SEmO: SEG+EDGE MINUS OTHERS TOKEN WITH PROGRESSIVE MASK DECODING FOR REASONING SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reasoning segmentation aims to generate precise segmentation masks from complex and implicit language queries. Prior approaches mainly rely on a single ⟨SEG⟩ token to represent all segmentation-relevant information and predict the final mask in a single decoding stage. This design makes it difficult to capture fine-grained target edges while suppressing false positives in the background. We propose SEmO, a novel framework that augments the conventional ⟨SEG⟩ token with two role-specific tokens, ⟨EDGE⟩ and ⟨OTHERS⟩, together with a Progressive Mask Decoder (PMD). The three tokens explicitly represent the target region, its edge, and the background, respectively, providing complementary information for mask prediction. At each resolution stage, PMD aggregates their role-specific predictions and propagates the refined result to the next stage for coarse-to-fine mask refinement. These designs enable SEmO to predict more accurate target edges while effectively suppressing background false positives. Experiments demonstrate that SEmO achieves state-of-the-art performance on ReasonSeg and improvements on referring segmentation benchmarks. Moreover, PMD requires only 0.69M parameters and achieves faster inference than existing mask decoders, showing that the performance gains are obtained without sacrificing efficiency. All code and model parameters of SEmO will be made available.

1 INTRODUCTION

Reasoning segmentation aims to generate a pixel-level mask for a target specified by a complex and implicit language query. Unlike referring segmentation, which locates a target explicitly described in language, reasoning segmentation requires language understanding and visual reasoning to infer the intended target. For example, the query "segment the heaviest object on the table" does not directly name the target but requires reasoning over the objects and their attributes to identify the corresponding region. To address such queries, reasoning segmentation combines the reasoning capabilities of vision-language models with pixel-level mask prediction.

LISA (Lai et al., 2024), which first defined this task, introduced the embedding-as-mask paradigm. It adds a special ⟨SEG⟩ token to the vocabulary of a vision-language model (VLM) and feeds its embedding into the mask decoder of the Segment Anything Model (SAM). Subsequent studies (Yang et al., 2023; Xia et al., 2024; Qian et al., 2025) have extended this paradigm in various directions, including handling multiple objects with multiple ⟨SEG⟩ tokens, identifying absent targets with a ⟨REJ⟩ token, and improving spatial alignment using point prompts.

However, these existing ⟨SEG⟩-token-based approaches share a structural limitation. Multiple ⟨SEG⟩ tokens may represent multiple targets, and a ⟨REJ⟩ token may indicate the absence of a target. Nevertheless, each ⟨SEG⟩ token must still jointly encode the target region, its edge, and the background. Moreover, by directly adopting the SAM mask decoder, these methods typically predict the final mask in a single decoding stage without explicit coarse-to-fine refinement. This design makes it difficult to capture fine-grained edges while effectively suppressing background responses. Consequently, existing methods often produce coarse or incomplete edges, as shown in Fig. 1(a), and generate false positives in unrelated regions or adjacent objects, as shown in Fig. 1(b).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To address these limitations, we propose SEmO, which improves edge quality and suppresses false positives while preserving the semantic reasoning capability of VLM tokens. SEmO introduces two key designs. First, it augments the single ⟨SEG⟩ token with two additional role-specific tokens: ⟨EDGE⟩ for the edge and ⟨OTHERS⟩ for the background, allowing ⟨SEG⟩ to focus on the target region alone rather than jointly encoding all three. The ground-truth labels for ⟨EDGE⟩ and ⟨OTHERS⟩ are derived directly from existing mask annotations, requiring no additional annotation. Second, we design a Progressive Mask Decoder (PMD) that progressively refines the mask from coarse to fine resolutions rather than predicting the final mask in a single step.

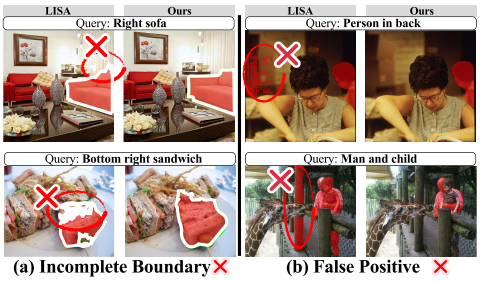

Figure 1: Limitations of existing reasoning segmentation methods compared with SEmO. Red masks and white contours denote predicted regions and boundaries. Existing methods often produce (a) incomplete boundaries and (b) frequent false positives.

SEmO achieves state-of-the-art performance on the ReasonSeg benchmark. Using a 7B VLM backbone, it obtains 65.9% gIoU and 69.1% cIoU on the validation set, surpassing the previous results by 1.2 and 6.0 percentage points, respectively. Despite introducing two additional tokens and progressive decoding, SEmO employs a substantially lighter decoder, PMD, with only 0.69M parameters, compared with 4.06M for LISA and 7.57M for PixelLM. It also reduces inference time, demonstrating that these performance gains do not come at the cost of efficiency.

Our contributions are summarized as follows:

- We propose a new role-specific token representation that augments the conventional ⟨SEG⟩ token with ⟨EDGE⟩ and ⟨OTHERS⟩ tokens, explicitly representing the target region, its edge, and the background. The supervision for the two new tokens is derived from existing mask annotations without additional labeling.
- We design a Progressive Mask Decoder (PMD) that refines the mask across coarse-to-fine resolutions by aggregating the three role-specific predictions at each stage and propagating the previous prediction to the next stage.
- We achieve state-of-the-art performance on ReasonSeg and consistent gains on RefCOCO, demonstrating effectiveness of role-specific representations and progressive refinement for both reasoning and referring segmentation.

2 Related Work

Reasoning Segmentation. Reasoning segmentation generates a pixel-level mask from a complex and implicit language query, in contrast to referring segmentation, which identifies a target through an explicit referring expression. LISA (Lai et al., 2024), which first introduced this task, predicts a mask by using the ⟨SEG⟩ token embedding generated by a vision-language model (VLM) as a prompt for a segmentation model. Follow-up studies have advanced LISA in various directions, including handling multiple or absent targets and improving segmentation quality through enhanced decoding and training strategies (Yang et al., 2023; Jang et al., 2025; Xia et al., 2024; Ren et al., 2024b; Wu et al., 2024b; Bao et al., 2024; Lu et al., 2025; Qian et al., 2025).

More recently, several studies have attempted to move beyond LISA's single-embedding prompting scheme in different ways: AnchorSeg (Qian et al., 2026) reformulates reasoning segmentation using a structured query bank, where anchor queries provide spatial localization and contextual queries encode semantic information. FlowSeg (Zhang et al., 2026) replaces static language prompting with bidirectional semantic coevolution that progressively refines condition embeddings and visual queries. SAM-Veteran (Du et al., 2026) replaces embedding-based mask prompting with iterative box-and-point interactions with the Segment Anything Model (SAM). However, all of the approaches described above still have a limitation in that they cannot explicitly disentangle and represent the target's region, object boundary, and background as separate pieces of information.

Vision Language Models (VLMs). The success of large language models (LLMs) has led to the development of VLMs that jointly process visual and textual information. Flamingo (Alayrac et al., 2022), one of the early VLMs pretrained on large-scale vision-language data, demonstrated strong image-text understanding. Subsequently, BLIP-2 (Li et al., 2023) connected a frozen image encoder to an LLM through a Querying Transformer (Q-Former), while MiniGPT-4 (Zhu et al., 2024) aligned visual features with the language space through a simple linear projection. LLaVA (Liu et al., 2023b) efficiently built a VLM with strong visual reasoning capabilities using GPT-4-generated instruction-following data and a simple architecture that connects a CLIP (Radford et al., 2021) vision encoder to an LLM through a linear projector. LLaVA-1.5 (Liu et al., 2024) further improved performance using a multi-layer perceptron projector and higher-resolution image inputs.

3 Motivation and Main Idea

Most existing reasoning segmentation methods (Lai et al., 2024; Qian et al., 2025) combine a vision-language model (VLM) with the Segment Anything Model (SAM). The VLM reasons over the input image and query, while SAM generates the segmentation mask. Specifically, the embedding of the ⟨SEG⟩ token generated by the VLM is fed into the SAM decoder for reasoning mask prediction. However, as shown in Fig. 1(a), a single ⟨SEG⟩ token, which mainly represents the target region, provides insufficient information to predict precise object boundaries and avoid false positives in the background.

To address these issues, we propose SEmO, which consists of two key designs. First, we divide the role of the conventional ⟨SEG⟩ token across three specialized tokens: ⟨SEG⟩ for the target region, ⟨EDGE⟩ for the object boundary, and ⟨OTHERS⟩ for the background. This design allows the tokens to specialize in complementary aspects of mask prediction, as illustrated in Fig. 2. Second, we design a Progressive Mask Decoder (PMD) that integrates the token embeddings across multiple resolutions instead of predicting the final mask in a single decoding step. Starting from a coarse estimate, each stage refines the previous-stage prediction.

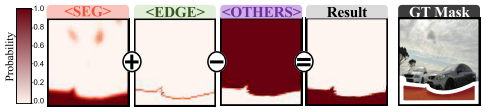

Figure 2: Example of role-specific logit aggregation. The aggregated logit map is obtained by adding the edge logits ⟨EDGE⟩ to the target-region logits ⟨SEG⟩ and subtracting the background logits ⟨OTHERS⟩.

4 Proposed Methods

As illustrated in Fig. 3, SEmO consists of two main components: a vision-language model (VLM) and a Progressive Mask Decoder (PMD). The VLM takes an input image and a text query and produces three specialized token embeddings: ⟨SEG⟩, ⟨EDGE⟩, and ⟨OTHERS⟩, representing the target region, object boundary, and background, respectively (Sec. 4.1). The PMD progressively integrates these token embeddings with multi-resolution image features to generate the final reasoning mask (Sec. 4.2).

4.1 Role-Specific Token Generation

A vision-language model (VLM), built on LLaVA (Liu et al., 2023b), takes an input image x_{img} and a text query x_{txt} and generates a textual response $\hat{y}_{\text{txt}}$, as illustrated in Fig. 3(a). While existing reasoning segmentation methods (Lai et al., 2024) generate a single ⟨SEG⟩ token and use its embedding as a prompt for mask prediction, SEmO generates three role-specific tokens, ⟨SEG⟩, ⟨EDGE⟩, and ⟨OTHERS⟩, in the following response:

"It is ⟨SEG⟩, its edge is ⟨EDGE⟩, and the rest is ⟨OTHERS⟩."

The three tokens represent complementary semantic roles: ⟨SEG⟩ represents the target region, ⟨EDGE⟩ its edge, and ⟨OTHERS⟩ the background. This explicit role assignment allows each token to learn a distinct semantic representation.

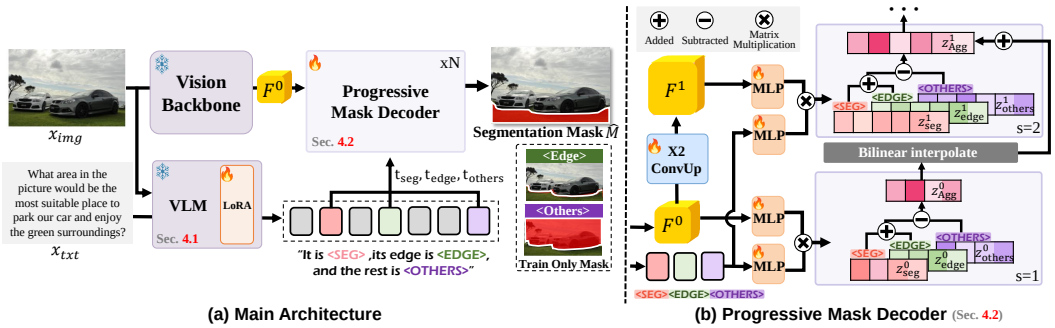

Figure 3: Overall architecture of SEmO. (a) The VLM generates three role-specific tokens representing the target region, ⟨SEG⟩, object edge ⟨EDGE⟩, and background ⟨OTHERS⟩, which are passed to the Progressive Mask Decoder (PMD). (b) The PMD predicts role-specific logit maps from the three tokens, aggregates them by reinforcing the target prediction with boundary cues while suppressing background responses, and progressively refines the mask across N resolution stages.

Recent studies (Qian et al., 2025; Xia et al., 2024) have improved reasoning segmentation by introducing auxiliary visual prompts or multiple segmentation tokens. However, these approaches leave the role of each ⟨SEG⟩ token unchanged, with each token serving as a unified representation of its corresponding target. In contrast, SEmO assigns the target region, object boundary, and background to separate role-specific tokens. Unlike the ⟨REJ⟩ token in GSVA (Xia et al., 2024), which indicates the absence of the queried object, the ⟨OTHERS⟩ token in SEmO explicitly represents the background even when the target object is present.

From the last hidden layer of the VLM, we extract the embeddings corresponding to the three tokens, denoted as t_k, where $k \in \{\text{seg}, \text{edge}, \text{others}\}$. These token embeddings are then fed into the Progressive Mask Decoder (PMD) (Sec. 4.2), where they are progressively fused with multi-resolution image features to predict the segmentation mask.

4.2 Progressive Mask Decoder (PMD)

Instead of predicting the final mask in a single decoding step, the Progressive Mask Decoder (PMD) adopts a coarse-to-fine decoding process in which each stage refines the prediction from the previous stage. PMD takes as input the three token embeddings t_k from the VLM (Sec. 4.1) together with the image features F^0 extracted by the vision backbone. Starting from the spatial resolution of F^0, PMD progressively integrates t_k with image features over N stages of increasing spatial resolution to produce the final segmentation mask. As shown in Fig. 3(b), each stage consists of three sequential operations: (1) computing role-specific logit maps from the token embeddings and image features, (2) aggregating the role-specific logit maps, and (3) propagating the aggregated logits to the next stage as a prior for progressive refinement.

Role-Specific Logit Computation. Given the image feature F^s at resolution stage s, where $s = 0, \ldots, N-1$, we project the three token embeddings and image features into a shared d-dimensional embedding space using multilayer perceptrons (MLPs). The projected token embeddings are stacked into a matrix T^s as

$$T^s = [\text{MLP}^s(t_{\text{seg}}); \text{MLP}^s(t_{\text{edge}}); \text{MLP}^s(t_{\text{others}})] \in \mathbb{R}^{3\times d}. \tag{1}$$

Next, the projected image feature is spatially flattened as

$$X^s = \text{Flatten}\left(\text{MLP}^s_{\text{img}}(F^s)\right) \in \mathbb{R}^{H_s W_s \times d}. \tag{2}$$

The role-specific logit maps are then computed through matrix multiplication and reshaping:

$$z^s = \text{Reshape}\left(X^s (T^s)^\top\right) \in \mathbb{R}^{H_s \times W_s \times 3}, \tag{3}$$

where each channel of z^s corresponds to the target region z^s_{seg}, object edge z^s_{edge}, and background z^s_{others}, respectively.

Role-Specific Logit Aggregation. The three role-specific logit maps contribute to the final prediction in different directions. We use z^s_{seg} as the primary target prediction, add z^s_{edge} to reinforce edge regions, and subtract z^s_{others} to suppress background responses. Accordingly, the edge logit z^s_{edge} is added to the primary target logit z^s_{seg}, while z^s_{others} is subtracted to suppress background responses by:

$$z^s_{\text{Agg}} = \alpha^s z^s_{\text{seg}} + \beta^s z^s_{\text{edge}} - \gamma^s z^s_{\text{others}}, \tag{4}$$

where α^s, β^s, and γ^s are learnable weights at stage s. Figure 2 visually illustrates the complementary contributions of the three role-specific logit maps to the proposed aggregation.

Progressive Logit Propagation and Refinement. The role-specific logit computation and aggregation described above are performed progressively across resolution stages.

At stage s, the role-specific logit maps are computed from F^s and aggregated into z^s_{Agg}. The image feature and aggregated logits are then propagated to stage $s{+}1$, where z^s_{Agg} serves as a prior for refining the current prediction. Specifically, the image feature F^s is upsampled using the transposed convolution as $F^{s+1} = \text{ConvUp}(F^s)$, providing higher-resolution image features for the next stage. Using F^{s+1}, stage $s{+}1$ performs the same role-specific matching and aggregation as described above to obtain z^{s+1}_{Agg}. The aggregated logits from the previous stage are then resized to the spatial resolution of the current stage using bilinear interpolation (Bilinear) and incorporated as a prior as

$$z^{s+1}_{Agg} \leftarrow z^{s+1}_{Agg} + \text{Bilinear}(z^s_{Agg}). \tag{5}$$

The propagated logits preserve the coarse target localization established at the previous stage, while the newly computed logits refine object edges and fine-grained structures using higher-resolution image features. Consequently, each stage progressively refines the prediction from the previous stage, rather than independently estimating the complete segmentation mask at every stage. The refinement process is repeated until the final resolution, and the resulting aggregated logits are converted into the final mask $\hat{M}$.

4.3 Training Objectives

The model is trained end-to-end using three objectives: a text generation loss, a mask loss, and a role supervision loss. Since the training datasets provide only ground-truth segmentation masks, the supervision targets for the three roles are derived from the ground-truth mask M:

$$M_{\text{seg}} = M,\ M_{\text{edge}} = M - \text{erode}(M),\ M_{\text{others}} = 1 - M, \tag{6}$$

where M_{edge} denotes the inner boundary band obtained through morphological erosion, and M_{others} denotes the complement of the target mask. For each resolution stage s, these targets are resized to the corresponding spatial resolution, yielding M^s_k.

Text generation loss. The text generation loss is the standard autoregressive cross-entropy loss computed between the predicted textual response $\hat{y}_{\text{txt}}$ and the ground-truth response y_{txt}, which contains the three role-specific tokens:

$$\mathcal{L}_{\text{txt}} = \mathcal{L}_{\text{CE}}(\hat{y}_{\text{txt}}, y_{\text{txt}}). \tag{7}$$

Mask loss. Following LISA (Lai et al., 2024), we use a combination of BCE and Dice losses:

$$\ell(\hat{M}, M) = \lambda_{\text{bce}}\mathcal{L}_{\text{BCE}}(\hat{M}, M) + \lambda_{\text{dice}}\mathcal{L}_{\text{DICE}}(\hat{M}, M). \tag{8}$$

The mask loss is then defined as $\mathcal{L}_{\text{mask}} = \ell(\hat{M}, M)$.

Role supervision loss. To encourage each role-specific prediction to capture its assigned semantics, we apply deep supervision to role-specific logit maps at every resolution stage:

$$\mathcal{L}_{\text{role}} = \sum_{s=0}^{N-1} \sum_{k} \ell(z^s_k, M^s_k), \tag{9}$$

where $k \in \{\text{seg}, \text{edge}, \text{others}\}$ and $\ell(\cdot,\cdot)$ is the loss function defined in Eq. 8. This deep supervision improves the intermediate predictions before they are propagated across resolution stages, while guiding each role-specific prediction to represent the target region, object boundary, or background.

Table 1: Performance comparison of SEmO with existing methods on ReasonSeg. "ft" denotes fine-tuning on reasoning segmentation data. The best and second-best results are highlighted in bold and underlined, respectively. SEmO achieves the best performance on both splits.

Method	Backbone	val (overall)		test (short)		test (long)		test (overall)	
		gIoU	cIoU	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU
(Generalist / specialist segmentation models)									
X-Decoder (2023a)	UniCL	22.6	17.9	20.4	11.6	22.2	17.5	21.7	16.3
Grounded-SAM (2024a)	–	26.0	14.5	17.8	10.8	22.4	18.6	21.3	16.4
SEEM (2023b)	UniCL	25.5	21.2	20.1	11.5	25.6	20.8	24.3	18.7
OVSeg (2023)	CLIP ViT-L	28.5	18.6	18.0	15.5	28.7	22.5	26.1	20.8
GRES (2023a)	BERT	22.4	19.9	17.6	15.0	22.6	23.8	21.3	22.0
(LMM-based methods)									
SESAME (2024b)	LLaVA-7B	34.8	39.1	28.3	27.6	31.6	32.7	30.5	30.4
LLaVA1.5-7B + OVSeg (2024; 2023)	LLaVA1.5-7B	38.2	23.5	24.2	18.7	44.6	37.1	39.7	31.8
LISA (2024)	LLaVA-7B	44.4	–	37.6	–	36.6	–	36.8	–
CoReS (2024)	LLaVA-7B	54.8	–	41.0	–	50.9	–	48.7	–
LISA (ft) (2024)	LLaVA-7B	52.9	54.0	40.6	40.6	49.4	51.0	47.3	48.4
CoReS (ft) (2024)	LLaVA-7B	59.4	–	44.2	–	55.0	–	52.4	–
LISA (ft) (2024)	LLaVA1.5-7B	61.3	62.9	48.3	46.3	57.9	59.7	55.6	56.9
READ (ft) (2025)	LLaVA1.5-7B	59.8	67.6	52.6	49.5	60.4	61.0	58.5	58.6
RSVP-GPT (2025)	GPT-4o	<u>64.7</u>	<u>63.1</u>	<u>55.4</u>	<u>50.4</u>	<u>61.9</u>	<u>62.5</u>	60.3	<u>60.0</u>
FlowSeg (2026)	Qwen3-4B	62.4	49.2	-	-	-	-	<u>60.5</u>	54.7
Ours (ft)	LLaVA1.5-7B	**65.9**	**69.1**	**55.8**	**51.3**	**63.2**	**63.8**	**62.0**	**61.1**

Overall loss. The overall loss is defined as

$$\mathcal{L} = \lambda_{\text{txt}}\mathcal{L}_{\text{txt}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} + \lambda_{\text{role}}\mathcal{L}_{\text{role}}, \tag{10}$$

where λ_{txt}, λ_{mask}, and λ_{role} balance the three objectives.

5 Experimental Results

5.1 Experiment Settings

Datasets. We train our model using datasets from three categories: (1) semantic segmentation datasets, including ADE20K (Zhou et al., 2017), COCO-Stuff (Caesar et al., 2018), PACO-LVIS (Ramanathan et al., 2023), and PASCAL-Part (Chen et al., 2014); (2) reasoning and referring segmentation datasets, including ReasonSeg (Lai et al., 2024), refCLEF, RefCOCO, RefCOCO+ (Kazemzadeh et al., 2014), and RefCOCOg (Mao et al., 2016); and (3) the visual instruction tuning dataset LLaVA-Instruct 150K (Liu et al., 2023b). We additionally incorporate FP-RefCOCO(+/g) (Wu et al., 2024b) and R-RefCOCO (Wu et al., 2024a) to improve robustness against false-premise queries and mismatched referring expressions.

Evaluation Metrics. Following the evaluation protocol of prior works (Lai et al., 2024; Qian et al., 2025), we report global intersection-over-union (gIoU) and cumulative IoU (cIoU). The gIoU is computed by averaging the IoU over individual samples, whereas the cIoU is computed by accumulating the intersections and unions over the entire dataset and then taking their ratio. Since cIoU is more influenced by samples containing large target regions, we use gIoU as the primary evaluation metric because it provides a more balanced assessment across samples with different target sizes.

Implementation Details. We adopt LLaVA-1.5-7B (Liu et al., 2024) as our vision-language model (VLM) and employ the ViT-H image encoder from the Segment Anything Model (SAM) (Kirillov et al., 2023) to extract image features for mask prediction. All experiments are conducted using the DeepSpeed (Rasley et al., 2020) on four NVIDIA H200 GPUs. We use the AdamW optimizer with an initial learning rate of 5×10^{-4} and a learning-rate scheduler with 500 warm-up steps. The loss weights are set to $\lambda_{\text{txt}} = 1$ for text generation, $\lambda_{\text{mask}} = 2$ for mask prediction, and $\lambda_{\text{role}} = 1$ for role-specific supervision. Following LISA (Lai et al., 2024), we set $\lambda_{\text{bce}} = 2.0$ and $\lambda_{\text{dice}} = 0.5$. We freeze the base parameters of the VLM and train LoRA adapters with 16 rank.

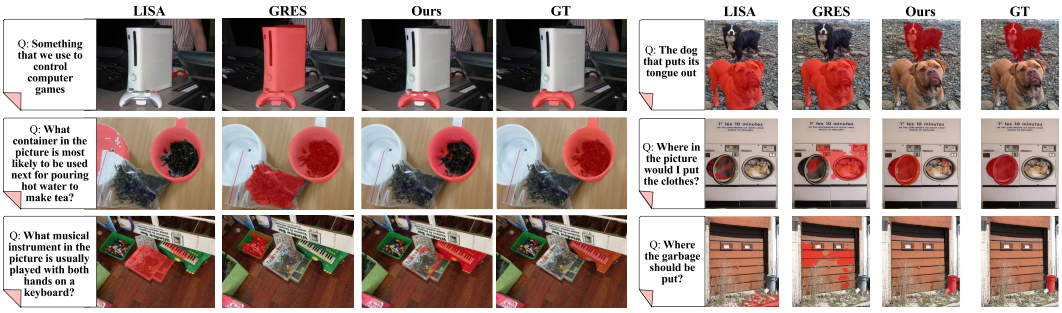

Figure 4: Qualitative comparison of LISA, GRES, and SEmO (Ours) on the ReasonSeg dataset. Ground-truth masks and model predictions are highlighted in red. SEmO predicts edges more accurately and suppresses false-positive regions more effectively (see supplementary for more results).



Figure 5: Qualitative comparison of LISA, GRES, and SEmO on the RefCOCO dataset. Ground-truth masks and model predictions are highlighted in red. SEmO more accurately localizes the referred objects and produces clear masks.

5.2 Comparison with State-of-the-Art Methods

Reasoning Segmentation. We compare SEmO with existing state-of-the-art methods on the ReasonSeg benchmark, as shown in Tab. 1. SEmO achieves the best performance in terms of both gIoU and cIoU on the validation and test sets. Compared with the previous best-performing method, RSVP-GPT (Lu et al., 2025), SEmO improves gIoU and cIoU by 1.2 and 6.0 points, respectively, on the validation set, and by 1.7 and 1.1 points on the overall test set. These results demonstrate the effectiveness of SEmO for reasoning segmentation. We further analyze the individual contributions of the role-specific tokens and the Progressive Mask Decoder.

Figure 4 shows qualitative comparisons on ReasonSeg, where complex reasoning is required to identify the target object. Compared with previous methods, SEmO produces masks that more accurately correspond to the reasoning query, with sharper object boundaries and fewer false-positive regions on visually similar distractors. Figure 5 further shows qualitative results on referring segmentation. Our method yields more precise masks, particularly for objects with thin structures or partial occlusion. Additional qualitative results are provided in the supplementary material.

Referring segmentation. We compare SEmO with existing state-of-the-art methods on the RefCOCO, RefCOCO+, and RefCOCOg benchmarks, as shown in Tab. 2. SEmO achieves a cIoU of 78.3% on the RefCOCO validation split and 68.9% on the RefCOCO+ validation split, surpassing the previous state of the art by 0.2 and 0.5 percentage points, respectively. It also achieves the best performance on both the validation and test splits of RefCOCOg. These results demonstrate that SEmO, although primarily designed for reasoning segmentation, generalizes effectively to referring expression segmentation.

Table 2: Performance comparison on referring segmentation datasets. Best and second-best in bold and underlined.

Method \ Dataset	RefCOCO			RefCOCO+			RefCOCOg	
	val	testA	testB	val	testA	testB	val(U)	test(U)
VLT (2022)	67.5	70.5	65.2	56.3	61.0	50.1	55.0	57.7
CRIS (2022)	70.5	73.2	66.1	62.3	68.1	53.7	59.9	60.4
LAVT (2022)	72.7	75.8	68.8	62.1	68.4	55.1	61.2	62.1
ReLA (2023a)	73.8	76.5	70.2	66.0	71.0	57.7	65.0	66.0
LISA (2024)	74.9	79.1	72.3	65.1	70.8	58.1	67.9	70.6
CoReS (2024)	76.0	78.6	72.5	65.1	70.0	58.6	69.0	70.7
READ (2025)	<u>78.1</u>	**80.2**	<u>73.2</u>	<u>68.4</u>	<u>73.7</u>	<u>60.4</u>	<u>70.1</u>	<u>71.4</u>
Ours	**78.3**	<u>80.1</u>	**73.7**	**68.9**	**74.4**	**61.3**	**70.8**	**72.2**

Table 3: Comparison of decoder parameters and inference time. SEmO achieves superior performance with the fewest decoder parameters and the shortest inference time.

Method	Decoder Params(M)	Inference Time(ms)
LISA (2024)	4.06	3.12
PixelLM (2024b)	7.57	2.81
Ours	**0.69**	**2.72**

Table 4: Ablation studies on ReasonSeg. (a) Effect of removing each role-specific token. (b) Effect of varying the number of refinement stages (N). Using all three role-specific tokens and three refinement stages achieves the best performance.

(a) Effect of removing each role-specific token

Variant	val (overall)		test (short)		test (long)		test (overall)	
	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU
w/o ⟨EDGE⟩	59.2	63.8	50.1	48.8	58.0	60.7	57.2	58.4
w/o ⟨OTHERS⟩	60.1	65.4	52.2	49.9	60.7	61.8	58.8	58.7
Full (Ours)	**65.9**	**69.1**	**55.8**	**51.3**	**63.2**	**63.8**	**62.0**	**61.1**

(b) Effect of the number of refinement stages (N)

N	val (overall)		test (short)		test (long)		test (overall)	
	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU
1	57.8	61.3	47.1	45.5	55.2	53.9	54.3	55.2
2	61.2	68.6	53.4	49.3	61.7	60.3	58.8	59.9
3	**65.9**	**69.1**	**55.8**	**51.3**	**63.2**	**63.8**	**62.0**	**61.1**
4	60.1	68.3	52.4	49.2	58.8	59.0	56.8	57.1

Decoder Efficiency. As shown in Tab. 3, we compare the proposed Progressive Mask Decoder (PMD) with the original SAM mask decoder (Kirillov et al., 2023) and the lightweight decoder of PixelLM (Ren et al., 2024b) in terms of the number of decoder parameters and inference time. The inference time is measured as the average per-image decoder processing time on the ReasonSeg validation set. Despite achieving superior segmentation performance, PMD requires only 0.69M decoder parameters, reducing the parameter count by 3.37M (83.0% ↓) compared with the original SAM decoder. PMD also achieves the shortest inference time of 2.72 ms, compared with 3.12 ms for the SAM decoder and 2.81 ms for the PixelLM decoder.

These results demonstrate that PMD achieves a better trade-off between segmentation accuracy and computational efficiency than the compared mask decoders. This efficiency is attributed to its lightweight architecture, in which each refinement stage performs only feature projection and similarity matching, without computationally expensive attention operations as shown in Fig. 3(b).

5.3 Ablation Studies and Model Analysis

Effect of role-specific tokens. To evaluate the contribution of each role-specific token, we conduct an ablation study by individually removing the ⟨EDGE⟩ and ⟨OTHERS⟩ tokens on the ReasonSeg validation set as shown in Tab. 4(a). For each variant, the corresponding similarity-matching operation, aggregation term, and role-specific supervision are removed, while all other components remain unchanged.

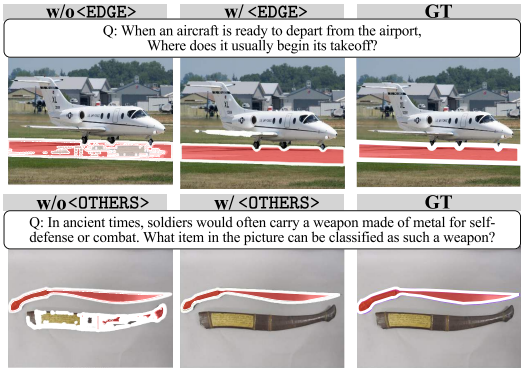

Figure 6: Effect of the ⟨EDGE⟩ and ⟨OTHERS⟩ tokens. Removing ⟨EDGE⟩ degrades edge quality (1st row). Removing ⟨OTHERS⟩ causes false activation (2nd row).

Removing the ⟨EDGE⟩ token results in the largest performance degradation, reducing gIoU from 65.9% to 59.2%, which highlights the importance of explicitly modeling object boundaries. Removing the ⟨OTHERS⟩ token also decreases gIoU to 60.1%, indicating that explicitly modeling non-target regions helps reduce false activations. Using all three role-specific tokens achieves the best performance, demonstrating that boundary modeling and background suppression provide complementary benefits. These effects are further illustrated in Fig. 6.

Table 5: Performance comparison on ReasonSeg across different image aspect ratios.

Method	Near-square		Landscape		Portrait	
	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU
LISA (2024)	0.44	0.39	0.48	0.62	0.44	0.44
Ours	**0.47**	**0.44**	**0.60**	**0.71**	**0.55**	**0.56**

Number of refinement stages. We analyze the effect of the number of progressive refinement stages, denoted by N. As shown in Tab. 4(b), increasing the number of stages from $N{=}1$, where the mask is predicted at a single resolution, to $N{=}2$ improves gIoU by 3.4 points, demonstrating the benefit of progressively refining coarse predictions using higher-resolution image features. The performance further improves and reaches its highest value at $N{=}3$. Increasing the number of stages to $N{=}4$ yields no further improvement, suggesting that three refinement stages are sufficient. Therefore, we use $N{=}3$ in all experiments.

Performance across image aspect ratios. To investigate performance across different image aspect ratios, we compare SEmO with LISA on the ReasonSeg test set by grouping images into near-square ($0.9 \leq \text{width}/\text{height} \leq 1.1$), landscape ($\text{width} > \text{height}$), and portrait ($\text{height} > \text{width}$) categories according to their aspect ratios. As shown in Tab. 5, SEmO consistently outperforms LISA across all groups. Notably, the performance gains are substantially larger for landscape and portrait images than for near-square images. Specifically, SEmO improves gIoU by 12.0 and 11.0 points for landscape and portrait images, respectively, compared with a 3.0-point improvement for near-square images.

These results suggest that SEmO is effective for images with non-square aspect ratios. We attribute this advantage in part to Progressive Mask Decoder (PMD), which progressively refines predictions across multiple resolutions, allowing coarse predictions to be successively corrected using finer spatial details. Such progressive refinement may be especially beneficial for landscape and portrait images, where spatial details can be distributed unevenly across the image.

6 Conclusions and Future Work

In this work, we presented SEmO, a reasoning segmentation framework that addresses the limitations of existing embedding-as-mask approaches through role-specific token representations and progressive mask refinement. By explicitly representing the target region, its edge, and the background with three specialized tokens and progressively refining their aggregated predictions across multiple resolutions, SEmO produces more accurate segmentation masks while effectively suppressing false positives. Experimental results demonstrate state-of-the-art performance on the ReasonSeg benchmark and consistent improvements on the RefCOCO, RefCOCO+, and RefCOCOg benchmarks.

Despite these promising results, several limitations remain. Our current framework uses a fixed edge-label generation strategy and a fixed number of refinement stages for all images, which may not be optimal for targets with diverse structures or scenes of varying complexity. Future work will explore adaptive edge modeling and dynamic refinement strategies that adjust the refinement process according to the input. We also plan to extend the proposed role-specific representation to more challenging reasoning segmentation settings, such as multi-instance reasoning and complex visual reasoning scenarios.

AI use statement

In this work, we used generative AI tools (large language models) for writing assistance, including improving the readability of the manuscript, condensing text to meet the page limit, suggesting a structure for the appendix, and adjusting LaTeX formatting of tables and figures. Additionally, we used generative AI tools to help identify the cause of coding errors encountered during implementation. We have not used generative AI tools to generate synthetic data, develop the theoretical framework, formulate or prove mathematical claims, design experiments, clean or reformat datasets,

analyze data, or interpret results. Translation and qualitative/thematic analysis are not applicable to this work. All AI-assisted debugging suggestions were verified and tested for correctness by the two co-first authors before being incorporated into the codebase. We have also reviewed all AI-assisted edits to ensure that the technical content, claims, and numerical results remained unchanged from the original author-written text. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

Reproducibility statement

We have made efforts to ensure that our results are reproducible. SEmO, including the role-specific token design (⟨SEG⟩, ⟨EDGE⟩, ⟨OTHERS⟩) and the Progressive Mask Decoder (PMD), is described in Sec. 4; the training objectives, including the mask loss combining BCE and Dice losses, are given in Sec. 4.3. Datasets, evaluation protocols, and implementation details (backbone architectures, LoRA configuration, optimizer, and hardware) are specified in Sec. 5.1. All experiments use the publicly available `ReasonSeg`, `RefCOCO`, `RefCOCO+`, and `RefCOCOg` datasets with standard evaluation protocols. Additional details, including pseudocode and the adaptive edge-label generation procedure, are provided in the Appendix. Source code and pretrained checkpoints will be released upon publication.

References

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

Xiaoyi Bao, Siyang Sun, Shuailei Ma, Kecheng Zheng, Yuxin Guo, Guosheng Zhao, Yun Zheng, and Xingang Wang. Cores: Orchestrating the dance of reasoning and segmentation. In *European Conference on Computer Vision*, pp. 187–204. Springer, 2024.

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

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

Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vlt: Vision-language transformer and query generation for referring segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Tianyuan Du, Haopeng Li, Zhen Fan, Jiarui Zhang, Panwang Pan, and Yang Zhang. Sam-veteran: An mllm-based human-like sam agent for reasoning segmentation. In *The Fourteenth International Conference on Learning Representations*, 2026.

Donggon Jang, Yucheol Cho, Suin Lee, Taehyeon Kim, and Dae Shik Kim. Mmr: A large-scale benchmark dataset for multi-target and multi-granularity reasoning segmentation. In *International Conference on Learning Representations*, volume 2025, pp. 67406–67431, 2025.

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

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

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9579–9589, 2024.

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

Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7061–7070, 2023.

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

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024.

Yi Lu, Jiawang Cao, Yongliang Wu, Bozheng Li, Licheng Tang, Yangguang Ji, Chong Wu, Jay Wu, and Wenbo Zhu. Rsvp: Reasoning segmentation via visual prompting and multi-modal chain-of-thought. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14699–14716, 2025.

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

Rui Qian, Xin Yin, and Dejing Dou. Reasoning to attend: Try to understand how ⟨seg⟩ token works. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24722–24731, 2025.

Rui Qian, Chuanhang Deng, Qiang Huang, Jian Xiong, Mingxuan Li, Yingbo Zhou, Wei Zhai, Jintao Chen, and Dejing Dou. Anchorseg: Language grounded query banks for reasoning segmentation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20490–20505, 2026.

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

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

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 3505–3506, 2020.

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

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

Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11686–11695, 2022.

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

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

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

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

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

Zekang Zhang, Guangyu Gao, Youyun Tang, ChengJing Wu, Xiaochao Qu, Chi Harold Liu, Jianbo Jiao, Yunchao Wei, Luoqi Liu, and Ting Liu. Flowseg: Dynamic semantic guidance for llm-conditioned segmentation. *arXiv preprint arXiv:2605.29461*, 2026.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 633–641, 2017.

Deyao Zhu, Xiaoqian Shen, Xiang Li, Mohamed Elhoseiny, et al. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations*, volume 2024, pp. 18378–18394, 2024.

Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15116–15127, 2023a.

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

A Overview

The appendix is organized as follows:

- Section B: Additional implementation details, including the complete forward process of PMD.
- Section C: Adaptive edge-label generation for defining the ⟨EDGE⟩ supervision target.
- Section D: Additional quantitative results on generalization across model scales.
- Section E: Additional qualitative results on `ReasonSeg`.
- Section F: Visualization and analysis of representative failure cases.
- Section G: Discussion of the limitations of SEmO and directions for future work.

B Additional Implementation Details

Complete Forward Process. At each resolution stage, Progressive Mask Decoder (PMD) computes role-specific logit maps, aggregates them into a local prediction, and combines the local prediction with the aggregated logits propagated from the preceding stage. The refinement process is repeated until the final resolution, and a sigmoid function converts the final aggregated logits into the mask probability map $\hat{M}$. Algorithm 1 summarizes the complete forward process across all N resolution stages.

During training, $\hat{M}$ is used as a mask probability map. During evaluation, the binary mask is obtained by thresholding $\hat{M}$ at 0.5.

C Adaptive Edge-Label Generation

The supervision target for the ⟨EDGE⟩ token is derived from the ground-truth mask M using morphological erosion with a fixed 3×3 kernel. Rather than using a fixed number of erosion steps, we adaptively select the number of iterations according to the relative area of the resulting inner boundary band. The notation $\mathrm{erode}(M)$ used in the main paper refers to the adaptively iterated erosion operation detailed in this section.

For a non-empty ground-truth mask, let $\mathrm{erode}^t(\cdot)$ denote t successive erosion operations. The candidate edge target and its relative area are defined as

$$M_{\mathrm{edge}}^{(t)} = M - \mathrm{erode}^t(M), \qquad r(t) = \frac{|M_{\mathrm{edge}}^{(t)}|}{|M|}. \tag{11}$$

Here, $|\cdot|$ denotes the number of foreground pixels. Since $r(t)$ is non-decreasing with respect to t, we select the largest number of iterations for which the relative edge area does not exceed $\tau = 0.3$, subject to a maximum of $T_{\max} = 12$:

$$t^* = \max\left(\{t \in \{1, \ldots, T_{\max}\} \mid r(t) \le \tau\} \cup \{1\}\right). \tag{12}$$

The final edge supervision target is defined as

$$M_{\mathrm{edge}} = M_{\mathrm{edge}}^{(t^*)} = M - \mathrm{erode}^{t^*}(M). \tag{13}$$

The inclusion of $\{1\}$ in Eq. 12 ensures that t^* remains defined even when the first erosion step already exceeds the threshold.

For an empty ground-truth mask, erosion is not performed, and the role-specific targets are defined as

$$M_{\mathrm{seg}} = \mathbf{0}, \qquad M_{\mathrm{edge}} = \mathbf{0}, \qquad M_{\mathrm{others}} = \mathbf{1}. \tag{14}$$

All role-specific supervision targets depend only on the ground-truth mask and are computed once during preprocessing.

Figure 7 illustrates three examples of the adaptive edge-label generation process. For the small and medium targets, the largest valid iteration numbers are $t^* = 2$ and $t^* = 5$, respectively. For the large

Algorithm 1 Progressive Mask Decoder (PMD) Forward Process

Require: Token embeddings $t_{\text{seg}}, t_{\text{edge}}, t_{\text{others}}$; initial image feature F^0; number of stages N
Ensure: Final mask probability map $\hat{M}$

for $s = 0$ **to** $N-1$ **do**
  $T^s \leftarrow [\text{MLP}^s(t_{\text{seg}}); \text{MLP}^s(t_{\text{edge}}); \text{MLP}^s(t_{\text{others}})]$
  $X^s \leftarrow \text{Flatten}(\text{MLP}^s_{\text{img}}(F^s))$
  $z^s \leftarrow \text{Reshape}(X^s (T^s)^\top)$
  $z^s_{\text{seg}}, z^s_{\text{edge}}, z^s_{\text{others}} \leftarrow z^s[:,:,0], z^s[:,:,1], z^s[:,:,2]$
  $\tilde{z}^s_{\text{Agg}} \leftarrow \alpha^s z^s_{\text{seg}} + \beta^s z^s_{\text{edge}} - \gamma^s z^s_{\text{others}}$
  if $s = 0$ **then**
    $z^s_{\text{Agg}} \leftarrow \tilde{z}^s_{\text{Agg}}$
  else
    $z^s_{\text{Agg}} \leftarrow \tilde{z}^s_{\text{Agg}} + \text{Bilinear}(z^{s-1}_{\text{Agg}}; H^s, W^s)$
  end if
  if $s < N-1$ **then**
    $F^{s+1} \leftarrow \text{ConvUp}(F^s)$
  end if
end for
$\hat{M} \leftarrow \sigma(z^{N-1}_{\text{Agg}})$
return $\hat{M}$

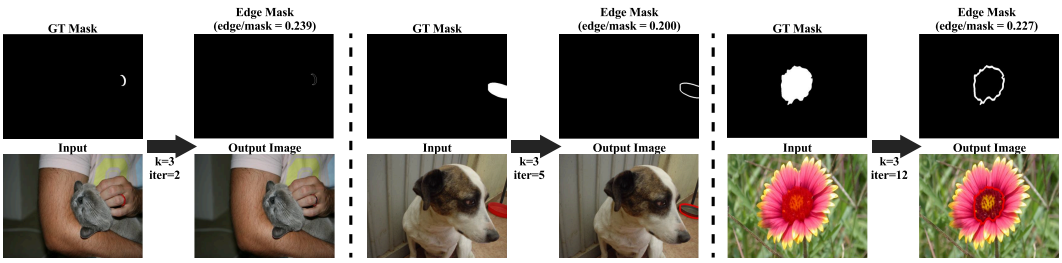

Figure 7: Adaptive edge-label generation using a fixed 3×3 erosion kernel. For each non-empty target mask, the number of erosion iterations is selected as the largest valid value for which the relative edge area does not exceed $\tau = 0.3$, with a maximum of $T_{\max} = 12$.

target, the relative edge area remains below τ until the maximum number of iterations, resulting in $t^* = T_{\max} = 12$. The absolute boundary-band width therefore adapts to the target size while its relative area remains bounded by the predefined threshold whenever possible.

D Additional Quantitative Results

Generalization across model scales (13B backbone).

Table 6: Generalization across model scales. Comparison between LISA and SEmO using a LLaVA-13B backbone on `ReasonSeg`.

Method	val (overall)		test (overall)	
	gIoU	cIoU	gIoU	cIoU
LLaVA1.5-13B + OVSeg	37.9	26.4	41.5	34.1
LISA-13B-LLaVA1.5 (ft)	65.0	72.9	61.3	62.2
READ-13B-LLaVA1.5 (ft)	-	-	62.2	62.8
Ours (ft)	**68.6**	**76.1**	**64.2**	**65.9**

Table 6 compares SEmO with LISA (Lai et al., 2024) and READ (Qian et al., 2025) using a larger LLaVA-13B backbone on `ReasonSeg`, complementing the 7B results reported in the main paper. SEmO continues to outperform all compared methods across both the validation and test sets,



Figure 8: Qualitative comparison of LISA and SEmO (Ours) on the `ReasonSeg` dataset. Ground-truth masks and model predictions are highlighted in red. Compared with LISA, SEmO predicts target boundaries more accurately and suppresses false-positive regions more effectively. Best viewed in color.

achieving gIoU and cIoU improvements of 3.6 and 3.2 points over the previous best-performing method on the validation set, and 2.0 and 3.1 points on the test set. These results indicate that the performance gains of SEmO are not specific to a particular backbone scale, and that the proposed role-specific token design and Progressive Mask Decoder generalize effectively to a larger vision-language model.

E Additional Qualitative Results

Figure 8 presents additional qualitative comparisons between LISA (Lai et al., 2024) and SEmO on the `ReasonSeg` dataset, complementing the results shown in the main paper. Across diverse reasoning queries, SEmO consistently produces masks that more precisely follow object boundaries and better suppress false-positive activations on visually similar or adjacent regions. For queries requiring fine-grained part-level reasoning, such as identifying the stem responsible for transporting nutrients within a plant, SEmO correctly localizes the thin, elongated target structure, whereas LISA (Lai et al., 2024) fails to capture such fine-grained parts and instead produces coarse or misplaced predictions. These results further demonstrate that the role-specific token design and the Progressive Mask Decoder (PMD) generalize effectively across a wide range of reasoning scenarios, including cases involving object function, spatial relations, and part-level structural details.

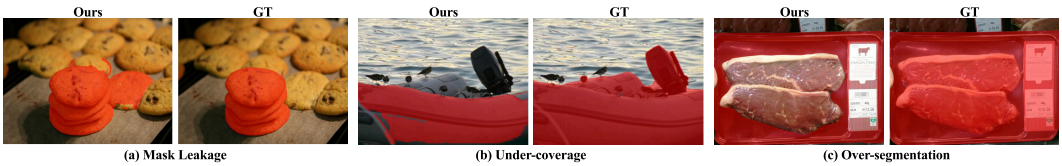

Figure 9: Representative failure cases of SEmO on ReasonSeg. (a) Mask leakage onto adjacent objects with similar appearance. (b) Under-coverage of the target region. (c) Over-segmentation extending beyond the target boundary.

F Failure Case Analysis

Figure 9 illustrates three representative failure patterns observed in SEmO's predictions. In Fig. 9(a), the predicted mask leaks onto an adjacent cookie with a highly similar appearance and texture to the target, indicating that the ⟨OTHERS⟩ token can fail to suppress background responses when distractor objects are visually near-identical to the target. In Fig. 9(b), the model under-covers the target object, missing a small protruding part; this suggests that thin or loosely connected structures remain challenging for the ⟨EDGE⟩ token to fully capture, especially when such parts are only weakly connected to the main body of the target. In Fig. 9(c), the predicted mask over-segments the target, extending beyond the true object boundary to include the surrounding packaging tray; this indicates that when the target and its surrounding context share similar color and texture, the ⟨SEG⟩ token may fail to precisely localize the object boundary. These failure cases suggest that further improving robustness to visually similar distractors, thin structures, and low-contrast boundaries remains an important direction for future work.

G Discussion of Limitations and Future Work

Despite its effectiveness, SEmO has three main limitations. First, SEmO follows a single-referent design in which each forward pass processes only one target expression, similar to the original ⟨SEG⟩ token formulation in LISA (Lai et al., 2024). Consequently, queries that refer to multiple targets within a single expression cannot be segmented simultaneously, unlike approaches that assign multiple ⟨SEG⟩ tokens to handle several targets at once. Second, SEmO does not incorporate an explicit rejection mechanism for queries referring to non-existent objects. While we incorporate FP-RefCOCO(+/g) (Wu et al., 2024b) and R-RefCOCO (Wu et al., 2024a) during training to improve robustness against false-premise queries, SEmO relies on implicit learning from these datasets rather than an explicit token, such as the ⟨REJ⟩ token in GSVA (Xia et al., 2024), that directly signals the absence of the queried object. Third, as discussed in Section F, SEmO can still produce imprecise masks when distractor objects are visually near-identical to the target, when the target contains thin or loosely connected parts, or when the target and its surrounding context share similar color and texture.

Addressing these limitations is a promising direction for future work. Specifically, we consider the following directions: (1) extending SEmO to support multiple role-specific token sets per expression; (2) training the model to generate an explicit natural-language response, such as "The queried object is not present in the image," when the target is absent, while encouraging the activation of z_{others} to increase accordingly; and (3) further improving robustness to visually similar distractors and low-contrast boundaries.