al., 2024). "ft" denotes that the model was fine-tuned on ReasonSeg's training split, others are tested under zero-shot. "CoT" means utilizing region proposal Chain-of-Thought strategy. Results in bold are the best metric among zero-shot models. Columns with gray backgrounds indicate training-free methods, while grayed-out columns are LISA models fine-tuned on ReasonSeg Training split.

The first 4 rows represent traditional zero-shot methods, which perform significantly worse than zero-shot or fine-tuned LISA models, and our approaches. It is worth mentioning that applying our RSVP framework to weaker models (e.g., LLaVA-7B + Multi-Modal CoT Visual Prompt + OVSeg) notably improves segmentation results.

Despite using models with comparable or larger parameter sizes, both 7B and 13B versions of LISA that are not fine-tuned on ReasonSeg-Training split underperform against RSVP, demonstrating the effectiveness of reasoning-driven segmentation. Although LISA models can be further fine-tuned on ReasonSeg to achieve competitive performance, they require substantial computational resources and retraining whenever new foundation models emerge. In contrast, RSVP achieves state-of-the-art gIoU and cIoU in zero-shot settings, outperforming fine-tuned models without additional training.

| Method | refCOCOg | |
	val(cIoU)	test(cIoU)
MCN (Luo et al., 2020)	49.2	49.4
VLT (Ding et al., 2021)	55.0	57.7
CRIS (Wang et al., 2021)	59.9	60.4
LAVT (Yang et al., 2021)	61.2	62.1
ReLA (Liu et al., 2023a)	65.0	66.0
X-Decoder (Zou et al., 2022)	64.6	-
SEEM (Zou et al., 2024)	65.7	-
LISA (Lai et al., 2024)	66.4	68.5
RSVP (VLSM only)	**65.5**	**66.4**

Table 3: cIoU metric report on refCOCOg dataset's validation and test split. "-": data is not reported by original work. Our result: highlighted with gray background and bold text.

Referring Segmentation. We further evaluate VLSM on refCOCOg (Table 3). RSVP achieves competitive results and demonstrates segmentation model remains effective in standard referring seg-

mentation tasks.

Open-World Segmentation.

Table 4 shows the result in SegInW. Although the reasoning part of the Segmentation in the Wild task is relatively straightforward, the target categories are more diverse, allowing us to validate the generalization capabilities of RSVP. Our method achieved a mean average precision (mAP) of 49.7 across 25 categories, demonstrating the effectiveness of our proposed approach.

4.4 Ablation Study

We conduct ablation studies to analyze design choices in RSVP's CoT reasoning process.

First-stage Multimodal Information Distillation. Table 2 compares different MLLMs. GPT-4o outperforms LLaVA by +6.4 cIoU, +5.5 gIoU, and Qwen2-VL by +14.6 cIoU, +6.1 gIoU, showing that stronger reasoning models yield better segmentation performance. However, Gemini-Flash does not outperform LLaVA, indicating that applying our multi-modal chain-of-thought region proposal with MLLMs which are stronger in reasoning, comprehending, and distilling information conveyed in multi-modal inputs is crucial for achieving powerful performance in reasoning segmentation tasks.

Distillation Modality	GPT-4o	LLaVA
Visual Only	33.7 (−29.6)	31.2 (−25.5)
Text Only	56.6 (−6.5)	53.5 (−3.2)
Both Modalities	**63.1**	**56.7**

Table 5: Comparison of our model's cIoU metric on ReasonSeg-Val split utilizing different Multi-Modal Information Distillation Approaches. Values in brackets indicate the amount of cIoU decrease of this approach compared to the "Both Modalities" approach.

Method	mean SegInW	Elephants	Hand-Metal	Watermelon	House-Parts	HouseHold-Items	Strawberry	Fruits	Nutely-Squirrel	Hand	Garbage	Chicken	Rail	Airplane-Parts	Brain-Tumor	Poles	Electric-Shaver	Bottles	Toolkits	Trash	Salmon-Fillet	Puppies	Tablets	Phones	Cows	Ginger-Garlic
X-Decoder-T (Zou et al., 2022)	22.6	65.6	22.4	16.2	5.5	50.6	41.6	66.5	62.1	0.6	28.7	12.0	0.7	10.5	1.1	3.6	1.2	19.0	9.5	19.3	15.0	48.9	15.2	29.9	12.0	7.9
X-Decoder-L-IN22K	26.6	63.9	20.3	13.5	4.9	50.5	74.4	79.1	58.8	0.0	24.3	3.5	1.3	12.3	0.5	13.4	18.8	43.2	14.6	20.1	12.3	57.3	6.9	43.4	12.3	15.6
X-Decoder-B	27.7	68.0	18.5	13.0	6.7	51.7	81.6	76.7	53.1	20.6	30.2	13.6	0.8	13.0	0.3	5.6	4.2	45.9	13.0	27.3	18.2	55.4	8.0	8.9	36.8	19.4
X-Decoder-L	32.2	66.0	42.1	13.8	7.0	53.0	67.1	79.2	68.4	75.9	33.0	8.6	2.3	13.1	2.2	20.1	7.5	42.1	9.9	22.3	19.0	59.0	22.5	15.6	44.9	11.6
ODISE-L (Xu et al., 2023)	38.7	74.9	51.4	37.5	9.3	60.4	79.9	81.3	71.9	41.4	**39.8**	84.1	2.8	15.8	2.9	0.4	18.3	37.7	15.8	28.6	30.2	65.4	9.1	43.8	41.6	23.0
UNINEXT-H (Yan et al., 2023)	42.1	72.1	57.0	56.3	0.0	54.0	80.7	81.1	**84.1**	**93.7**	16.9	75.2	0.0	15.1	2.6	13.4	71.2	46.1	10.8	**44.4**	64.6	64.6	21.0	6.1	**52.7**	23.7
Grounded-SAM (B+H)	48.7	77.9	**81.2**	64.2	8.4	60.1	**83.5**	82.3	71.3	70.0	24.0	84.5	8.7	37.2	11.9	23.3	71.7	65.4	20.4	30.0	32.9	50.1	29.8	35.4	47.5	**45.8**
Grounded-SAM (L+H)	46.0	78.6	75.2	61.5	7.2	35.0	82.5	**86.9**	70.9	90.7	22.8	**84.6**	7.2	38.4	10.2	17.4	59.7	43.7	**26.3**	22.4	27.1	63.2	38.6	3.4	49.4	40.0
RSVP	**49.7**	**84.7**	61.6	**69.1**	**42.3**	**90.7**	81.6	84.3	79.6	90.2	34.7	82.3	**34.1**	**61.2**	**13.4**	**52.4**	**75.6**	**83.8**	12.6	41.1	**76.4**	**91.7**	**70.6**	**45.7**	45.1	45.6

Table 4: mAP reported on SegInW Open-world segmentation dataset. Metrics in bold represent the top 1 result for each subtask, our result is highlighted with the gray background.

Multimodal Information Distillation Strategy. We explore different approaches to distilling multimodal input using MLLMs for reasoning segmentation tasks. Specifically, we adopt different methodologies on our model with two representative MLLMs as the first-stage model, GPT-4o and LLaVA: (a) The MLLM comprehends and distills the input's textual modality only. (b) The MLLM comprehends and distills the input's visual modality only. (c) The MLLM comprehends and distills both of the input's modalities.

Results in Table 5 show that excluding textual reasoning leads to a 29.6 cIoU drop, while omitting visual information decreases performance by 6.5 cIoU, while distilling both modalities yields the most superior performance in reasoning segmentation, confirming that reasoning about object attributes as well as generating rough object grounding proposal is crucial for reasoning segmentation.

VLSM Combination	gIoU (%)	cIoU (%)
RSVP-OVSeg	43.5	37.2
RSVP-LLaVA	55.9	50.7
RSVP-GPT	**60.3**	60.0
RSVP-GPT (ft)	57.5	**61.6**

Table 6: Ablation study on the VLSM Module in RSVP, evaluated on the ReasonSeg Test split. Best result rows are highlighted in gray.

Importance of Reasoning and Segmentation Modules. We examined the importance of the modules of both stages in the performance of the reasoning segmentation task. Table 6 demonstrates the ablation result, where RSVP-GPT (ft) refer to finetuned VLSM using GPT-4o's reasoning on ReasonSeg-Train, RSVP-GPT and RSVP-LLaVA are zero-shot models trained on refCOCOg. RSVP-LLaVA uses LLaVA-7B's reasoning for inference and RSVP-GPT (ft) and RSVP-GPT use GPT-4o's reasoning on ReasonSeg-Test. Replacing VLSM with OVSeg reduces cIoU by -22.8 on ReasonSeg-Test, aligning with OVSeg's weaker referring seg-

Visual Modality Distill Strategy	LLaVA	GPT-4o
9×9 Grid	53.1	56.8
5×5 Split	54.3	59.6
9×9 Split	**56.7**	**63.1**
13×13 Split	52.3	57.6
No Visual Prompt	53.5	56.6

Table 7: Visual Modality Distillation Strategy. **Grid**: Grid visual prompt. **Split**: Region-aware visual prompting with varying densities (5×5, 9×9, 13×13 regions). Best result is highlighted in gray background and bold text.

mentation capability. Further, RSVP with weaker reasoning models (LLaVA vs. GPT-4o) underperforms despite an identical segmentation module, emphasizing that a strong MLLM is essential for strong, top-tier performance.

Visual Modality Distillation Strategy. We explore two types of visual prompting strategies to assist in visual modality distillation: (A) Grid-based Visual Prompt. (B) Region-aware Visual Prompt. For (B), we further experimented with three densities: (a) 5 horizontal and vertical regions, (b) 9 horizontal and vertical regions, and (c) 13 horizontal and vertical regions. The tests were carried out on two representative models: open-source LLaVA and close-source GPT4o. Neither of them is trained in comprehending our region-aware horizontal/vertical separation visual prompts or grid visual prompts. Results in Table 7 shows:

(1) With the appropriate density, 9×9 being set, region-aware visual prompting led to notable improvements in cIoU, which is +3.2 cIoU for LLaVA and +6.5 cIoU for GPT-4o.

(2) Visual markers' density may strongly impact models' performance. Too many visual markers degrade performance (e.g., 13×13 results in up to -6.3 cIoU drop), which could be explained as overdetailed visual prompts exceed MLLMs' capacity for recognizing fine details and impact the semantic information's quality obtained by the model, while too few markers fail to filter irrelevant regions, lead-

Multi-Modal CoT Design	RSVP-LLaVA	RSVP-GPT
Prompt (A)	**50.7**	**60.0**
Prompt (B)	45.5	55.7

Table 8: Multi-Modal Chain-of-Thought Prompt Design. **(A)**: manual-crafted multi-step prompt. **(B)**: simple CoT prompt. Best result rows are highlighted in gray, while the best results are highlighted in bold.

ing to segmentation accuracy declines.

Multi-Modal Chain-of-Thought Prompting Design. An ablation test on CoT prompt design is conducted to demonstrate the importance of manually designed Chain-of-Thought prompts. RSVP's cIoU results on ReasonSeg dataset's test split across different models are examined with two distinct prompt designs:

(A) A hierarchical Chain-of-Thought prompt instructing the model to comprehend both intricate queries and provided images in a structuralized, step-by-step manner.

(B) A plain Chain-of-Thought prompt describing the overall problem, and provide plain instructions on generating structuralized answers.

As Table 8 shows, both models demonstrated stronger performance on the human-designed CoT prompt, while a suboptimal prompt has a noticeable negative impact on model's reasoning segmentation performance. For both models, significant decreases in cIoU are observed, which is -5.2 cIoU for RSVP-LLaVA and -4.3 for RSVP-GPT. This finding suggests that carefully designed reasoning principles could activate models' inherent inference capability and world knowledge, providing more robust and credible answers.

In conclusion, these results highlight the importance of balanced visual prompts for optimal multi-modal grounding.

5 Conclusion

In this work, we identified that the core challenge of reasoning segmentation is query comprehension and object localization. To address this, we introduced RSVP, a two-stage multi-modal reasoning framework that leverages MLLMs' intrinsic reasoning and visual localization capabilities through Multi-Modal Chain-of-Thought Visual Prompting. By explicitly modeling the interaction between step-wise reasoning and segmentation, RSVP achieves state-of-the-art performance on zero-shot reasoning segmentation tasks.

Apart from performance improvements, we demonstrate the potential of integrating multi-

modal reasoning with visual localization, providing new insights into bridging the gap between cognitive inference and fine-grained visual perception. We hope our work could lay the foundation for future research on reasoning segmentation, visual prompting, and broader integration of MLLMs in vision-language tasks.

6 Limitations

Although RSVP effectively integrates reasoning-driven object localization with structured segmentation, several challenges remain.

Dependence on MLLMs. RSVP relies on MLLMs for reasoning, making its performance sensitive to the capabilities of the underlying model. While stronger models like GPT-4o produce high-quality region proposals and nuanced query understanding, weaker or smaller MLLMs may struggle with complex reasoning and detailed visual interpretation. Future work could explore model distillation techniques or mixture-of-experts to enhance performance with lightweight models.

Visual Prompting Strategy. While our region-aware visual prompting effectively guides MLLMs for spatial reasoning, the optimal visual prompt design design remains an open question. Future directions include discovering diverse visual prompt designs or fine-tuned visual prompting mechanisms to further boost MLLMs' spatial understanding and visual grounding capabilities.

Computational Overhead. RSVP eliminates the need for fine-tuning while multiple processing steps are still involved, leading to potential latency in real-time applications. Further work could investigate efficient model compression techniques to improve inference speed.

Data Bias and Generalization. While RSVP performs well on ReasonSeg and SegInW, its robustness on broader real-world datasets remains underexplored. The reliance on MLLM may introduce bias inherent from pretraining data, affecting the fairness and reliability of the model's output. Future work could investigate domain adaptation techniques to improve generalization beyond curated benchmarks and explore MLLM guardrails to maintain the safety of model output.

Despite these limitations, RSVP establishes a scalable and interpretable framework for reasoning-based segmentation, offering new insights for future improvements in multimodal vision-language grounding.

References

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *ArXiv*, abs/2404.18930.

Jiawang Cao, Yongliang Wu, Weiheng Chi, Wenbo Zhu, Ziyue Su, and Jay Wu. 2024. Reframe anything: Llm agent for open world video reframing. *ArXiv*, abs/2403.06070.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. 2025. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *ArXiv*, abs/2503.09567.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. 2024. M3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. In *Annual Meeting of the Association for Computational Linguistics*.

Anurag Das, Xinting Hu, Li Jiang, and Bernt Schiele. 2024. Mta-clip: Language-guided semantic segmentation with mask-text alignment. In *European Conference on Computer Vision*, pages 39–56. Springer.

Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. 2021. Vision-language transformer and query generation for referring segmentation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16301–16310.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Xinting Hu, Li Jiang, and Bernt Schiele. 2024. Training vision transformers for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4007–4017.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. 2023. Segment anything. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3992–4003.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

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

Bozheng Li, Yongliang Wu, Yi Lu, Jiashuo Yu, Licheng Tang, Jiawang Cao, Wenqing Zhu, Yuyang Sun, Jay Wu, and Wenbo Zhu. 2025. Veu-bench: Towards comprehensive understanding of video editing. *ArXiv*, abs/2504.17828.

Weifeng Lin, Xinyu Wei, Ruichuan An, Peng Gao, Bocheng Zou, Yulin Luo, Siyuan Huang, Shanghang Zhang, and Hongsheng Li. 2024. Draw-and-understand: Leveraging visual prompts to enable mllms to comprehend what you want. *ArXiv*, abs/2403.20271.

Chang Liu, Henghui Ding, and Xudong Jiang. 2023a. Gres: Generalized referring expression segmentation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23592–23601.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual Instruction Tuning. *Preprint*, arXiv:2304.08485.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chun yue Li, Jianwei Yang, Hang Su, Jun-Juan Zhu, and Lei Zhang. 2023c. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *ArXiv*, abs/2303.05499.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. 2020. Multi-task collaborative network for joint referring expression comprehension and segmentation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10031–10040.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana-Maria Camburu, Alan Loddon Yuille, and Kevin P. Murphy. 2015. Generation and comprehension of unambiguous object descriptions. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11–20.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, and Lama Ahmad. 2024. GPT-4 Technical Report. *Preprint*, arXiv:2303.08774.

Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. 2025. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. 2024. Grounding multimodal large language models to the world. In *The Twelfth International Conference on Learning Representations*.

Leigang Qu, Haochuan Li, Tan Wang, Wenjie Wang, Yongqi Li, Liqiang Nie, and Tat-Seng Chua. 2024. Unified text-to-image generation and retrieval. *ArXiv*, abs/2406.05814.

Muhammad Faza Ilmanuddin Rachmadi, Salman Shalahuddin, Heru Permana, Urip Teguh Setijohatmo, and Jonner Hutahaean. 2023. Xlm-roberta model for key information extraction on military document. In *2023 10th International Conference on ICT for Smart Society (ICISS)*, pages 1–6. IEEE.

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

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. 2024. Grounded sam: Assembling openworld models for diverse visual tasks. *ArXiv*, abs/2401.14159.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *Neural Information Processing Systems*.

Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. 2023. What does CLIP know about a red circle? Visual prompt engineering for VLMs. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11953–11963. IEEE Computer Society.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, and Jiahui Yu. 2024. Gemini: A Family of Highly Capable Multimodal Models. *Preprint*, arXiv:2312.11805.

Haoxiang Wang, Pavan Kumar Anasosalu Vasu, Fartash Faghri, Raviteja Vemulapalli, Mehrdad Farajtabar, Sachin Mehta, Mohammad Rastegari, Oncel Tuzel, and Hadi Pouransari. 2023. Sam-clip: Merging vision foundation models towards semantic and spatial understanding. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3635–3647.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. 2022. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*.

Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yan Guo, Ming Gong, and Tongliang Liu. 2021. Cris:

Clip-driven referring image segmentation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11676–11685.

Zhaoqing Wang, Xiaobo Xia, Ziye Chen, Xiao He, Yandong Guo, Mingming Gong, and Tongliang Liu. 2024. Open-vocabulary segmentation with unpaired mask-text supervision. *arXiv preprint arXiv:2402.08960*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

Yifan Wu, Pengchuan Zhang, Wenhan Xiong, Barlas Oğuz, James C. Gee, and Yixin Nie. 2023. The role of chain-of-thought in complex vision-language reasoning task. *ArXiv*, abs/2311.09193.

Yixuan Wu, Yizhou Wang, Shixiang Tang, Wenhao Wu, and Tong He. 2024a. DetToolChain: A New Prompting Paradigm to Unleash Detection Ability of MLLM. *Preprint*, arXiv:2403.12488.

Yongliang Wu, Xinting Hu, Yuyang Sun, Yizhou Zhou, Wenbo Zhu, Fengyun Rao, Bernt Schiele, and Xu Yang. 2024b. Number it: Temporal grounding videos like flipping manga. *arXiv preprint arXiv:2411.10332*.

Yongliang Wu, Wenbo Zhu, Jiawang Cao, Yi Lu, Bozheng Li, Weiheng Chi, Zihan Qiu, Lirian Su, Haolin Zheng, Jay Wu, and Xu Yang. 2024c. Video repurposing from user generated content: A large-scale dataset and benchmark. *ArXiv*, abs/2412.08879.

Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. 2023. Open-vocabulary panoptic segmentation with text-to-image diffusion models. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2955–2966.

Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. 2023. Universal instance perception as object discovery and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15325–15336.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyue Li, and Jianfeng Gao. 2023a. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *ArXiv*, abs/2310.11441.

Senqiao Yang, Tianyuan Qu, Xin Lai, Zhuotao Tian, Bohao Peng, Shu Liu, and Jiaya Jia. 2023b. An improved baseline for reasoning segmentation with large language model. *arXiv e-prints*, pages arXiv–2312.

Xu Yang, Yongliang Wu, Mingzhuo Yang, Haokun Chen, and Xin Geng. 2023c. Exploring diverse in-context configurations for image captioning. *Advances in Neural Information Processing Systems*, 36:40924–40943.

Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Heng-shuang Zhao, and Philip H. S. Torr. 2021. Lavt: Language-aware vision transformer for referring image segmentation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18134–18144.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023d. The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision). *Preprint*, arXiv:2309.17421.

Congzhi Zhang, Linhai Zhang, Jialong Wu, Yulan He, and Deyu Zhou. 2025. Causal prompting: Debiasing large language model prompting based on front-door adjustment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25842–25850.

Congzhi Zhang, Linhai Zhang, and Deyu Zhou. 2024a. Causal walk: Debiasing multi-hop fact verification with front-door adjustment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19533–19541.

Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, and Joyce Chai. 2024b. Groundhog: Grounding large language models to holistic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14227–14238.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alexander J. Smola. 2023. Multimodal chain-of-thought reasoning in language models. *Trans. Mach. Learn. Res.*, 2024.

Qiji Zhou, Ruochen Zhou, Zike Hu, Panzhong Lu, Siyang Gao, and Yue Zhang. 2024. Image-of-Thought Prompting for Visual Reasoning Refinement in Multimodal Large Language Models. *Preprint*, arXiv:2405.13872.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *Preprint*, arXiv:2304.10592.

Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. 2023. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15116–15127.

Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Singh Behl, Jianfeng Wang, Lu Yuan, Nanyun Peng, Lijuan Wang, Yong Jae Lee, and Jianfeng Gao. 2022. Generalized decoding for pixel, image, and language. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15116–15127.

Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. 2024. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems*, 36.

A Appendix

B Supplementary Ablation Experiment Results

In this section, we present more ablation experiments designed to validate our pipeline's design choices as complementary material to ablation experiments in previous sections.

B.1 Padding Ablation Study.

To investigate the performance impact of padding size, we experimented on padding ratios set to 0%, 20% and 40% relative to the height/width of the cropped image region. Our main results reported are obtained with the 20% padding. With three diverse padding sizes, we report the following results on ReasonSeg-Test split:

Model	Padding Size (%)	gIoU	cIoU
RSVP-GPT	0%	58.8	57.1
	20%	**60.3**	**60.0**
	40%	60.0	59.3
RSVP-LLaVA	0%	52.6	49.9
	20%	**55.9**	**50.7**
	40%	53.7	50.1

Table 9: Ablation study on the effect of padding ratio in the visual prompt. Best-performing rows for each model are highlighted in gray, and best values are shown in bold.

As the above result reveals, the moderate-sized padding provides the most optimal balance between object boundaries preservation and minimizing irrelevant background inclusion.

B.2 CoT method Ablation Study.

We conducted an ablation experiment for investigating the impact of Multi-modal CoT design on our system's performance. Our design is compared with the prompt design adapting (Wu et al., 2023), which force the MLLM to first summarize all perceptable visual cues in the given media, then perform analysis according to the summarization.

Model	gIoU	cIoU
RSVP-LLaVA	**55.6**	**50.9**
RSVP-LLaVA, with (Wu et al., 2023)	52.5	46.2
RSVP-GPT	**60.3**	**60.0**
RSVP-GPT, with (Wu et al., 2023)	57.6	56.4

Table 10: Comparison of RSVP performance with/without integrating the CoT framework (Wu et al., 2023). Best-performing rows are highlighted in gray and best values are highlighted in bold.

Adopting CoT design from literature Wu et al. (2023) degrades performance, likely due to fundamental differences between our reasoning-and-localization task and the caption-based selection tasks focused by the literature, demonstrating the importance of choosing appropriate CoT frameworks for achieving satisfiable reasoning segmentation performance. Prompt is attached in Figure 10.

B.3 Model Temperature Ablation Study.

Temperature can significantly impact the consistency of MLLM's output. We evaluated our pipeline under different temperature settings on non-MoE models where the randomness of outputs are controllable by temperature parameter, to evaluate our system's robustness. Experiments on the ReasonSeg-Test split yield the following results:

Model	Inference Temperature	gIoU	cIoU
RSVP-LLaVA	0.0	**55.9**	**50.7**
	0.4	55.8	50.5
	0.8	55.4	50.9
RSVP-Qwen	0.0	**56.6**	**51.6**
	0.4	56.4	51.7
	0.8	56.5	51.4

Table 11: Effect of varying temperature in RSVP-LLaVA and RSVP-Qwen. Rows with the best gIoU or cIoU for each model are highlighted in gray, and best values are bolded.

Although slight variations are observed across settings, the overall performance remains stable. Temperature 0.0's result are what we reported in our paper (bold text in the table). These findings demonstrate that our method is robust to changes in the temperature parameter.

B.4 Dependency on First Stage MLLM Ablation Study

To further explore our system's performance reliance on first-stage MLLM choice, we experimented an additional ablation experiment on Qwen2-VL with its two variants (7B and 2B parameter size) on the ReasonSeg-Test split (with temperature set to 0) varying significantly in model parameter sizes.

Model	gIoU	cIoU
RSVP-Qwen (7B First-Stage MLLM)	**56.6**	**51.6**
RSVP-Qwen (2B First-Stage MLLM)	45.8	43.7
LISA (7B MLLM, Zero-shot)	48.7	48.8

Table 12: Comparison of First-stage MLLM choices with different model sizes and types. RSVP-Qwen (7B) achieves the highest performance.

The performance drop for the smaller 2B model is likely caused by its limited reasoning abilities, along with its reduced capacity to fully comprehend task requirements. It is worth notice that the performance difference between the RSVP-Qwen (2B) and LISA-7B is only up to 5.1% in terms of IoU metrics. In practice, deploying the pipeline with models smaller than 7B with the cost of a small level of performance dropping in reason segmentation tasks is practical and acceptable.

C Reasoning Segmentation Case Study

In this section, we demonstrate case studies regarding our model's results with LISA-13B on the ReasonSeg Dataset Test split. We mainly report two types of demonstrative results: Incorrect localization and low-quality segmentation masks.
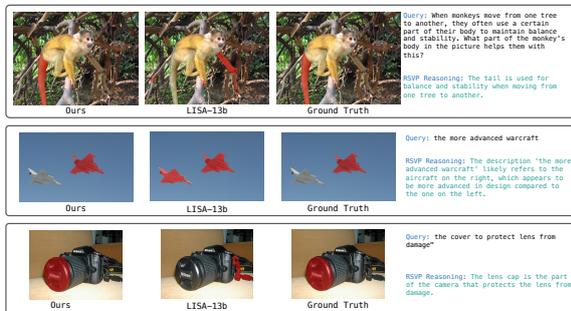
Figure 4: Illustration of Incorrect Localization cases produced by LISA.
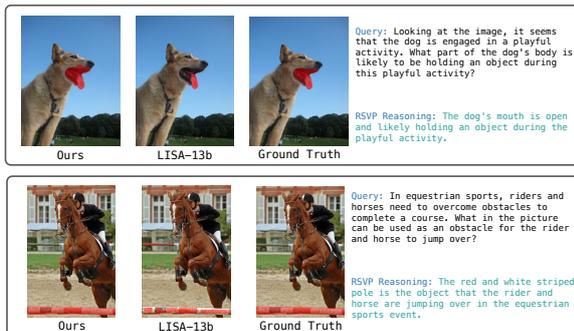
Figure 5: Illustration of Low-quality Segmentation mask cases produced by LISA.

C.1 Case: Incorrect Localization

Fig. 4 shows three distinct cases that demonstrated the limitation of LISA in correctly comprehending the query and localizing the correct object in the image. In all three examples, our RSVP successfully identified the object of interest by reasoning about the provided query, while LISA failed to detect the correct object of interest. In the first case, LISA segmented out the tree branch which has nothing to

do with the query, while in the second case, LISA was not aware of the requirement of distinguishing the more advanced fighter, while RSVP correctly captured this subtle requirement and identified the double-engine jet. For the final example, LISA mistook the strap as the item that protects the lens.

Figure 6: Illustration of Low-quality Segmentation mask cases produced by LISA.

C.2 Case: Low-Quality Segmentation Mask

Fig. 5 and Fig. 6 demonstrated another limitation of LISA, which is the possibility of producing poorly shaped segmentation masks. In the first case, LISA only managed to segment out the dog's tongue, while RSVP correctly segmented out the entire mouth along with the tongue. For the horse case, our model precisely identified the pole on the front, while LISA produced a shattered segmentation mask for the front pole, and also incorrectly segmented out the unrelated pole in the background as well. The final example shown in Fig. 6 of LISA segmented a part of the tire, but the segmentation mask is deformed and only covers a very small portion of the object of interest.

C.3 Analysis and Summary

The above case study demonstrated two main limitations of LISA. On the one hand, LISA may incorrectly reason about the provided query, therefore producing off-the-track localized results that lead to the segmentation of unrelated objects. On the other hand, the latent segmentation proposal token that is produced by LISA's front-end MLLM may not be able to efficiently guide the second-stage segmentation model to generate high-quality, complete segmentation masks. The generated visual prompt and the produced region proposal are visualized in Fig. 12.

C.4 Bad Case Analysis

In this subsection, we conduct a qualitative analysis of failure cases. Our observations reveal two primary categories of error:

False Localization or Misinterpretation. THe first bad case type originates from the first-stage false interpretation of user prompt or incorrect/imprecise localization, where the initial object localization or comprehension of object type is suboptimal.
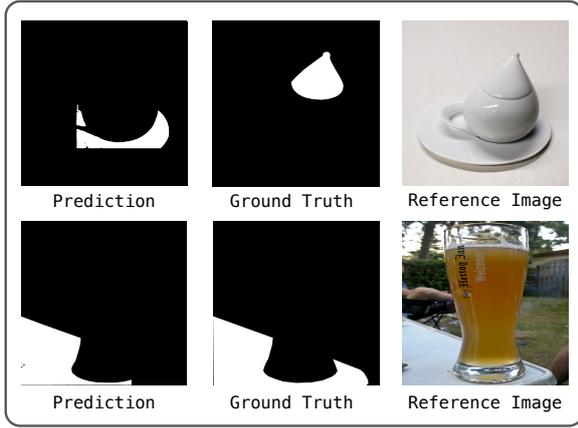
Figure 7: Illustration of bad cases caused by false localization or prompt misinterpretation.

As shown in Fig. 7, in the first example, the query referred to the ceramic cup lid, but the system misinterpreted it as the plate, resulting in completely incorrect localization and segmentation. In the second example, the system only captured the left half of the table because the first-stage MLLM failed to include the right half in its localization.

Suboptimal Segmentation. This type of bad cases are mainly caused by the second-stage VLSM module which produces unsatisfactory segmentation masks with holes or incomplete/hard edges, resulting in suboptimal gIoU and cIoU metrics.
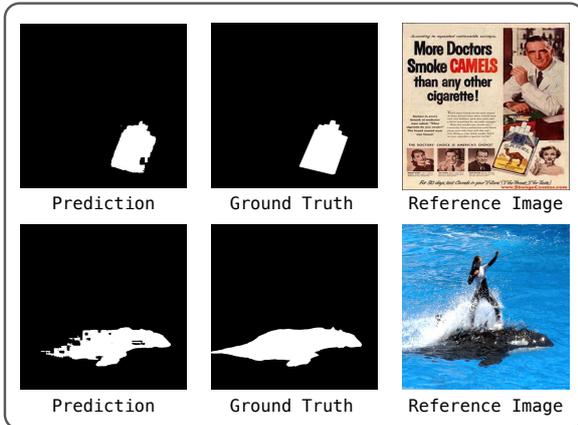
Figure 8: Illustration of bad cases caused by suboptimal segmentation masks.

As shown in Fig. 8, in both cases, our system correctly identifies the target object but fails to fully segment it. In the cigarette advertisement, the system recognizes the cigarette box but the mask contains a hole. In the whale case, the whale is mostly captured by RSVP, but the tail is not fully segmented, leaving holes and hard edges on the mask, partly due to the challenging visual conditions, as the whale's tail is obscured by intense waves.

D Implementation Detail of Region-aware Visual Prompt

During the experimentation, evaluation, and ablation study, we designed the following prompt universally as demonstrated in Fig. 9. For "vertically-segmented image", we refer to the image that is being processed by a horizontally-dividing visual prompt, while for "horizontally-segmented image", it refers to the image being processed by the vertically-dividing visual prompt. The algorithm is demonstrated as in Algorithm 1. The example of a generated visual prompt is shown in Fig. 11.

Each image is resized to a resolution of 1000×1000 during processing, and the margin width/height is dynamically set as 20% of the width / height of the uniformly divided vertical / horizontal regions.

Algorithm 1 Visual Prompt Generation

1: **procedure** GENERATEVISUALPROMPT(I, n) ▷ Generate prompts for I with n segments
2: $h, w \leftarrow$ HEIGHT(I), WIDTH(I) ▷ Get image dimensions
3: $s_h \leftarrow h/n, s_w \leftarrow w/n$ ▷ Compute segment sizes
4: ▷ **Vertical Segmentation (Horizontal Lines)**
5: $I_h, I_v = I, I$
6: **for** $i \leftarrow 1$ to n **do**
7:   DRAWHORIZONTALLINE($I_h, y = (i-1) \cdot s_h$, color=green) ▷ Draw horizontal line
8:   PLACENUMERICLABEL($I_h, x = w/2, y = (i-0.5) \cdot s_h$, color=white)▷ Center label
9: ▷ **Horizontal Segmentation (Vertical Lines)**
10: **for** $j \leftarrow 1$ to n **do**
11:   DRAWVERTICALLINE($I_v, x = (j-1) \cdot s_w$, color=green) ▷ Draw vertical line
12:   PLACENUMERICLABEL($I_v, x = (j-0.5) \cdot s_w, y = h/2$, color=white)▷ Center label
13: **return** I_h, I_v ▷ Return image with visual prompts

Algorithm 1: Algorithm used for generating visual prompts.
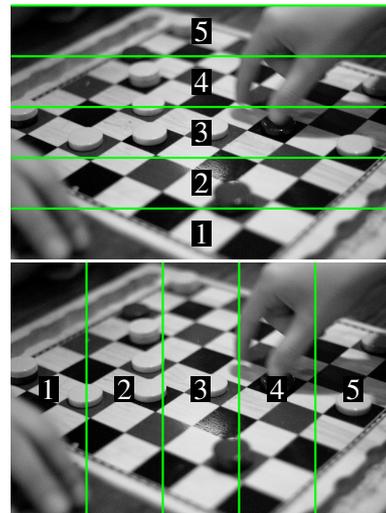

Figure 11: Visual prompt example: the image is evenly divided into vertical/horizontal regions. The dividing density is set to five for demonstrative purposes.

⚙ CoT Prompt Applied in Evaluations and Experiments

You are now acting as an image annotator. I will give you several abstract descriptions which refer to the same object which may appear in the given image and you need to locate them if they exists.

Now I will provide you with two images featuring exactly the same scene but annotated with vertical and horizontal visual segmentation prompt.

The first provided image is divided into <<segment_density>> segments vertically, and the second one is divided horizontally into <<segment_density>> segments. Each segment is labeled with a number from 1 to <<segment_density>>.

Now please carefully follow the below instruction step by step:
1. Infer what is (are) the objects that the questions are jointly referring to, carefully review the given image and find the object(s) that is the most probable candidate(s).

2. Denote from which labeled segment to which contain any parts of the object(s) that the questions described, both vertically and horizontally.

Tell me these segment ids in json as follows:
{
    "instance": "<words(adj+noun) short summarative description of the answer object>",
    "ids_vertical": [1,2,3,...],
    "ids_horizontal": [1,2,3...],
    "reason": "<your rationale about what is(are) being referred>"
}

Attention: you MUST give all segment numbers that contain any parts of the object(s) that the questions described. The answer object may not be the salient object in the frame, and may even not appear in the frame.

If the answer is not in the image, please return empty ids like:
{
    "instance": "<words(adj+noun) short summarative description of the answer object>",
    "ids_vertical": [],
    "ids_horizontal": [],
    "reason": "<your rationale about what is being referred and why none of the object in the scene matches>"
}

Now, the questions are presented as follows, do NOT output anything other than the required json.

questions:
<<questions>>

Figure 9: The prompt utilized to query MLLMs in our implementation.

⚙ Prompt Applied in CoT Framework Ablation

You are now acting as an image annotator. I will give you several abstract descriptions which refer to the same object which may appear in the given image and you need to locate them if they exists.
Now I will provide you with two images featuring exactly the same scene but annotated with vertical and horizontal visual segmentation prompt.
The first provided image is divided into <<segment_density>> segments vertically, and the second one is divided horizontally into <<segment_density>> segments. Each segment is labeled with a number from 1 to <<segment_density>>.

Now please carefully follow the below instruction step by step:
1. Determine if the provided descriptions are long, complex or not. If the descriptions are long, abstract and complex, please perform the following task:
1.1. Infer what is (are) the objects that the questions are jointly referring to, carefully review the given image and find the object(s) that is the most probable candidate(s).
1.2. Describe this object in a short, concise, summarative way.
1.3. Denote from which labeled segment to which contain any parts of the object(s) that the questions described, both vertically and horizontally.
1.4. Set "is_complex" flag to True.
2. Otherwise, perform the following task instead:
2.1 Infer what is (are) the objects that the questions are jointly referring to, carefully review the given image and find the object(s) that is the most probable candidate(s).
2.2 Describe this object in a short, concise way, also include its relative location (top/bottom, left/right, center, ...) in the image.
2.3 Set "is_complex" flag to False.

Tell me these segment ids in json as follows:
{
    "is_complex": <true/false>,
    "instance": "<words(adj+noun) short summarative description of the answer
                object>",
    "ids_vertical": [1,2,3,...],
    "ids_horizontal": [1,2,3...],
    "reason": "<your rationale about what is(are) being referred>"
}

Attention: you MUST give all segment numbers that contain any parts of the object(s) that the questions described. The answer object may not be the salient object in the frame, and may even not appear in the frame.
If the answer is not in the image, please return empty ids like:
{"instance": "<words(adj+noun) short summarative description of the answer object>", "ids_vertical": [], "ids_horizontal": [], "reason": "<your rationale about what is being referred and why none of the object in the scene matches>"}

Now, the questions are presented as follows, do NOT output anything other than the required json.
questions:
<<questions>>

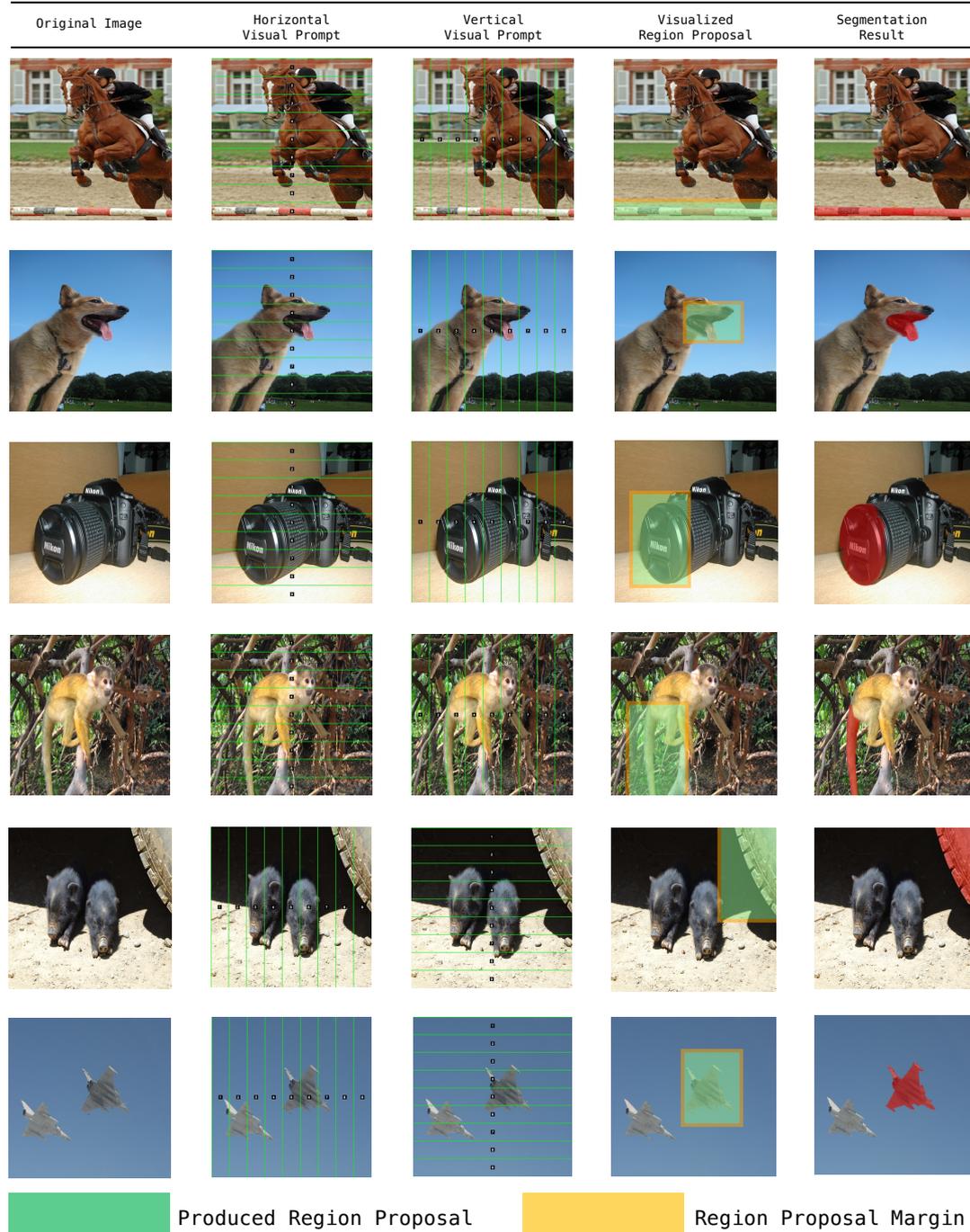**Figure 10:** The prompt utilized to experiment on CoT framework ablation experiment.

Figure 12: Visualization of produced visual prompts and region proposals for demonstrated cases.