

000 SEG-AGENT: IMPROVING LANGUAGE-GUIDED SEG- 001 MENTATION VIA EXPLICIT CHAIN-OF-REASONING 002 CONSTRUCTION

006 **Anonymous authors**

007 Paper under double-blind review

011 ABSTRACT

013 Language-guided segmentation breaks through the scope limitations of traditional
014 semantic segmentation, enabling models to segment any target region in an image
015 based on user instructions. Existing methods are typically two-stage frameworks:
016 they first employ multimodal large language models (MLLMs) to understand the
017 textual instruction and generate visual prompts from the image, and then use foun-
018 dational segmentation models such as SAM to produce high-quality masks. How-
019 ever, due to the limited spatial grounding capability of the base models, they usu-
020 ally require training on large-scale datasets to achieve improved segmentation accu-
021 racy. In this paper, we propose **Seg-Agent**, a completely training-free language-
022 guided segmentation method. By constructing an explicit reasoning chain: genera-
023 tion, selection, and refinement, Seg-Agent achieves performance comparable
024 to training-based approaches. Additionally, to evaluate the generalization ability
025 of Seg-Agent, we collect a diverse dataset covering various language-guided seg-
026 mentation scenarios, named **Various-LangSeg**. Extensive experiments demon-
027 strate the effectiveness of our proposed method. The code and dataset will be
028 made publicly available.

029 1 INTRODUCTION

031 The rapid development of multimodal large language models (MLLMs) (Liu et al., 2023b; Achiam
032 et al., 2023; Bai et al., 2025) and foundational segmentation models (Cheng et al., 2021; 2022; Kir-
033 illov et al., 2023; Ravi et al., 2024) has driven significant progress in language-guided segmentation
034 (Ren et al., 2024a; Lai et al., 2024). Unlike traditional segmentation methods (Xie et al., 2021;
035 Zheng et al., 2021; Liu et al., 2023a; Hao et al., 2025), which are limited to predefined categories
036 and scenarios, language-guided segmentation models can segment any target region of interest based
037 on textual instruction. This makes it an open and domain-unrestricted segmentation approach. As
038 shown in Figure 1, we categorize common segmentation tasks into three types: explicit semantic seg-
039 mentation (ESS), generic object segmentation (GOS), and reasoning-guided segmentation (RGS).
040 Traditional segmentation models can typically handle only a limited subset of these scenarios. How-
041 ever, with the powerful understanding capabilities of MLLMs and the flexible configuration of text
042 prompts, language-guided segmentation models are capable of addressing all three categories.

043 Most existing language-guided segmentation models follow a two-stage approach (Lai et al., 2024;
044 Ren et al., 2024b; Chen et al., 2024; Liu et al., 2025). First, they use MLLMs to understand the
045 instruction and perceive the image, generating visual prompts (typically in the form of bounding
046 boxes or points). Then, a foundational segmentation model such as SAM is employed to produce
047 high-quality segmentation masks based on these visual prompts. However, due to limitations such
048 as the MLLM’s relatively weak spatial perception and grounding capabilities, the visual prompts it
049 generates directly are often of low quality (Lai et al., 2024; Yang et al., 2023). As a result, these
050 models typically require training on large-scale datasets to improve performance.

051 However, these training-based methods have several notable limitations. First, it requires collecting
052 large datasets for training. Due to the diversity of segmentation scenarios, it is difficult to fully cover
053 all possible cases in the training data, which limits the model’s generalization ability and leads to
054 poor performance in out-of-distribution (OOD) scenarios. Second, training models requires sub-

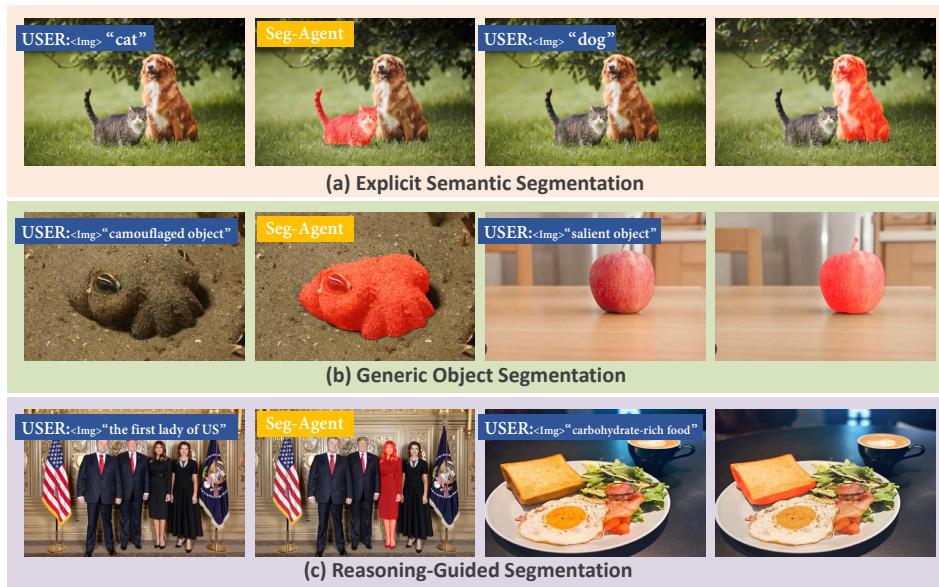


Figure 1: Given an image and a textual target, Seg-Agent can handle segmentation tasks across various scenarios: (a) Explicit Semantic Segmentation: segmenting objects with clearly defined semantics (e.g., “cat”, “dog”). (b) Generic Object Segmentation: segmenting conceptually defined objects without specific categories (e.g., “camouflaged object”, “salient object”). (c) Reasoning-Guided Segmentation: segmenting targets based on prompts that require commonsense or factual reasoning (e.g., “the first lady of US”, “carbohydrate-rich food”).

stantial computational resources, especially for MLLMs. And as newer, more powerful base models become available, training-based methods cannot be directly integrated with them. Therefore, in this paper, we propose **Seg-Agent**, a completely training-free language-guided segmentation framework. As shown in Figure 2, existing methods are all training-based, either fully trained or partially trained, which inevitably introduces the aforementioned inherent limitations. In fact, regardless of whether training is involved, the core objective is to improve the quality of the generated visual prompts and the final segmentation mask. Instead of enhancing model capabilities through training, we directly construct an explicit reasoning chain to guide the model. Specifically, we propose a three-step reasoning process: generation, selection, and refinement, through which the MLLM progressively improves the generated visual prompts, ultimately leading to better segmentation masks.

Furthermore, considering the limited scene diversity in existing language-guided segmentation datasets (Lai et al., 2024; Kazemzadeh et al., 2014; Mao et al., 2016) and to better validate the generalization capability of segmentation models, we collect a multi-scenario evaluation dataset called **Various-LangSeg**. Specifically, Various-LangSeg includes the three types of tasks illustrated in Figure 1: explicit semantic segmentation, generic object segmentation, and reasoning-guided segmentation, which collectively cover the majority of common language-guided segmentation scenarios. We evaluate the performance of Seg-Agent and several related language-guided segmentation models (Lai et al., 2024; Chen et al., 2024; Liu et al., 2025) on Various-LangSeg.

We summarize our contributions as follows:

- We propose Seg-Agent, a completely training-free framework for language-guided segmentation. By constructing an explicit reasoning chain for guidance, Seg-Agent achieves performance comparable to training-based methods.
- We have collected a comprehensive evaluation dataset named Various-LangSeg, which covers nearly all common scenarios of language-guided segmentation and effectively assesses models’ generalization ability.
- Extensive experiments demonstrate the effectiveness of our proposed method and provide a low-cost, simple, and effective design paradigm for the community.

108 2 RELATED WORK

110 **Multimodal Large Language Models.** In recent years, MLLMs have achieved revolutionary
 111 progress in vision-language tasks. Models such as GPT-4 (Achiam et al., 2023) and Qwen-VL2.5
 112 (Bai et al., 2025) have demonstrated outstanding capabilities in understanding multimodal content,
 113 giving them a natural advantage in tasks requiring joint image-text reasoning, thus driving the de-
 114 velopment of numerous downstream tasks (Rawles et al., 2024; Cheng et al., 2024). MLLMs have
 115 shown remarkable performance in visual question answering (Agrawal et al., 2016), image caption-
 116 ing (Ghandi et al., 2023), and multimodal reasoning (Zhang et al., 2024b). However, they typically
 117 lack fine-grained spatial perception and grounding abilities (Wu et al., 2024; Lai et al., 2024), which
 118 poses challenges for dense prediction tasks such as segmentation. Previous approaches mostly en-
 119 hance grounding capabilities by training on task-specific datasets (Zhang et al., 2024a; Ren et al.,
 120 2024b). In contrast, our work leverages off-the-shelf, native MLLMs to generate visual prompts, but
 121 circumvents their limitations in spatial understanding through an explicit reasoning chain.

122 **Foundational Segmentation Models.** Found-
 123 dational segmentation models such as Mask-
 124 Former (Cheng et al., 2021; 2022) and SAM
 125 (Kirillov et al., 2023; Ravi et al., 2024) are
 126 general-purpose models trained on large-scale
 127 data for universal segmentation. In particular,
 128 SAM introduces a promptable interface that en-
 129 ables segmentation with sparse visual cues like
 130 points or bounding boxes. These models offer
 131 strong generalization ability and high-quality
 132 mask prediction across various domains. In our
 133 framework, we use SAM2 as the backend seg-
 134 mentation module and focus on improving its
 135 performance by enhancing the quality of the vi-
 136 sual prompts generated by MLLMs.

136 **Language-Guided Segmentation.** Early
 137 methods typically employ a text encoder (De-
 138 vlin et al., 2018; Radford et al., 2021) to ex-
 139 tract textual features for guiding the segmenta-
 140 tion model (Ren et al., 2024a; Liu et al., 2023c;
 141 Yang et al., 2022). With the advancement of
 142 MLLMs and foundational segmentation mod-
 143 els, most existing approaches have evolved into the two-stage framework described earlier, although
 144 they differ in subtle aspects. PixELLM (Ren et al., 2024b) and OMG-LLaVA (Zhang et al., 2024a)
 145 use their own lightweight decoders to generate masks. However, such decoders generally underper-
 146 form compared to SAM, which is pretrained on massive-scale data. As a result, current mainstream
 147 methods directly integrate SAM as the segmentation backbone. LISA (Lai et al., 2024), Sa2VA
 148 (Yuan et al., 2025) and GSVA (Xia et al., 2024) do not generate explicit visual prompts, instead,
 149 they introduce a special $\langle\text{SEG}\rangle$ token to compress textual information, requiring fine-tuning of
 150 SAM so that it can interpret this novel type of prompt. In contrast, SAM4MLLM (Chen et al.,
 151 2024) and Seg-Zero (Liu et al., 2025) adopt a more intuitive strategy: they keep SAM frozen and
 152 instead post-train MLLMs to generate more accurate bounding boxes or points prompts—formats
 153 that SAM can directly understand, thereby achieving improved performance. These two works are
 154 the closest to our Seg-Agent, except that we enhance the quality of generated visual prompts through
 155 a manually designed, explicit reasoning chain without any training.

155 3 VARIOUS-LANGSEG: A COMPREHENSIVE EVALUATION BENCHMARK

156 **Motivation** Existing language-guided segmentation datasets primarily focus on referring seg-
 157 mentation (Kazemzadeh et al., 2014; Mao et al., 2016) and reasoning segmentation (Lai et al., 2024).
 158 However, these datasets cover a limited spectrum of task types, which hinders the thorough eval-
 159 uation of general-purpose models. To address this gap, we introduce **Various-LangSeg**, a unified
 160 and diverse benchmark designed to evaluate language-guided segmentation methods across a broad
 161 range of scenarios.

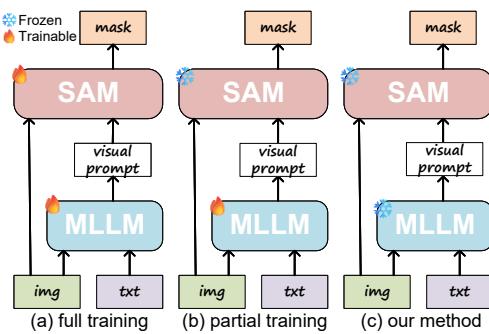


Figure 2: Comparison between the proposed Seg-Agent and existing methods. (a) Full training: both the MLLM and SAM are trained simultaneously, with representative methods including LISA (Lai et al., 2024) and Sa2VA (Yuan et al., 2025). (b) Partial training: only the MLLM is trained while SAM is kept fixed, with representative methods including SAM4MLLM (Chen et al., 2024) and Seg-Zero (Liu et al., 2025). (c) Our method is entirely training-free.

162 **Task Categorization** As shown in Figure 1, we categorize language-guided segmentation tasks into
 163 three representative scenarios: Explicit Semantic Segmentation, Generic Object Segmentation, and
 164 Reasoning-Guided Segmentation. These three scenarios jointly span most common language-guided
 165 segmentation tasks, enabling comprehensive evaluation of model generalization.

166 **Dataset Construction** Given an input image and a textual instruction, the goal is to generate the
 167 corresponding binary mask. Since our method does not require any training, Various-LangSeg is
 168 designed solely for evaluation purposes.

169 To construct the dataset, we proceed as follows:

171 1. **Image Collection**: We sample images and corresponding masks from existing public datasets (Lv
 172 et al., 2021; Li et al., 2014; Vicente et al., 2016; Dong et al., 2013; Lin et al., 2015).

173 2. **Instruction Annotation**: For each image-mask pair $(x_{\text{img}}, y_{\text{mask}})$, we manually annotate a textual
 174 instruction x_{txt} to form the complete triplet $(x_{\text{img}}, x_{\text{txt}}, y_{\text{mask}})$.

175 Note that although the images and their corresponding masks are sourced from other datasets, we
 176 must manually analyze the relationship between each image and its corresponding mask, assign
 177 accurate textual instructions, and carefully select high-quality image-mask pairs. This process is not
 178 merely a simple sampling and assembly to form a new dataset.

179 For each scenario:

- 182 • **Explicit Semantic Segmentation**: 20 common object categories (e.g., cat, dog, bird) with
 183 7 images per category, totaling 140 samples.
- 184 • **Generic Object Segmentation**: 4 binary segmentation tasks: salient object detection
 185 (SOD) (Borji et al., 2015), camouflaged object detection (COD) (Fan et al., 2020), shadow
 186 detection (SD) (Vicente et al., 2016), and image tampering detection (ITD) (Dong et al.,
 187 2013). Each task includes 16 samples. Textual input directly uses the task name (e.g.,
 188 “salient object”).
- 189 • **Reasoning-Guided Segmentation**: 40 samples are annotated using complex descriptions
 190 requiring implicit reasoning. These are inspired by ReasonSeg (Lai et al., 2024), such as
 191 “carbohydrate-rich food”, where identification requires reasoning and domain knowledge.

192 In total, Various-LangSeg contains 244 evaluation samples, with 140, 64, and 40 samples in three
 193 scenarios respectively. Its scale is comparable to that of the validation set of the ReasonSeg dataset
 194 (Lai et al., 2024). We visually present the statistics of Various-LangSeg in Figure 5 in the Appendix.

196 **Evaluation Metrics** We follow established works (Lai et al., 2024; Liu et al., 2025) and adopt two
 197 standard metrics: **gIoU (global IoU)**: the average IoU over all samples. **cIoU (cumulative IoU)**: the
 198 ratio of the total intersection area to the total union area across the entire dataset.

200 4 METHOD

202 We propose Seg-Agent, a training-free framework for language-guided segmentation. Our method
 203 is a two-stage approach that first uses MLLMs to generate visual prompts and then employs a base
 204 segmentation model to produce the final mask. However, unlike previous methods that directly
 205 generate visual prompts (Chen et al., 2024; Liu et al., 2025) in a single step, we construct an explicit
 206 reasoning chain to guide the MLLM’s prompt generation process, comprising generation, selection,
 207 and refinement, mimicking an iterative procedure of progressively localizing and finetuning the
 208 target boundaries. In contrast to traditional end-to-end approaches that rely on learned features, Seg-
 209 Agent explicitly constructs and updates visual prompts, guiding the segmentation model through
 210 interpretable, step-by-step interactions. The overall pipeline is illustrated in Figure 3.

212 4.1 PROBLEM FORMULATION

214 Let $\mathbf{X}_{\text{img}} \in \mathbb{R}^{H \times W \times 3}$ be an input image and $\mathbf{X}_{\text{txt}} \in \mathcal{T}$ be a natural language instruction describing
 215 the target object (e.g., “the man in red”). The goal is to predict a binary mask $\mathbf{Y}_{\text{mask}} \in \{0, 1\}^{H \times W}$
 that segments the described object.

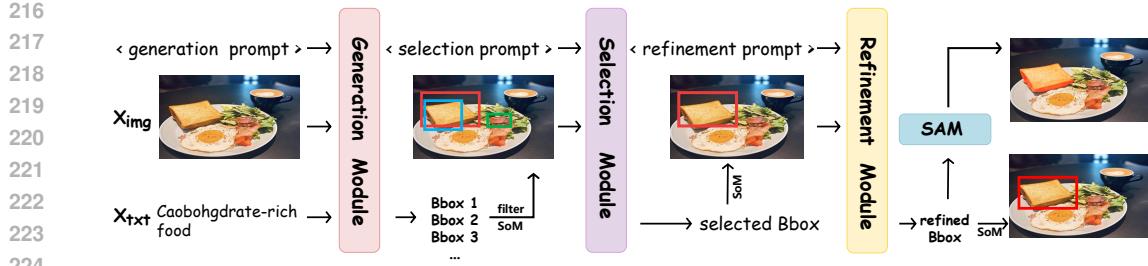


Figure 3: Illustration of Seg-Agent. By constructing an explicit reasoning chain: generation, selection, and refinement, the MLLM is able to improve the quality of generated visual prompts, thereby enabling SAM to produce more accurate target segmentation masks. SoM here indicates Set-of-Mark prompt (Yang et al., 2023).

4.2 OVERVIEW OF SEG-AGENT PIPELINE

Seg-Agent consists of four modules:

1. **Generation Module:** Proposes diverse bounding boxes using image augmentations.
2. **Selection Module:** Selects the most appropriate box via visual comparison.
3. **Refinement Module:** Fine-tunes the selected box to better align with the object boundaries.
4. **Segmentation:** Applies a pretrained segmenter such as SAM to produce the final mask.

4.3 THE FORWARD PASS OF SEG-AGENT

Generation Module To ensure robustness across views and scales, we apply a set of augmentations $\mathcal{A} = \{a_i\}_{i=1}^N$ to the input image (including flipping, scaling, etc):

$$\mathbf{X}^{(i)} = a_i(\mathbf{X}_{\text{img}}), \quad i = 1, \dots, N. \quad (1)$$

Each augmented image $\mathbf{X}^{(i)}$ is paired with \mathbf{X}_{txt} and sent to an MLLM with a task-specific prompt (denoted as <generation prompt>, see Appendix) to localize the object:

$$\mathbf{B}^{(i)} = \text{MLLM}_{\text{gen}}(\mathbf{X}^{(i)}, \mathbf{X}_{\text{txt}}, \text{<generation prompt>}). \quad (2)$$

This yields a set of bounding box proposals (coordinate format: $[x_1, y_1, x_2, y_2]$):

$$\mathcal{B}_{\text{gen}} = \left\{ \mathbf{B}^{(i)} \right\}_{i=1}^N. \quad (3)$$

Selection Module To consolidate candidate boxes back into the original image frame, we invert the augmentations:

$$\tilde{\mathbf{B}}^{(i)} = a_i^{-1}(\mathbf{B}^{(i)}). \quad (4)$$

We perform Non-Maximum Suppression (NMS) to filter redundant boxes:

$$\mathcal{B}_{\text{sel}} = \text{NMS} \left(\left\{ \tilde{\mathbf{B}}^{(i)} \right\}, \theta_{\text{IoU}} \right). \quad (5)$$

Using a visualization strategy such as Set-of-Mark (SoM) (Yang et al., 2023), we render the candidate boxes onto the original image. This method has been shown to enhance the spatial perception and grounding capability of MLLMs (Yang et al., 2023; Rawles et al., 2024), and also allows for intuitive visualization of the spatial relationship between the bounding box and the target object, as illustrated in Figure 3. The MLLM receives the SoM-marked image, textual instruction, and a comparison prompt (denoted as <selection prompt>, see Appendix) to choose the most relevant box:

$$\mathbf{B}_{\text{sel}} = \text{MLLM}_{\text{sel}}(\text{SoM}(\mathbf{X}_{\text{img}}, \mathcal{B}_{\text{sel}}), \mathbf{X}_{\text{txt}}, \mathcal{B}_{\text{sel}}, \text{<selection prompt>}). \quad (6)$$

Refinement Module The selected box \mathbf{B}_{sel} may still require fine-tuning for optimal spatial coverage. We invoke a final reasoning step using another refinement prompt (denoted as <refinement prompt>, see Appendix) to refine it:

$$\mathbf{B}_{\text{refined}} = \text{MLLM}_{\text{refine}}(\text{SoM}(\mathbf{X}_{\text{img}}, \mathbf{B}_{\text{sel}}), \mathbf{X}_{\text{txt}}, \mathbf{B}_{\text{sel}}, \text{<refinement prompt>}). \quad (7)$$

270 Through this reasoning process, the MLLM is able to carefully examine the alignment between the
 271 current bounding box and the target object, and fine-tune it based on semantic and visual context, for
 272 example, by expanding, shrinking, translating, or adjusting its boundaries to achieve more precise
 273 coverage of the target region.

274 The final output $\mathbf{B}_{\text{refined}}$ is a well-refined bounding box that serves as a high-quality visual prompt
 275 for the subsequent segmentation task. This module significantly improves the accuracy of boundary
 276 localization and is a key step toward achieving high-precision segmentation.

277 **Segmentation Module** The refined bounding box is then used as the visual prompt input to a seg-
 278 mentation model such as SAM:

$$280 \quad \mathbf{Y}_{\text{mask}} = \text{SAM}(\mathbf{X}_{\text{img}}, \mathbf{B}_{\text{refined}}). \quad (8)$$

281 SAM uses the bounding box as a spatial prompt to precisely identify and segment the described tar-
 282 get region in the image, producing a high-quality, pixel-level binary mask \mathbf{Y}_{mask} , where 1 indicates
 283 the target region and 0 indicates the background.

284 This step completes the final transformation from a language instruction to an accurate segmenta-
 285 tion, serving as the last stage of the entire Seg-Agent framework. Thanks to the progressive refine-
 286 ment of visual prompts in the previous three stages, SAM receives more accurate guidance, thereby
 287 significantly improving the final segmentation accuracy.

289 4.4 WHY SEG-AGENT MATTERS

290 Seg-Agent decomposes language-guided segmentation into interpretable sub-tasks, enabling robust
 291 performance across diverse conditions without any task-specific training. By explicitly engaging in
 292 step-wise reasoning, Seg-Agent avoids common failure modes of end-to-end systems and provides
 293 traceable decision-making, all while fully leveraging the generalization power of MLLMs.

294 Unlike previous works relying on parameter updates, Seg-Agent operates in a zero-shot and training-
 295 free setting, relying solely on step-wise reasoning within MLLMs. This has several advantages:

- 297 • **Generalization:** Augmentation-enriched proposals increase robustness across unseen dis-
 298 tributions. We perform no post-training on MLLMs, thus avoiding any potential negative
 299 impacts on their performance.
- 300 • **Interpretability:** Each reasoning step is explicit and traceable, enabling transparent de-
 301 bugging and user intervention.
- 302 • **Modularity:** Seg-Agent can be instantly adapted to newer and stronger MLLMs or seg-
 303 mentation models without retraining.

305 5 EXPERIMENTS

307 5.1 EXPERIMENTAL SETTING

308 **Implementation Details.** We em-
 309 ploy QwenVL-2.5 (Bai et al., 2025)
 310 as the base MLLM for generating
 311 visual prompt, the generation mod-
 312 ule, selection module, and refine-
 313 ment module are all built upon it,
 314 which can be deployed locally or
 315 accessed via API services. Addi-
 316 tionally, we use SAM2-Large (Ravi
 317 et al., 2024) to generate segmen-
 318 tation masks. Seg-Agent operates in a
 319 training-free manner, and the entire
 320 inference process can be completed
 321 on a single NVIDIA RTX 4090 GPU
 322 with 24 GB of memory. For the NMS
 323 step, we set the IoU threshold to 0.8.
 During inference, the prompt for Seg-
 Agent is pre-defined (see Appendix),

324 Table 1: Referring segmentation results. Methods marked
 325 with “*” are traditional approaches, while the other methods
 326 are based on MLLMs. We compare cIoU in this table. Best
 327 results are in **bold**.

Method	RefCOCO testA	RefCOCO+ testA	RefCOCOg test
training-based methods			
CRIS*	73.2	68.1	60.4
LAVT*	75.8	68.4	62.1
ReLA*	76.5	71.0	66.0
LISA-7B	76.5	67.4	68.5
PixelLM-7B	76.5	71.7	70.5
PerceptionGPT-7B	78.6	73.9	71.7
Seg-Zero-3B	79.3	73.7	71.5
Seg-Zero-7B	80.3	76.2	72.6
training-free methods			
Qwen2.5-VL-3B + SAM2-L	75.9	71.5	70.1
Qwen2.5-VL-7B + SAM2-L	77.8	73.5	71.2
Seg-Agent-3B (Ours)	79.0	73.2	71.4
Seg-Agent-7B (Ours)	79.9	76.0	72.2

requiring only the input of the target object. For other methods, we set the text input according to the templates provided in their works.

Datasets. Since our method does not require training, we only select benchmark datasets to evaluate model performance. Following prior related work (Lai et al., 2024; Chen et al., 2024; Liu et al., 2025), we adopt three datasets for the referring segmentation task: refCOCO, refCOCO+ (Kazemzadeh et al., 2014), and refCOCOg (Mao et al., 2016). These datasets involve simple textual descriptions such as “the man wearing white clothes”, belonging to the explicit semantic segmentation scenario. We also include the ReasonSeg (Lai et al., 2024) dataset for the reasoning segmentation task, which falls under the reasoning-guided segmentation scenario and contains referring expressions that require reasoning, such as “the food with the most Vitamin C”. Finally, we introduce the Various-LangSeg dataset proposed in this paper, which covers three common scenarios in language-guided segmentation and effectively evaluates the model’s generalization and versatility.

Baseline Methods. We compare Seg-Agent with two groups of methods: (1) Training-based methods, including traditional methods that do not use MLLMs such as grounded-sam (Ren et al., 2024a), ReLA (Liu et al., 2023a) etc., and MLLM-based methods such as LISA (Lai et al., 2024), SAM4MLLM (Chen et al., 2024), etc.; (2) Training-free methods, we mainly adopt the Qwen2.5-VL + SAM2 (Bai et al., 2025; Ravi et al., 2024) baseline method introduced in Seg-Zero (Liu et al., 2025), with related settings and prompts kept consistent with that paper, i.e., letting MLLMs directly output bounding box coordinates in a single-step method to prompt SAM, rather than having an explicit thinking process like Seg-Agent.

Evaluation Metrics. We follow prior work (Lai et al., 2024; Liu et al., 2025) in adopting two evaluation metrics: gIoU and cloU, which have been described before. Since cloU tends to be heavily biased toward large objects and exhibits high variability, gIoU is generally preferred as the primary metric.

Table 3: Results on Various-LangSeg. We report the performance across three scenarios and the overall performance.

Method	Various-LangSeg							
	Explicit Semantic		Generic Object		Reasoning-Guided		Overall	
training-based methods								
ReLA*	76.8	77.7	20.1	22.1	25.2	21.2	53.4	53.8
OVSeg*	77.1	76.0	23.2	23.0	25.0	22.1	54.4	53.2
LISA-7B	81.9	83.1	32.4	30.8	46.8	36.7	63.2	64.2
LISA-13B	82.8	83.9	35.3	40.3	54.4	47.9	65.8	67.9
PixelLLM-7B	81.5	83.5	31.2	31.5	45.2	40.1	62.3	62.7
SAM4MLLM	82.1	83.5	32.0	31.8	45.2	37.1	62.9	62.3
Seg-Zero-7B	81.8	81.0	41.4	43.5	74.5	67.1	70.0	68.9
training-free methods								
Qwen2.5-VL-3B + SAM2-L	80.1	77.4	33.3	33.7	61.0	49.5	64.7	61.3
Qwen2.5-VL-7B + SAM2-L	80.8	79.7	39.5	38.3	70.1	58.9	68.2	65.4
Seg-Agent-3B (Ours)	82.3	81.5	40.8	31.5	66.1	62.6	68.8	60.9
Seg-Agent-7B (Ours)	83.0	83.7	41.0	42.1	75.2	66.7	70.6	68.5

5.2 COMPARISON WITH OTHER METHODS

In this subsection, we conduct a comparative analysis of the performance between Seg-Agent and several most relevant methods. We compare CRIS (Wang et al., 2022), LAVT (Yang et al., 2022),



Figure 4: Visual comparison between the proposed Seg-Agent and existing related methods. We show three common scenarios of language-guided segmentation here.

OVSeg (Liang et al., 2023), X-Decoder (Zou et al., 2022), SEEM (Zou et al., 2023), ReLA (Liu et al., 2023a), LISA (Lai et al., 2024), PixelLLM (Ren et al., 2024b), PerceptionGPT (Pi et al., 2024), SAM4MLLM (Chen et al., 2024) and Seg-Zero (Liu et al., 2025).

Referring Segmentation. As shown in Table 1, we evaluate Seg-Agent and other related methods on the test sets of refCOCO, refCOCO+, and refCOCOg. As mentioned earlier, the target objects in these datasets are described using simple and direct text descriptions, which falls under the explicit semantic segmentation scenario. Due to the relatively simple nature of this task, all methods achieve good performance, and there is little difference between traditional methods and MLLM-based methods. Notably, the training-free baseline method Qwen2.5VL+SAM2-L also achieves favorable results, while Seg-Agent provides a certain improvement over this baseline and achieves performance comparable to the SOTA training-based method Seg-Zero.

Reasoning Segmentation. As shown in Table 2, we evaluate Seg-Agent and other related methods on the validation and test sets of ReasonSeg (Lai et al., 2024). As previously mentioned, the target objects in this dataset are described using reasoning-based textual expressions, categorized into long queries and short queries, falling under the reasoning-guided segmentation scenario. Compared to explicit semantic segmentation, this task requires reasoning to first identify the target object, making it significantly more challenging. As can be seen from the table, traditional methods that do not employ MLLMs perform poorly in this scenario, as their text encoders are typically better at extracting textual features but lack reasoning capabilities. In contrast, MLLM-based methods achieve substantial performance improvements. Moreover, the training-free baseline method Qwen2.5VL+SAM2-L also achieves favorable performance, while Seg-Zero achieves significant gains over this baseline through GRPO (Shao et al., 2024) training. Similarly, our Seg-Agent achieves notable improvement over the same baseline and attains performance comparable to the SOTA training-based method Seg-Zero, surpassing earlier approaches such as LISA.

Various-LangSeg. As shown in Table 3, we evaluate Seg-Agent and related methods on Various-LangSeg. It can be observed that traditional methods perform poorly on this dataset, achieving acceptable results only in the explicit semantic segmentation scenario, and even then, they are outperformed by MLLM-based approaches. Their performance is also weak in the general object segmentation scenario and particularly poor in the reasoning-guided segmentation scenario. In contrast, MLLM-based methods are applicable across all three scenarios. However, due to the lack of training on relevant data and the absence of an explicit reasoning process, early methods such as LISA perform relatively poorly on the general object segmentation task, underperforming compared to their results on the other two tasks. Seg-Zero achieves the best performance among training-based methods, especially in the reasoning-guided scenario, benefiting from task-specific training. For training-free methods, Seg-Agent consistently improves upon the Qwen2.5-VL+SAM2-L baseline in all three scenarios. Notably, Seg-Agent-7B achieves the best overall performance among training-free methods, and even outperforms many training-based approaches, highlighting its strong generalization and reasoning ability without task-specific training.

Visual Comparison. Figure 4 presents a visual comparison of Seg-Agent with several of the most relevant methods, including LISA, Seg-Zero, and Qwen2.5-VL-7B+SAM2-L. In the first row (explicit semantic segmentation scenario), the target object is a sunflower, other methods either over-

segment or under-segment the object, while only Seg-Agent produces an accurate and precise segmentation. In the second row (general object segmentation scenario), other methods fail to fully capture the main structure or produce overly coarse boundaries, whereas Seg-Agent successfully preserves the overall structure without edge blurring. In the final row (reasoning-guided segmentation scenario), only Seg-Agent correctly identifies the first lady of US (second from the right), while all other methods select incorrect objects. These visual results demonstrate the strong generalization capability and segmentation accuracy of Seg-Agent. We present more visualization results in Figure 6 in the Appendix.

5.3 ABLATION STUDY

Effectiveness of Each Module. As shown in Table 4, the performance when combining GM with SM or RM is better than that of the baseline method. Only when all three modules are used together does the performance reach its maximum, demonstrating the effectiveness of the explicitly constructed reasoning chain in our approach. As can be observed, compared to the baseline method, the combination of different modules yields the largest performance gain in the reasoning-guided segmentation scenario, likely because this task inherently requires segmentation is relatively simple, and the base resulting in only marginal improvements. From scenario appears to be the most challenging; thus, only when all three modules are used together. Qwen2.5-VL-7B + SAM2-L. In the first row, according to the baseline method mentioned earlier when used together, only one candidate box needs to be

Generalization across Base Models. We conduct this experiment to demonstrate that Seg-Agent can be directly adapted to different base models. As shown in Table 5, under the same configuration, Seg-Agent, which employs an explicit reasoning chain, achieves performance improvements compared to direct inference. Moreover, it can be observed that Seg-Agent is compatible with different MLLMs (Bai et al., 2025; Zhu et al., 2025) and segmentation models (Kirillov et al., 2023; Ravi et al., 2024), and the stronger the base model, the better Seg-Agent performs. This highlights the advantage of our proposed training-free approach: as newer and more powerful base models emerge, Seg-Agent can directly integrate with them to a is a significant advantage over training-based m

6 CONCLUSION

In this paper, we propose Seg-Agent, a completely training-free language-guided segmentation model. By constructing explicit reasoning chains of generation, selection, and refinement to guide the model in generating more accurate visual prompts, Seg-Agent achieves segmentation performance comparable to training-based methods. Additionally, we construct the Various-LangSeg dataset containing multiple scenarios, which can comprehensively evaluate the generalization capability of language-guided segmentation models. Extensive experiments demonstrate the effectiveness of our approach. We hope that our simple yet effective method can provide rich inspiration to the community.

Table 4: Ablation study on each module. GM, SM and RM denote generation module, selection module and refinement module, respectively. We compare gIoU here.

GM	SM	RM	Various-LangSeg			
			ESS	GOS	RGS	Overall
X	X	X	80.8	39.5	70.1	68.2
✓	✓	X	81.2	39.8	72.0	68.8
✓	X	✓	81.5	40.0	71.8	69.0
✓	✓	✓	83.0	41.0	75.2	70.6

a reasoning process. In contrast, explicit semantic line method already achieves strong performance, the metrics, the generic object segmentation scene noticeable performance improvement is achieved. Notably, the configuration used here is consistently one of the three modules are employed, correspond- no improvements applied. When GM and RM are e generated.

Table 5: Ablation study on base models. a, b, and c represent different combinations of MLLMs and base segmentation models, directly using a single-step reasoning approach. Seg-Agent (x) denotes using the configuration described in x to replace the corresponding part of Seg-Agent. We compare gIoU in this table.

Setting	Various-LangSeg			
	ESS	GOS	RGS	Overall
a: InternVL3-8B + SAM2-L	79.2	35.3	67.1	65.7
b: Qwen2.5-VL-7B + SAM-L	80.1	37.5	69.9	67.3
c: Qwen2.5-VL-7B + SAM2-L	80.8	39.5	70.1	68.2
Seg-Agent (a)	79.8	36.3	67.5	66.4
Seg-Agent (b)	81.1	39.5	71.2	68.6
Seg-Agent (c)	83.0	41.0	75.2	70.6

486 ETHICS STATEMENT
487488 This work adheres to the ICLR Code of Ethics. This study involved no human subjects or animal
489 experimentation. All datasets used were sourced in compliance with relevant usage guidelines, en-
490 suring no violation of privacy. We have taken care to avoid any biases or discriminatory outcomes
491 in our research process. No personally identifiable information was used, and no experiments were
492 conducted that could raise privacy or security concerns. We are committed to maintaining trans-
493 parency and integrity throughout the research process.
494495 REPRODUCIBILITY STATEMENT
496497 To facilitate the reproduction of the Seg-Agent proposed in this paper, we provide implementation
498 details in Section B and an initial version of the code in the supplementary materials. Additionally,
499 we include the JSON files describing the Various-LangSeg dataset in the supplementary materials,
500 and we will publicly release the full code and dataset in the near future.
501502 REFERENCES
503504 Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Ale-
505 man, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical
506 report. *arXiv preprint arXiv:2303.08774*, 2023.507 Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv
508 Batra, and Devi Parikh. Vqa: Visual question answering, 2016. URL <https://arxiv.org/abs/1505.00468>.
509510 Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang,
511 Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*,
512 2025.513 Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark.
514 *IEEE transactions on image processing*, 24(12):5706–5722, 2015.515 Yi-Chia Chen, Wei-Hua Li, Cheng Sun, Yu-Chiang Frank Wang, and Chu-Song Chen. Sam4mllm:
516 Enhance multi-modal large language model for referring expression segmentation. In *European
517 Conference on Computer Vision*, pp. 323–340. Springer, 2024.518 Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you
519 need for semantic segmentation. 2021.520 Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-
521 attention mask transformer for universal image segmentation. 2022.522 Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Yantao Li, Jianbing Zhang, and Zhiyong
523 Wu. Seeclick: Harnessing gui grounding for advanced visual gui agents, 2024. URL <https://arxiv.org/abs/2401.10935>.
524525 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep
526 bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.527 Jing Dong, Wei Wang, and Tieniu Tan. Casia image tampering detection evaluation database. In
528 *2013 IEEE China Summit and International Conference on Signal and Information Processing*,
529 pp. 422–426, 2013. doi: 10.1109/ChinaSIP.2013.6625374.530 Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Cam-
531 ouflaged object detection. In *Proceedings of the IEEE/CVF conference on computer vision and
532 pattern recognition*, pp. 2777–2787, 2020.533 Taraneh Ghandi, Hamidreza Pourreza, and Hamidreza Mahyar. Deep learning approaches on image
534 captioning: A review. *ACM Computing Surveys*, 56(3):1–39, October 2023. ISSN 1557-7341.
535 doi: 10.1145/3617592. URL <http://dx.doi.org/10.1145/3617592>.
536

540 Chao Hao, Zitong Yu, Xin Liu, Jun Xu, Huanjing Yue, and Jingyu Yang. A simple yet effective
 541 network based on vision transformer for camouflaged object and salient object detection. *IEEE*
 542 *Transactions on Image Processing*, 2025.

543 Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to
 544 objects in photographs of natural scenes. In Alessandro Moschitti, Bo Pang, and Walter Daele-
 545 mans (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Pro-
 546 cessing (EMNLP)*, pp. 787–798, Doha, Qatar, October 2014. Association for Computational Lin-
 547 guistics. doi: 10.3115/v1/D14-1086. URL <https://aclanthology.org/D14-1086/>.

548 Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete
 549 Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceed-
 550 ings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

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

554 Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille. The secrets of salient object
 555 segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*,
 556 pp. 280–287, 2014.

557 Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang,
 558 Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted
 559 clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*,
 560 pp. 7061–7070, 2023.

561 Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro
 562 Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects
 563 in context, 2015. URL <https://arxiv.org/abs/1405.0312>.

564 Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation,
 565 2023a. URL <https://arxiv.org/abs/2306.00968>.

566 Haotian Liu, Chunyuan Li, Qingsheng Wu, and Yong Jae Lee. Visual instruction tuning. *Advances
 567 in neural information processing systems*, 36:34892–34916, 2023b.

568 Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei
 569 Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for
 570 open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023c.

571 Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-
 572 zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint
 573 arXiv:2503.06520*, 2025.

574 Yunqiu Lv, Jing Zhang, Yuchao Dai, Aixuan Li, Bowen Liu, Nick Barnes, and Deng-Ping Fan.
 575 Simultaneously localize, segment and rank the camouflaged objects. In *Proceedings of the
 576 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11591–11601, 2021.

577 Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy.
 578 Generation and comprehension of unambiguous object descriptions. In *2016 IEEE Conference
 579 on Computer Vision and Pattern Recognition (CVPR)*, pp. 11–20. IEEE, June 2016. doi: 10.1109/
 580 cvpr.2016.9. URL <http://dx.doi.org/10.1109/CVPR.2016.9>.

581 Renjie Pi, Lewei Yao, Jiahui Gao, Jipeng Zhang, and Tong Zhang. Perceptiongpt: Effectively fusing
 582 visual perception into llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and
 583 Pattern Recognition (CVPR)*, pp. 27124–27133, June 2024.

584 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agar-
 585 wal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya
 586 Sutskever. Learning transferable visual models from natural language supervision, 2021. URL
 587 <https://arxiv.org/abs/2103.00020>.

594 Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham
 595 Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images
 596 and videos. *arXiv preprint arXiv:2408.00714*, 2024.

597 Christopher Rawles, Sarah Clinckemaillie, Yifan Chang, Jonathan Waltz, Gabrielle Lau, Marybeth
 598 Fair, Alice Li, William Bishop, Wei Li, Folawiyo Campbell-Ajala, Daniel Toyama, Robert Berry,
 599 Divya Tyamagundlu, Timothy Lillicrap, and Oriana Riva. Androidworld: A dynamic benchmark-
 600 ing environment for autonomous agents, 2024. URL <https://arxiv.org/abs/2405.14573>.

601 Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang,
 602 Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual
 603 tasks. *arXiv preprint arXiv:2401.14159*, 2024a.

604 Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie
 605 Jin. Pixellm: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF
 606 Conference on Computer Vision and Pattern Recognition*, pp. 26374–26383, 2024b.

607 Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang,
 608 Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathe-
 609 matical reasoning in open language models, 2024. URL <https://arxiv.org/abs/2402.03300>.

610 Tomás F Yago Vicente, Le Hou, Chen-Ping Yu, Minh Hoai, and Dimitris Samaras. Large-scale
 611 training of shadow detectors with noisily-annotated shadow examples. In *Computer Vision–
 612 ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016,
 613 Proceedings, Part VI 14*, pp. 816–832. Springer, 2016.

614 Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang
 615 Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference
 616 on computer vision and pattern recognition*, 2022.

617 Jiannan Wu, Muyan Zhong, Sen Xing, Zeqiang Lai, Zhaoyang Liu, Zhe Chen, Wenhai Wang,
 618 Xizhou Zhu, Lewei Lu, Tong Lu, Ping Luo, Yu Qiao, and Jifeng Dai. Visionllm v2: An end-
 619 to-end generalist multimodal large language model for hundreds of vision-language tasks, 2024.
 620 URL <https://arxiv.org/abs/2406.08394>.

621 Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Gen-
 622 eralized segmentation via multimodal large language models. In *Proceedings of the IEEE/CVF
 623 Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3858–3869, June 2024.

624 Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Seg-
 625 former: Simple and efficient design for semantic segmentation with transformers. In *Neural
 626 Information Processing Systems (NeurIPS)*, 2021.

627 Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark
 628 prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*,
 629 2023.

630 Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip H. S. Torr. Lavt:
 631 Language-aware vision transformer for referring image segmentation, 2022. URL <https://arxiv.org/abs/2112.02244>.

632 Haobo Yuan, Xiangtai Li, Tao Zhang, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi,
 633 Jiashi Feng, and Ming-Hsuan Yang. Sa2va: Marrying sam2 with llava for dense grounded under-
 634 standing of images and videos. *arXiv preprint arXiv:2501.04001*, 2025.

635 Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Chen Change Loy,
 636 and Shuicheng Yan. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and
 637 understanding. *Advances in neural information processing systems*, 37:71737–71767, 2024a.

638 Zhusong Zhang, Aston Zhang, Mu Li, hai zhao, George Karypis, and Alex Smola. Multimodal
 639 chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*,
 640 2024b. ISSN 2835-8856. URL <https://openreview.net/forum?id=y1pPWFVfvR>.

648 Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu,
 649 Jianfeng Feng, Tao Xiang, Philip H.S. Torr, and Li Zhang. Rethinking semantic segmentation
 650 from a sequence-to-sequence perspective with transformers. In *CVPR*, 2021.

651 Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen
 652 Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu,
 653 Xinguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen
 654 Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi,
 655 Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong
 656 Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min
 657 Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhui Wang. In-
 658 ternvl3: Exploring advanced training and test-time recipes for open-source multimodal models,
 659 2025. URL <https://arxiv.org/abs/2504.10479>.

660 Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat
 661 Behl, Jianfeng Wang, Lu Yuan, Nanyun Peng, Lijuan Wang, Yong Jae Lee, and Jianfeng Gao.
 662 Generalized decoding for pixel, image, and language, 2022. URL <https://arxiv.org/abs/2212.11270>.

663 Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng
 664 Gao, and Yong Jae Lee. Segment everything everywhere all at once, 2023. URL <https://arxiv.org/abs/2304.06718>.

669 A THE USE OF LARGE LANGUAGE MODELS

670 LLMs were used only during the writing phase, including for polishing the text and providing suggestions to improve the paper's figures.

674 B MORE DETAILS ABOUT SEG-AGENT

675 In this section, we present some implementation details of Seg-Agent. We also provide a basic
 676 version of the code in the supplementary materials, and the complete code will be released shortly.

679 B.1 GENERATION MODULE

680 In our experiments, we observed that the object identified in a single inference pass can be entirely
 681 incorrect. Inspired by object segmentation approaches that generate multiple candidate boxes, we
 682 found that if multiple candidate boxes are produced, it becomes easier for the LLM to select the
 683 one that correctly localizes the target object. To this end, we adopted a multi-view input strategy
 684 to encourage the generation of multiple candidate boxes. Specifically, we employed common im-
 685 age augmentation techniques from computer vision, including flipping (horizontal and vertical) and
 686 scaling (zooming in by $2\times$ or zooming out to $0.5\times$). In each case, we primarily selected two such
 687 augmented views along with the original image, resulting in three candidate boxes.

688 The <generation prompt> we used is as follows:

```
689
690 def build_generation_prompt(query):
691     template = """Locate "{query}", report the bboxes
692     coordinates in JSON format."""
693     return template.format(query=query)
694
```

695 Here, the query refers to the textual description of the target object we wish to segment. The output
 696 coordinates are in the format $[x_1, y_1, x_2, y_2]$, representing the pixel coordinates of the top-left and
 697 bottom-right corners of the bounding box, e.g., $[100, 100, 200, 200]$.

699 B.2 SELECTION MODULE

700 The Set-of-Marks (SoM) (Yang et al., 2023) used here has been shown to effectively enhance
 701 MLLMs' perception of specific objects in images and is widely employed in tasks such as screen-

702 shot grounding (Rawles et al., 2024). The approach involves overlaying visual bounding boxes on
 703 the original image according to the box coordinates, explicitly illustrating the spatial relationship
 704 between the target object and the generated box to help the MLLM evaluate the quality of candidate
 705 boxes. Different candidate boxes are distinguished by distinct colors, and the MLLM is provided
 706 with corresponding prompts to guide its selection.

707 The <selection prompt> we used is as follows:

```
709 def build_select_prompt(query, coords):
710     prompt = f'Please analyze the image provided
711     below and determine which of the bounding boxes
712     better captures the "{query}".\n'
713
714     prompt += """
715 Your task is to:
716 1. Identify which bounding box more
717     accurately includes the entire target object.
718 2. Provide a brief explanation for your choice.
719
720 Note: The format of the bounding box is
721 [x_min, y_min, x_max, y_max], representing the
722 top-left and bottom-right coordinates.
723
724 The coordinates for each bounding box are as follows:
725 """
726     d = 0
727     colors = ['red', 'green', 'blue', 'yellow']
728
729     for a in coords:
730         prompt += f'- **Bbox {d+1} ({colors[d]})**: {a}\n'
731         d += 1
732     prompt += """
733 Return your answer in the following format:
734
735 Best Box: <Box Number>
736 Reasoning: <Explanation>"""
737
738 Here, "coords" refers to the list of candidate bounding boxes.
```

739 B.3 REFINEMENT MODULE

740 This step mimics the human annotation process, where candidate bounding boxes are further refined
 741 to better enclose the target object.

743 The <refinement prompt> we used is as follows:

```
744 def build_optimize_prompt(query, current_box):
745     prompt = f"""Please analyze the image provided below
746     and evaluate whether the current bounding box
747     accurately captures the "{query}".\n"""
748
749     prompt += f"The current bounding box
750     coordinates are: {current_box},
751     where:\n"
752     prompt += "- 'x_min' = {:.2f}
753     (left edge)\n".format(current_box[0])
754     prompt += "- 'y_min' = {:.2f}
755     (top edge)\n".format(current_box[1])
756     prompt += "- 'x_max' = {:.2f}
757     (right edge)\n".format(current_box[2])
```

```
756     prompt += "- 'y_max' = {:.2f}\n"
757     (bottom edge)\n".format(current_box[3])
758
759     prompt += """
760
761 Your task is to:
762 1. Assess whether the current bounding
763 box adequately includes the entire target object.
764 2. If the current box does not perfectly
765 capture the target object or leaves unnecessary
766 margins, suggest an optimized bounding box
767 with improved coordinates.
768
769 Note: The current bounding box may
770 not be accurate. Please carefully
771 analyze the image and improve the
772 coordinates if necessary.
773
774 Return your response in the following format:
775
776 Current Box: [x_min, y_min, x_max, y_max]
777 Optimized Box:
778 [x_min_optimized, y_min_optimized,
779 x_max_optimized, y_max_optimized]
780 Reasoning: <Explanation of why the optimization was made>
781 """
782
783     return prompt
```

B.4 SEGMENTATION MODULE

We primarily use SAM2-L (Ravi et al., 2024) to generate binary masks, which only requires an input image and corresponding visual prompts. In our case, the visual prompts are the pixel coordinates of bounding boxes ($[x_1, y_1, x_2, y_2]$) output by the MLLM in the previous step.

C MORE INFORMATION ABOUT VARIOUS-LANGSEG

We provide an overview JSON file of Various-LangSeg in the supplementary materials, which contains detailed information about the entire dataset. Due to file size limitations, we did not upload the images themselves; we will release the full dataset and evaluation code upon paper acceptance.

All images were selected from external publicly available datasets, including NC4K (Lv et al., 2021), PASCAL-S (Li et al., 2014), SBU (Vicente et al., 2016), CASIA (Dong et al., 2013), and COCO (Lin et al., 2015), from which we collected both images and their corresponding masks, and manually constructed the textual prompts. Specifically, the explicit semantic segmentation scenario includes 20 subcategories, such as cat, dog, and other objects with clear semantic meanings, with 7 samples per category, resulting in a total of 140 samples. This task is relatively simple. The reasoning-guided segmentation scenario contains 40 samples, for which we designed textual prompts requiring reasoning based on the spatial and contextual relationships between the masks and the original images. This task has moderate difficulty. Finally, the generic object segmentation scenario includes four popular binary segmentation tasks: camouflaged object detection (COD) (Fan et al., 2020), salient object detection (SOD) (Borji et al., 2015), shadow detection (SD) (Vicente et al., 2016), and image tampering detection (ITD) (Dong et al., 2013). Each subtask contains 16 samples, totaling 64 samples. This task is relatively difficult. The entire dataset comprises 244 samples in total.

Since all segmentation tasks can essentially be guided by language, Various-LangSeg certainly cannot cover all possible scenarios. However, compared to existing datasets (Lai et al., 2024; Kazemzadeh et al., 2014; Mao et al., 2016), Various-LangSeg is more comprehensive and better suited for evaluating the generalization ability of language-guided segmentation models. We also plan to extend it in the future.

Our main effort involved selecting high-quality images and their corresponding masks from existing datasets and manually crafting appropriate textual descriptions based on a careful understanding of the correspondence between each image and its mask. Regarding the dataset scale, since our goal was to establish an evaluation benchmark rather than a large-scale training set, the size of our dataset is relatively modest—comparable to the validation set of ReasonSeg (Lai et al., 2024). We plan to further expand this dataset in the future by incorporating additional segmentation scenarios.

Figure 5 shows the detailed statistics of the Various-LangSeg dataset.

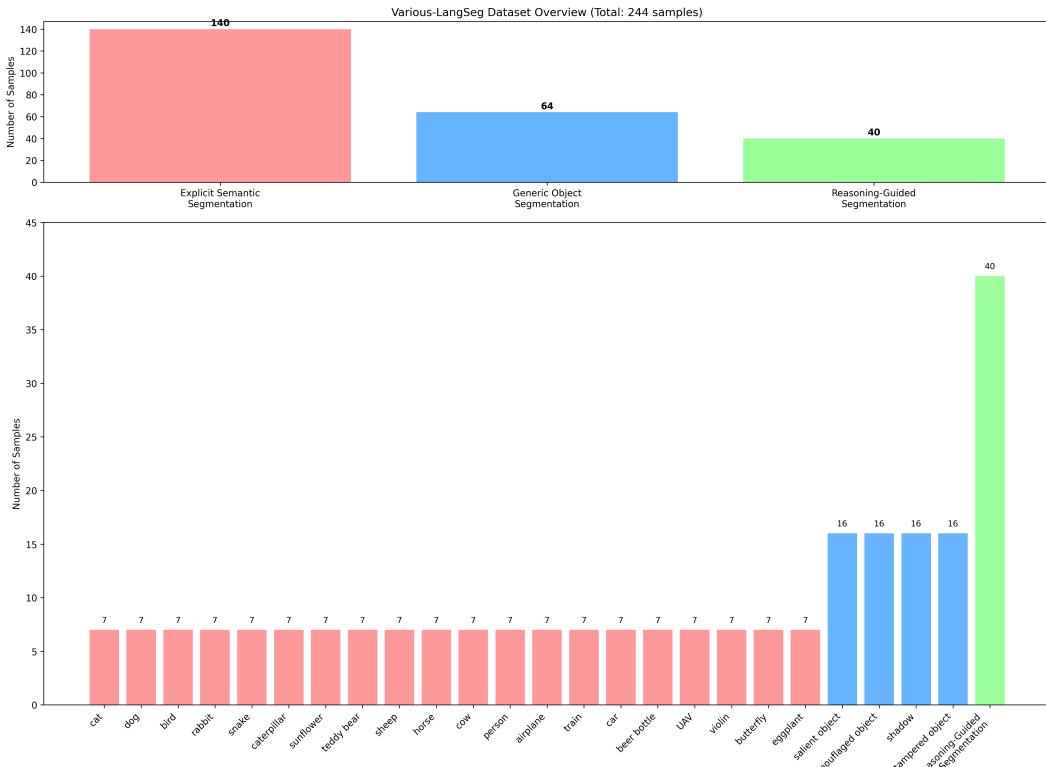


Figure 5: Overview of the Various-LangSeg dataset, comprising 244 evaluation samples across three segmentation scenarios. (Top) High-level distribution: Explicit Semantic Segmentation (140 samples), Generic Object Segmentation (64 samples), and Reasoning-Guided Segmentation (40 samples). (Bottom) Detailed breakdown: 20 explicit object categories (7 samples each), 4 generic tasks: salient object, camouflaged object, shadow, and tampered object detection (16 samples each), and the reasoning-guided category (40 samples).

D MORE VISUALIZATION RESULTS

We provide more visualization results in Figure 6. As can be seen, Seg-Agent is capable of handling inputs in various forms. It supports multilingual inputs, with a primary demonstration of Chinese, English, and their mixed usage. Furthermore, we present a variety of image types, including images from web news, datasets, screenshots, cartoons, photographs taken by cameras, and AI-generated images. The examples also cover the three types of language prompts introduced in the main text: explicit semantic segmentation, generic object segmentation, and reasoning-guided segmentation.

These high-quality segmentation results demonstrate the strong generalization capability and broad applicability of Seg-Agent. We have also released preliminary inference code in the supplementary materials, and welcome readers to try it out.

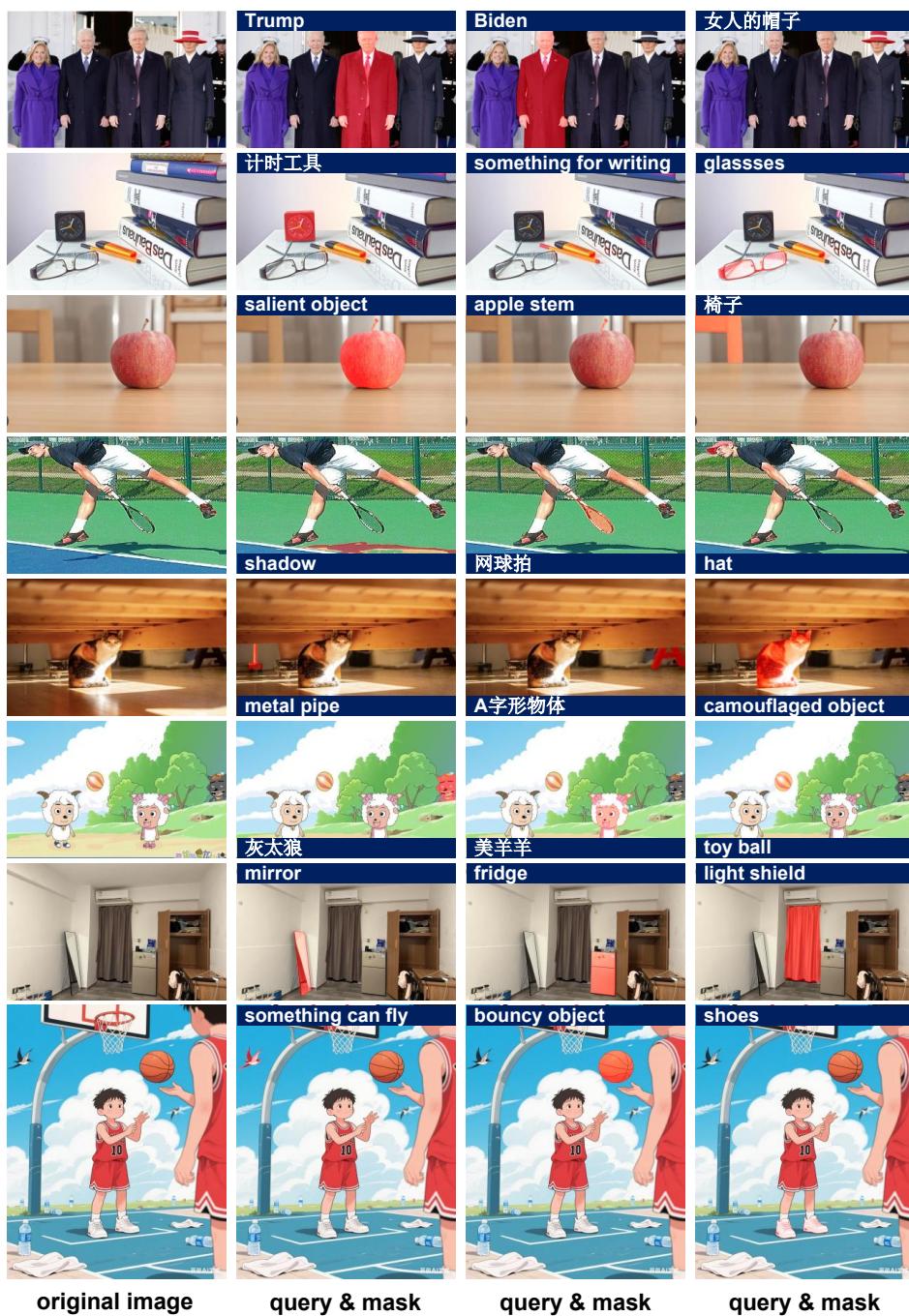


Figure 6: More visualization results of Seg-Agent. Seg-Agent can handle language inputs in various forms, including both Chinese and English. It is also capable of processing different types of images, such as real-world photos, captured photographs, cartoon images, and AI-generated images. These images demonstrate Seg-Agent’s strong generalization ability and its broad range of application scenarios. Please zoom in for a better view.

E INFERENCE PROCESS

The inference process of Seg-Agent is fully transparent and can be observed in real time. As shown in Figure 7, we present a visualization of the complete reasoning pipeline. Given an input image and a target text prompt “shadow”, the generation module first produces multiple candidate bounding boxes through image augmentation and prompt-guided generation, which are then visualized on the



918
919
920
921
922
923
924
925 Figure 7: The inference process of Seg-Agent. The current target object is “shadow”. GM, SM
926 and RM indicate generation module, selection module and refinement module, respectively. Please
927 zoom in for a better view.
928

929 original image in the form of Set-of-Mark (SoM) (Yang et al., 2023). Next, the selection module
930 chooses the most appropriate box based on the SoM-formatted image and a selection prompt. Fi-
931 nally, the refinement module further improves the selected box using a refinement prompt and the
932 SoM input. The final high-quality segmentation mask is then generated by the segmentation model
933 using the refined bounding box as a visual prompt. The entire process is fully open and transparent.
934 By carefully designing this reasoning pipeline, Seg-Agent effectively enhances the quality of the
935 generated masks.

936 Compared to directly generating visual prompts in a single step using MLLMs, we improve the
937 quality of visual prompts through a manually designed reasoning chain. In contrast, training-based
938 approaches (Lai et al., 2024; Liu et al., 2025) enhance output quality by training the model to adjust
939 its weights. While the methodologies differ, the underlying goal is essentially the same: to produce
940 more accurate and reliable visual prompts.

941 F ANALYSIS AND DISCUSSION

942 The experimental results across multiple benchmarks demonstrate the effectiveness, generalizabil-
943 ity, and flexibility of Seg-Agent. Compared with both training-based and training-free baselines,
944 Seg-Agent consistently achieves competitive or superior performance, particularly in more complex
945 scenarios such as reasoning-guided segmentation and mixed-scene generalization (e.g., Various-
946 LangSeg). This validates the strength of our explicit multi-stage reasoning design, which encourages
947 step-by-step visual grounding rather than relying on a single forward pass.

948 Several key observations emerge from the experiments:

- 949 • **Progressive reasoning improves localization.** Our ablation studies confirm that adding
950 selection and refinement steps leads to significant improvements over single-step genera-
951 tion. This highlights the importance of decomposing the task into interpretable subtasks,
952 especially for ambiguous or complex queries.
- 953 • **Seg-Agent is robust across tasks and data distributions.** The model performs well not
954 only on traditional referring segmentation datasets but also on reasoning-intensive and
955 multi-domain scenarios. This suggests that our prompt-based design allows the MLLM
956 to adapt flexibly without task-specific training.
- 957 • **Zero-shot and modular design is practically valuable.** Unlike many training-based
958 methods, Seg-Agent requires no fine-tuning and can easily integrate with newer MLLMs
959 or segmentation models. This makes it a highly deployable and maintainable system in
960 real-world applications.

961 G LIMITATIONS AND FUTURE WORK

962 Despite these advantages, Seg-Agent still has several limitations. First, its multi-stage reasoning
963 process requires multiple calls to MLLMs, involving three steps: generation, selection, and refine-
964 ment, which may result in higher latency compared to single-step methods. Second, the model’s
965 performance heavily depends on the quality of prompt engineering and the inherent capabilities of
966 the MLLM itself; even with a well-designed reasoning pipeline, suboptimal results may arise if the
967 underlying MLLM lacks sufficient reasoning or perception ability. Furthermore, the model is prone
968 to over-reasoning, where the model generates increasingly complex prompts without improving the
969 results. Future work will focus on addressing these limitations and improving the model’s performance
970 and efficiency.

972 to failure in challenging cases, such as when the target object is small or visually ambiguous, when
973 multiple objects are present, or when the language description is vague or indirect. Of course, these
974 issues are common challenges faced by all language-guided segmentation models. Finally, since we
975 do not fine-tune or modify the segmentation model, performance bottlenecks in the segmentation
976 model itself can lead to failures. For example, as shown in Figure 7, the bounding box fully encloses
977 the shadow, yet the segmentation model fails to fully capture it.

978 In future work, we plan to explore more efficient prompting strategies to reduce computation while
979 retaining accuracy, and investigate adaptive reasoning depth based on the complexity of the in-
980 put query. Incorporating visual feedback loops or confidence-based control could further enhance
981 robustness. Finally, expanding Seg-Agent to support multi-turn interaction or conversational seg-
982 mentation is a promising direction for broader applicability.

983

984

985

986

987

988

989

990

991

992

993

994

995

996

997

998

999

1000

1001

1002

1003

1004

1005

1006

1007

1008

1009

1010

1011

1012

1013

1014

1015

1016

1017

1018

1019

1020

1021

1022

1023

1024

1025