REFERENTIAL RECALIBRATION FOR FINE-GRAINED REFERRING EXPRESSION SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-grained referring expression segmentation extends classic RES from segmenting an entire referred object to localizing a specific part within a host-object. Most existing methods formulate this task as a direct mapping from a holistic expression to a target mask, which works well for object-level grounding. However, when the referred target is a fine-grained part of an object, the model is prone to assigning the referential focus to the host-object rather than the target-part. Moreover, even when the target-part is partially distinguished, the resulting score map may remain incomplete or drift into irrelevant regions. To address these issues, we introduce *Referential Attribution Reformulation*, which turns the host-object from a source of interference into useful guidance for target-part localization, and *Bidirectional Collaborative Rectification*, which further refines the initial target-part score map by recovering missing regions and suppressing drift-prone activations using internal target-part support and structural affinity. Our framework is training-free and improves VLM-based segmentation without additional fine-grained annotations or parameter updates. Extensive experiments demonstrate consistent improvements across different vision-language segmentation paradigms, supporting the effectiveness and broad compatibility of the proposed inference-time framework.

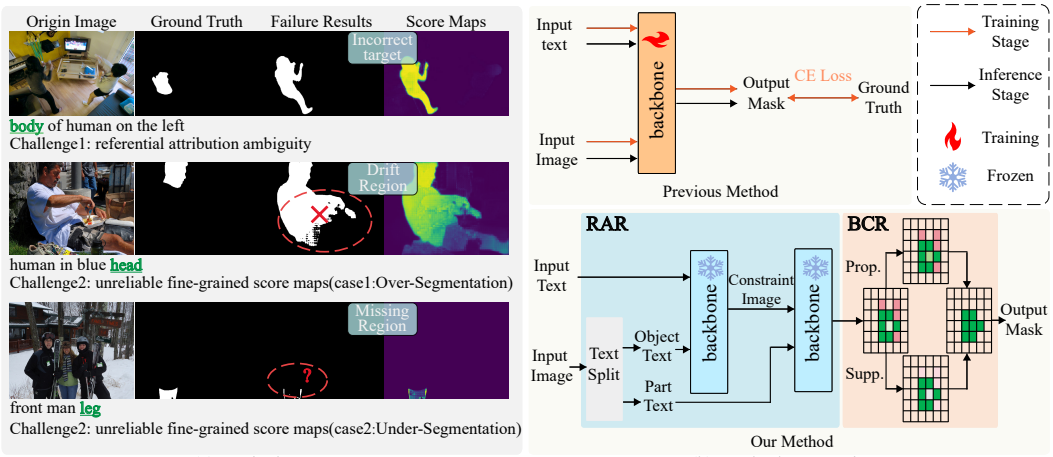
1 INTRODUCTION

Referring Expression Segmentation (RES) aims to predict pixel-level masks corresponding to objects described by natural language expressions. Benefiting from advances in vision–language pretraining models (i.e., CLIP Radford et al. (2021), ALBEF Li et al. (2021)) and Transformer-based architectures Han et al. (2022); Hatamizadeh & Kautz (2025), recent RES models have achieved substantial progress in object-level grounding Luo et al. (2024; 2020); Xia et al. (2024); Liu et al. (2023c); Wang et al. (2022); Yang et al. (2022). However, in many real-world scenarios, the referred target may correspond to only a local region or a fine-grained component of an object. This setting extends object-level grounding by requiring the model to understand fine-grained expressions and localize subtle spatial regions within the host-object.

Under the fine-grained RES setting, the standard expression-to-mask paradigm presents two major challenges. The first challenge is **referential attribution ambiguity**. As shown in the first row of Fig. 1(a), when the expression refers to the *body* of the *human* with sunglasses, the model tends to predict the entire person instead of the target-part. This ambiguity is caused by holistic text encoding, where semantically salient object words such as *human* can dominate the representation, while part words such as *body* receive insufficient attention. The second challenge is **unreliable fine-grained score maps**. As shown in the second and third rows of Fig. 1(a), even when part-related score maps are present, they are often incomplete or drift into irrelevant regions, resulting in under-segmentation or over-segmentation. The limited discriminative visual evidence provided by target-parts makes existing models prone to relying on a few salient local features or nearby host-object regions for localization. Consequently, fine-grained score maps may either cover only a salient local region or drift beyond the target-part boundary.

Despite the above challenges, most RES methods still formulate fine-grained localization as a direct mapping from the expression to the target mask Yang et al. (2022); Huang et al. (2020); Liu & Li (2025); Ding et al. (2026); Luo et al. (2025), mainly improving performance through stronger vision-language fusion or mask decoding. More recent methods further enhance local alignment by introducing large-scale fine-grained data and denser supervision Wang et al. (2024); Jang et al. (2025), as shown in the upper panel of Fig. 1(b). However, these methods still largely inherit the holistic expression-to-mask paradigm, where the host-object and target-part are not explicitly separated during inference and the final score map lacks dedicated correction. Consequently, existing methods

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.



(a) Typical Errors (b) Method Comparison

Figure 1: (a) Failure cases induced by two key challenges in fine-grained RES. The green bold text beneath each image denotes the referred target, and the score maps visualize the prediction responses in these failed results. These highlight referential attribution ambiguity and unreliable fine-grained score maps. (b) Compared to previous trained expression-to-mask methods, our framework employs RAR and BCR for explicit task decomposition and mask refinement.

remain limited in handling both referential attribution ambiguity and unreliable fine-grained score maps.

To address these unsolved limitations, we propose a method for fine-grained RES that improves the fine-grained localization ability of existing base models without requiring additional fine-grained annotations or retraining, as illustrated in the lower panel of Fig. 1(b). Our method consists of two components. *Referential Attribution Reformulation* (RAR) resolves referential attribution ambiguity by reformulating a fine-grained expression into a host-object-guided target-part localization process, thereby simplifying host-object understanding and constraining the search space for the target-part. *Bidirectional Collaborative Rectification* (BCR) further performs bidirectional correction on the resulting target-part score map by recovering missing target regions and suppressing drift into irrelevant areas. Together, these two components reduce host-object dominance over the target-part and improve the reliability of fine-grained localization.

We evaluate our method on fine-grained RES benchmarks using diverse representative models. The results show that our method consistently improves fine-grained part localization and alignment performance on pre-trained RES models as well as recent general-purpose MLLMs such as Qwen3-VL Bai et al. (2025), without any retraining. Moreover, when applied to models originally trained on classic RES datasets, our framework can achieve competitive performance with methods trained using large-scale fine-grained data. These results demonstrate that explicitly correcting referential attribution ambiguity and unreliable fine-grained score maps is an effective pathway for advancing fine-grained RES.

2 Related Work

Our work is related to three aspects of prior research, including referring expression segmentation, structured reasoning and task decomposition, and mask optimization. These three directions respectively provide the task background, the motivation for reformulating fine-grained expressions, and the basis for refining unreliable score maps.

Referring Expression Segmentation. RES was first introduced by Hu Hu et al. (2016) to segment image regions described by natural language. Existing RES methods have evolved from conventional two-branch designs with multimodal fusion and mask decoding Ding et al. (2021); Zhu et al. (2022); Dai et al. (2024; 2025); Xiao et al. (2024); Yu et al. (2024a) to recent MLLM-based segmentation frameworks Xia et al. (2024); Chen et al. (2025); Liu et al. (2023b); Lai et al. (2024). To extend RES toward finer granularity, subsequent studies further introduce fine-grained annotations Wang et al. (2024); Jang et al. (2025). However, most existing methods still follow a holistic expression-to-mask paradigm, where final mask supervision does not explicitly ensure that the encoded expression correctly captures the host–part referential relation. Even with fine-grained annotations, attribution errors may persist and lead the model to localize the wrong target. Moreover, due to the limited visual evidence of target-parts, fine-grained score maps remain prone to incompleteness and spatial drift.

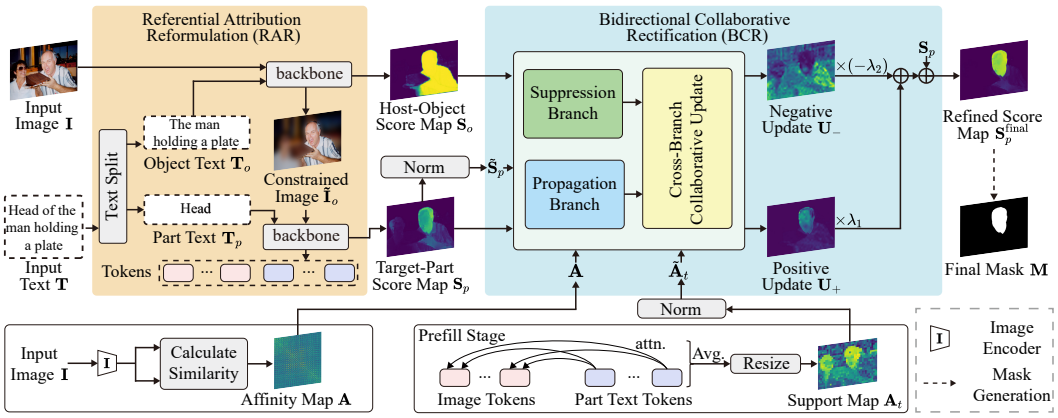


Figure 2: **Overall framework.** RAR decomposes the input expression into a host-object description $\mathbf{T}_o$ and a target-part description $\mathbf{T}_p$, and leverages the host-object prediction to obtain an initial target-part score map $\mathbf{S}_p$. BCR then collaboratively refines $\mathbf{S}_p$ through positive and negative updates, guided by a target-part support map $\mathbf{A}_t$ derived from internal prefill attention and a structural affinity matrix $\mathbf{A}$ constructed from visual patch representations, producing the final mask $\mathbf{M}$.

Structured Reasoning and Task Decomposition. Decomposing complex problems into more specific intermediate steps is a general strategy for improving robustness and reducing ambiguity Wei et al. (2022). This idea has also been explored in vision-language tasks, where semantic decomposition, staged reasoning Huang et al. (2021); Kao et al. (2025), and structured priors help replace direct holistic prediction with more manageable intermediate processes Lu et al. (2025); Bao et al. (2024); Zheng et al. (2025); Hsu et al. (2025); Li et al. (2026). While these works demonstrate the value of task simplification for complex problems, they mainly target general reasoning scenarios rather than the specific challenges of fine-grained RES, and thus do not explicitly address referential attribution ambiguity or unreliable fine-grained score maps. In contrast, our method tailors task decomposition to fine-grained RES through referential reformulation and subsequent score-map rectification.

Mask Optimization. Refining coarse masks is a common strategy for improving segmentation quality. Traditional methods improve boundary quality by explicitly exploiting the contrast between interior and exterior pixels Krähenbühl & Koltun (2011); Tang et al. (2021); Shen et al. (2021); Yuan et al. (2020). More recent works further adopt proposal selection or generic mask generators, such as SAM Kirillov et al. (2023) and Mask2Former Cheng et al. (2022) to refine an initial coarse prediction by selecting or combining candidate masks Lin et al. (2025); Keuth et al. (2024). However, these methods assume that the coarse mask already provides a reasonably reliable localization prior. This assumption is often violated in fine-grained localization, where coarse masks are more prone to host-object drift and under-activation in target regions. Consequently, such refinements may fail to produce effective corrections when the initial part mask is too weak. By contrast, our method directly rectifies fine-grained score maps through bidirectional correction under target-part support and structural affinity.

3 Overview

Given an image $\mathbf{I}$ and a referring expression $\mathbf{T}$, our goal is to segment the target-part referred to by $\mathbf{T}$ and produce its corresponding binary mask $\mathbf{M}$. Our framework consists of two components, RAR and BCR. RAR reformulates the original complex expression into a host-guided localization process, thereby alleviating referential attribution ambiguity and introducing an explicit spatial constraint for target-part localization. BCR performs collaborative bidirectional rectification on the initial target-part score map $\mathbf{S}_p$ using two complementary internal cues derived from the base model: a target-part support map $\mathbf{A}_t$ extracted from the model's prefill attention and a patch-to-patch structural affinity matrix $\mathbf{A}$ constructed from its visual patch representations.

As illustrated in Fig. 2, RAR rewrites the original expression into a host-object description $\mathbf{T}_o$ and a target-part description $\mathbf{T}_p$. This design aims to alleviate referential attribution ambiguity in fine-grained expressions, where the referential focus can be biased toward the host-object rather than the target-part. RAR first queries the base model with the host-object description $\mathbf{T}_o$ to obtain a host-object score map $\mathbf{S}_o$. This score map is then used to construct a constrained image $\tilde{\mathbf{I}}_o$, on which the target-part is localized with $\mathbf{T}_p$ to produce an initial target-part score map $\mathbf{S}_p$. In this way, holistic

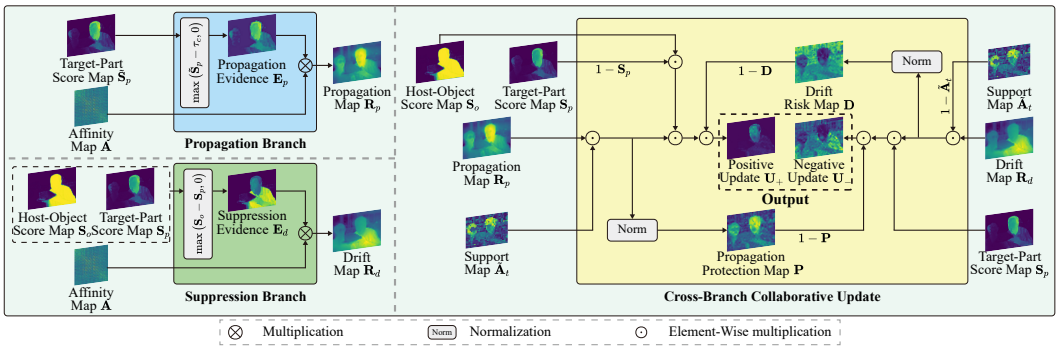


Figure 3: **Overview of the Positive and Negative Update Generation module.** The propagation and suppression branches generate complementary updates to recover missing target-part regions and suppress drift-prone responses, respectively. Cross-branch protection and risk maps coordinate the two updates for reliable target-part score refinement.

expression-to-mask prediction is reformulated as a host-guided target-part localization process, where the search space is naturally restricted to the relevant host-object.

BCR further refines the initial target-part score map $\mathbf{S}_p$. Since target-parts usually provide weaker discriminative visual evidence than complete objects, the initial score map is prone to incomplete activation and spatial drift into host-object or irrelevant regions. To improve its reliability, BCR exploits two complementary guidance signals already available within the base model. The target-part support map $\mathbf{A}_t$ is derived from the prefill attention of target-part text tokens to image tokens, providing target-part semantic support for the referred region. Meanwhile, the structural affinity matrix $\mathbf{A}$ is constructed from the visual patch representations produced by the image encoder, capturing structural correspondence among image regions. Guided by these internal semantic and structural cues, BCR performs bidirectional rectification, where one branch recovers missing target-part regions and the other suppresses activations drifting toward the host-object or irrelevant regions. Their complementary effects jointly produce the final refined score map $\mathbf{S}_p^{\text{final}}$, which is then converted into the final mask $\mathbf{M}$ following the original mask generation procedure of the base method.

4 METHOD

4.1 REFERENTIAL ATTRIBUTION REFORMULATION

Fine-grained referring expressions often involve both a host-object and a target-part, making it difficult for the model to correctly assign the referential focus. To alleviate this issue, we reformulate the original expression $\mathbf{T}$ into two more specific descriptions

$$\mathbf{T} \rightarrow (\mathbf{T}_o, \mathbf{T}_p), \tag{1}$$

where $\mathbf{T}_o$ denotes the host-object description and $\mathbf{T}_p$ denotes the target-part description.

We first use the host-object description $\mathbf{T}_o$ to obtain the host-object score map $\mathbf{S}_o$ from the backbone. Compared with the original mixed expression $\mathbf{T}$, $\mathbf{T}_o$ provides a clearer object-level localization target and thus yields a more reliable host-object constraint. We then construct a constrained image $\tilde{\mathbf{I}}_o$ from the original image $\mathbf{I}$ and the host-object score map $\mathbf{S}_o$. Specifically, regions outside the host-object score map are blurred, reducing interference from irrelevant instances and background regions. The target-part description $\mathbf{T}_p$ is then paired with $\tilde{\mathbf{I}}_o$ to produce the initial target-part score map $\mathbf{S}_p$.

Overall, RAR reformulates fine-grained localization from holistic expression-to-mask prediction into host-guided target-part localization. The reformulated descriptions clarify the referential relation between the host-object and the target-part, while the host-object score map provides an explicit spatial constraint for target-part localization. This encourages the model to focus on the target-part rather than the host-object, producing an initial target-part score map $\mathbf{S}_p$ for subsequent rectification.

4.2 BIDIRECTIONAL COLLABORATIVE RECTIFICATION

The first stage produces an initial target-part score map $\mathbf{S}_p$ under the guidance of the host-object. However, this score map may still contain two unresolved issues. Some target-part regions may remain weakly activated due to limited fine-grained visual evidence, while some activations may drift toward the host-object or irrelevant regions. To address these remaining errors, we perform BCR, which optimizes $\mathbf{S}_p$ from two complementary directions. The propagation branch recovers missing

target-part regions, while the suppression branch removes drift-prone activations. These two branches are coordinated by protection and risk maps constructed from target-part support and structural cues.

4.2.1 RECTIFICATION INPUTS

BCR operates on the initial target-part score map $\mathbf{S}_p$, the host-object score map $\mathbf{S}_o$, the target-part support map $\mathbf{A}_t$, and the structural affinity matrix $\mathbf{A}$.

The target-part support map $\mathbf{A}_t$ is derived from the model's internal attention during the prefill stage. Specifically, let $\mathcal{P}$ denote the set of token indices corresponding to the target-part description $\mathbf{T}_p$, and let $\mathbf{W}^{l,h}$ denote the attention matrix at the h-th attention head of layer l, where $\mathbf{W}^{l,h}_{k,i}$ represents the attention weight from the k-th text token to the i-th image token. We obtain the target-part support for each image token by averaging the corresponding attention responses,

$$\mathbf{A}_t(i) = \frac{1}{|\mathcal{L}||\mathcal{H}||\mathcal{P}|} \sum_{l\in\mathcal{L}} \sum_{h\in\mathcal{H}} \sum_{k\in\mathcal{P}} \mathbf{W}^{l,h}_{k,i}, \quad i = 1, \ldots, N, \tag{2}$$

where $\mathcal{L}$ and $\mathcal{H}$ denote the selected layers and attention heads, respectively. The resulting responses are reshaped according to the spatial layout of the image tokens to form $\mathbf{A}_t$.

To obtain the structural affinity matrix $\mathbf{A}$, we directly reuse the visual patch representations produced by the image encoder. Let $\mathbf{F} = \{\mathbf{f}_i\}_{i=1}^N$ denote the visual patch representations. The affinity between two image patches is computed as

$$\mathbf{A}(i,j) = \frac{\mathbf{f}_i^\top \mathbf{f}_j}{\|\mathbf{f}_i\|_2 \|\mathbf{f}_j\|_2}, \quad i,j = 1, \ldots, N. \tag{3}$$

The resulting matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$ captures patch-level structural correspondence and supports score propagation across visually coherent regions.

4.2.2 COLLABORATIVE POSITIVE AND NEGATIVE UPDATE GENERATION

BCR generates two complementary updates to rectify the initial target-part score map. The positive update aims to recover missing target-part regions through propagation from confident target-part responses, while the negative update aims to suppress drift-prone activations where the host-object response dominates the target-part response.

Propagation Branch. The propagation branch aims to recover missing target-part regions by aggregating propagation evidence over structurally coherent regions. Given the confidence threshold $\tau_c \in [0, 1]$, we define the propagation evidence as $\mathbf{E}_p(j) = \max\big(\tilde{\mathbf{S}}_p(j) - \tau_c, 0\big)$, which retains only confident target-part responses. For each location $i \in \Omega$, where Ω denotes the spatial location set of the score map, the propagation map is computed as

$$\mathbf{R}_p(i) = \sum_{j\in\Omega} \mathbf{A}(i,j)\mathbf{E}_p(j). \tag{4}$$

$\mathbf{R}_p(i)$ measures the aggregated propagation evidence at location i, where a high value indicates strong structural support from confident target-part regions. We use the normalized score map $\tilde{\mathbf{S}}_p$ to enhance the relative contrast among responses, making confident target-part regions easier to identify.

Suppression Branch. The suppression branch aims to identify drift-prone activations by aggregating suppression evidence over structurally coherent regions. We define the suppression evidence as $\mathbf{E}_d(j) = \max\big(\mathbf{S}_o(j) - \mathbf{S}_p(j), 0\big)$, which captures host-object-dominated responses where the host-object score exceeds the target-part score. For each location $i \in \Omega$, the drift map is computed as

$$\mathbf{R}_d(i) = \sum_{j\in\Omega} \mathbf{A}(i,j)\mathbf{E}_d(j). \tag{5}$$

$\mathbf{R}_d(i)$ measures the aggregated suppression evidence at location i, where a high value indicates strong structural support from drift-prone regions.

Cross-Branch Collaborative Update. Although the propagation and suppression branches provide complementary evidence, applying them independently may introduce conflicting updates. The propagation branch may over-expand into host-object regions, whereas the suppression branch may

Table 1: Comparison with existing methods on RefCOCOm. Results are reported in terms of mIoU.

Type	Methods	val		testA		testB	
		Part	Obj. & Part	Part	Obj. & Part	Part	Obj. & Part
Classic Methods	SeqTR Zhu et al. (2022)	13.9	28.2	12.1	22.8	18.1	34.7
	CRIS Wang et al. (2022)	10.6	25.4	10.1	21.2	12.9	30.0
	LAVT Yang et al. (2022)	15.3	29.9	13.2	24.4	18.7	35.5
	X-Decoder Zou et al. (2023)	16.2	29.5	13.6	23.6	20.3	33.8
	PolyFormer Liu et al. (2023c)	18.8	32.9	15.7	26.5	25.1	40.5
	LISA-7B Lai et al. (2024)	21.3	34.3	18.5	28.6	25.7	40.1
	LISA-7B+Ours	**24.4**	**37.2**	**21.5**	**31.8**	**27.9**	**41.6**
Fine-Grained Methods	UniRES Wang et al. (2024)	19.6	34.3	16.4	27.8	25.2	41.7
	M^{2}SA-Llama2-13B Jang et al. (2025)	24.5	37.3	21.9	31.9	28.5	42.7
	M^{2}SA-Llama2-13B+Ours	**27.4**	**39.2**	**24.4**	**35.2**	**30.8**	**44.6**

remove target-part regions with weak initial activations. Therefore, cross-branch collaboration is required so that propagation can be inhibited in drift-prone regions and suppression can be weakened in reliable target-part regions. For this purpose, we construct two cross-branch maps as

$$\mathbf{P} = \mathrm{Norm}\Big(\mathbf{R}_p \odot \tilde{\mathbf{A}}_t\Big), \quad \mathbf{D} = \mathrm{Norm}\Big(\mathbf{R}_d \odot (1 - \tilde{\mathbf{A}}_t)\Big), \tag{6}$$

where $\odot$ denotes element-wise multiplication. The propagation protection map $\mathbf{P}$ highlights locations that should be protected from over-suppression, while the drift risk map $\mathbf{D}$ highlights locations where propagation should be inhibited due to high drift risk.

With these two maps, the propagation and suppression updates are defined as

$$\begin{aligned} \mathbf{U}_+(i) &= \mathbf{S}_o(i)\big(1 - \mathbf{S}_p(i)\big)\mathbf{R}_p(i)\tilde{\mathbf{A}}_t(i)\big(1 - \mathbf{D}(i)\big), \\ \mathbf{U}_-(i) &= \mathbf{S}_p(i)\mathbf{R}_d(i)\big(1 - \tilde{\mathbf{A}}_t(i)\big)\big(1 - \mathbf{P}(i)\big). \end{aligned} \tag{7}$$

The factor $\mathbf{S}_o(i)$ imposes a host-object constraint on $\mathbf{U}_+$, preventing target-part propagation from leaking to similar parts of other host-objects when $\tilde{\mathbf{A}}_t$ produces ambiguous semantic responses. Thus, $\mathbf{U}_+$ is activated in host-object regions that are under-estimated by $\mathbf{S}_p$, supported by structural propagation $\mathbf{R}_p$, consistent with target-part support $\tilde{\mathbf{A}}_t$, and not assigned high drift risk by $\mathbf{D}$. In contrast, $\mathbf{U}_-$ is activated on existing target-part responses that receive drift support from $\mathbf{R}_d$, lack target-part support, and are not protected by $\mathbf{P}$.

4.2.3 Final Score Map Refinement

To integrate the complementary effects of propagation and suppression, we update $\mathbf{S}_p$ by adding the weighted positive correction and subtracting the weighted negative correction,

$$\mathbf{S}_p^{\mathrm{final}} = \mathbf{S}_p + \lambda_1 \mathbf{U}_+ - \lambda_2 \mathbf{U}_-, \tag{8}$$

where λ_1 and λ_2 control the strengths of positive propagation and negative suppression, respectively. The refined score map is then converted into the final mask $\mathbf{M}$ following the original mask generation procedure of each base method. This weighted additive refinement integrates propagation and suppression within a unified score map update rather than applying them as separate post-processing steps. The positive update $\mathbf{U}_+$ enhances missing target-part evidence, while the negative update $\mathbf{U}_-$ removes drift-prone activations. With the cross-branch constraints provided by $\mathbf{P}$ and $\mathbf{D}$, the update reduces both over-expansion and over-suppression, leading to a more reliable target-part score map.

5 Experiments

5.1 Experimental Setup

Datasets. Following the fine-grained RES task setting Wang et al. (2024), we evaluate our method on RefCOCOm Wang et al. (2024), a referring expression segmentation benchmark that contains fine-grained expressions. Following the official protocol, we report results on the val, testA, and testB splits. For each split, we separately evaluate **Part** expressions that focus on local target-parts and **Obj. & Part** expressions that involve both the host-object and the target-part. We further evaluate on standard object-level RES benchmarks to examine whether the proposed inference-time framework remains compatible with conventional object-level referring expressions.

Table 2: Comparison with existing methods on RefCOCO, RefCOCO+, and RefCOCOg using cIoU. "–" indicates that no additional training is required when applying our framework to the base method.

Methods	Training	RefCOCO			RefCOCO+			RefCOCOg	
		val	testA	testB	val	testA	testB	val(U)	test(U)
LAVT Yang et al. (2022)	✓	72.7	75.8	68.8	62.1	68.4	55.1	61.2	62.1
ReLA Liu et al. (2023a)	✓	73.8	76.5	70.2	66.0	71.0	57.7	65.0	66.0
LISA-7B Lai et al. (2024)	✓	74.1	76.5	71.1	62.4	67.4	56.5	66.4	68.5
LISA-Llama2-13B Lai et al. (2024)	✓	73.6	77.3	70.5	63.2	68.2	57.0	67.0	68.4
M^2SA-Llama2-13B Jang et al. (2025)	✓	74.6	77.6	71.0	64.0	68.1	57.6	69.0	69.3
M^2SA-Llama2-13B+Ours	–	**75.4**	**78.1**	**72.3**	**68.0**	**70.5**	**59.6**	**69.8**	**69.5**
Pseudo-RIS Yu et al. (2024b)	×	41.1	48.2	33.5	44.3	51.4	35.1	46.0	46.7
SAM3 Agent (Qwen3-VL-8B) Carion et al. (2025)	×	68.6	72.3	63.9	59.3	64.5	55.6	66.4	66.9
Kang et al. (LLaVA-1.5-13B) Kang et al. (2025)	×	76.1	78.9	72.8	64.1	67.1	57.3	67.7	69.0
Qwen3-VL-8B-Instruct Bai et al. (2025)	×	77.2	81.1	73.3	73.3	77.9	66.5	73.5	74.7
Qwen3-VL-8B-Instruct+Ours	–	**78.7**	**82.8**	**75.4**	**74.5**	**78.6**	**68.3**	**76.1**	**75.3**

Evaluation metric. Following the implementation of the fine-grained referring expression segmentation works Wang et al. (2024); Jang et al. (2025), we adopt mIoU scores to assess the quality of the segmentation mask. This metric directly measures the overlap between the predicted mask and the ground-truth target mask, and is suitable for assessing fine-grained part localization quality.

Compared methods. To evaluate the effectiveness of our method across different types of RES models, we organize the compared methods into two groups. The first group, denoted as **Classic Methods**, includes conventional RES models trained on standard object-level referring expression segmentation datasets, including RefCOCO Kazemzadeh et al. (2014), RefCOCO+ Kazemzadeh et al. (2014), and RefCOCOg Nagaraja et al. (2016). These methods are directly evaluated on fine-grained expressions without explicit fine-grained supervision, which allows us to examine whether our method can activate their latent fine-grained localization ability at inference time. The second group, denoted as **Fine-Grained Methods**, includes methods specifically designed or adapted for multi-grained or fine-grained segmentation Wang et al. (2024); Jang et al. (2025).

Implementation details. All experiments are conducted in a training-free inference setting, where the parameters of the base models are kept fixed. For RAR, we use Llama-3-8B-Instruct Grattafiori et al. (2024) as a representative open-source instruction-following language model to determine whether a referring expression describes a target-part. For expressions identified as fine-grained, it further decomposes the original expression into a host-object description and a target-part description. For BCR, we derive both target-part support and structural affinity directly from intermediate representations of the base model, without introducing an additional vision-language model. Specifically, the target-part support map $\mathbf{A}_t$ is obtained by aggregating the prefill attention from target-part text tokens to image tokens over layers 1–8, which exhibit relatively high image attention ratios; detailed layer-wise analysis is provided in the appendix. The structural affinity matrix $\mathbf{A}$ is constructed from the visual patch representations produced by the base model's image encoder. The confidence threshold and rectification weights are set to $\tau_c = 0.4$, $\lambda_1 = 0.7$, and $\lambda_2 = 0.5$, respectively, in all experiments. All experiments are conducted on a single NVIDIA A6000 GPU.

5.2 Comparison on RES Benchmarks

Table 1 reports the main comparison on RefCOCOm. The upper block contains classic RES methods mainly trained on standard object-level RES datasets, while the lower block contains methods designed for multi-grained or fine-grained segmentation. This comparison evaluates whether a method can handle fine-grained expressions that require not only locating the correct host-object but also focusing the prediction on the referred target-part.

Classic RES methods show limited performance on RefCOCOm, especially on the more challenging **Part** subset. We highlight the val-Part setting for a clearer comparison, where the model is required to precisely localize the referred target-part. As shown in Table 1, our method consistently improves both classic RES models and fine-grained methods. These results demonstrate that our inference-time framework consistently improves fine-grained localization across both conventional base models and methods specifically designed for fine-grained segmentation.

As shown in Table 2, we further evaluate our method on standard RES benchmarks, including RefCOCO Kazemzadeh et al. (2014), RefCOCO+ Kazemzadeh et al. (2014), and RefCOCOg Nagaraja et al. (2016). Although our framework is primarily designed for fine-grained localization, it also improves object-level segmentation performance. These results show that the proposed inference-time

Table 3: Framework evaluation across different base models. Segmentation results are reported in terms of mIoU. Inference efficiency is measured by GPU memory usage and inference time per item.

Methods	Segmentation Results						Inference Efficiency	
	val		testA		testB		GPU (GB)	Time (s/item)
	Part	Obj. & Part	Part	Obj. & Part	Part	Obj. & Part		
LISA-7B Lai et al. (2024)	21.3	34.3	18.5	28.6	25.7	40.1	17.8	0.5 ± 0.3
LISA-7B+Ours	**24.4**	**37.2**	**21.5**	**31.8**	**27.9**	**41.6**	28.4	1.1 ± 0.3
LISA-13B Lai et al. (2024)	22.1	35.2	19.4	29.7	27.2	41.6	29.2	0.6 ± 0.3
LISA-13B+Ours	**25.3**	**37.9**	**23.4**	**33.2**	**29.3**	**44.0**	31.0	1.2 ± 0.3
M^2SA-Llama2-13B Jang et al. (2025)	24.5	37.3	21.9	31.9	28.5	42.7	30.1	0.6 ± 0.3
M^2SA-Llama2-13B+Ours	**27.4**	**39.2**	**24.4**	**35.2**	**30.8**	**44.6**	31.2	1.6 ± 0.3
Qwen3-VL-8B-Instruct Bai et al. (2025)	33.4	44.6	30.1	39.2	37.4	49.3	23.4	1.2 ± 0.3
Qwen3-VL-8B-Instruct+Ours	**39.8**	**49.3**	**37.2**	**45.2**	**38.4**	**49.5**	25.3	2.5 ± 0.3

framework remains compatible with conventional object-level RES and does not trade object-level performance for improved fine-grained localization.

5.3 Framework Gains across Segmentation Paradigms

We evaluate our method on representative vision-language models spanning different segmentation paradigms, including General VLM-based segmentation models, Fine-grained VLM-based segmentation methods, and pipelines that combine general-purpose MLLMs with segmentation models. For the latter, we use Qwen3-VL Bai et al. (2025) to predict the target bounding box from the referring expression and then use the predicted box as the prompt for SAM to obtain the segmentation mask.

As shown in Table 3, our method consistently improves different VLM-based segmentation models and MLLM-based segmentation pipelines without additional training. These results demonstrate the compatibility of our inference-time framework across different vision-language segmentation paradigms, without relying on a specific model architecture or segmentation pipeline.

The additional inference time mainly comes from the sequential host-object and target-part localization in RAR, while BCR introduces additional memory overhead for constructing the target-part support map and structural affinity matrix. Despite the additional inference overhead, the entire framework requires neither retraining nor additional fine-grained annotations.

5.4 Ablation Study

Component-level Ablation. We examine the contributions of RAR and BCR in Table 4. Starting from the original LISA-7B baseline (see "LISA"), RAR (see "LISA + RAR") reformulates the expression into a host-object description and a target-part description. Compared with directly using the original expression, this reformulation provides a clearer referential structure and spatially constrains target-part localization within the relevant host-object. BCR (see "LISA + BCR") retains the original expression formulation but rectifies the initial target-part score map using internal target-part support and structural affinity, helping recover incomplete regions and suppress drift-prone activations. Combining RAR and BCR (see "LISA + Full Model") achieves the best performance, showing that the two components provide complementary benefits.

Table 4: Component-level ablation on the val-Part subset of RefCOCOm.

Setting	RAR	BCR	mIoU
LISA	×	×	21.3
LISA + RAR	✓	×	22.7
LISA + BCR	×	✓	23.1
LISA + Full Model	✓	✓	**24.4**

Rectification Ablation. We analyze the design of the rectification module in Table 5. Starting from LISA + RAR (see "RAR only"), the model benefits from referential reformulation. Adding propagation and suppression without collaboration (see "RAR + Prop. & Supp.") brings limited improvement, as the two branches operate independently and cannot coordinate where propagation should be restrained and where suppression should be weakened. In contrast, introducing cross-branch collaboration (see "RAR + Collaborative") further improves performance. Specifically, the drift risk map restrains propagation in drift-prone regions, while the propagation protection map weakens suppression in regions with reliable target-part support. This collaboration reduces conflicting updates between branches and produces a more reliable target-part score map.

Table 5: Rectification branch ablation on the val-Part subset of RefCOCOm.

Setting	Prop. & Supp.	Collab.	mIoU
RAR only	×	×	22.7
RAR + Prop. & Supp.	✓	×	23.0
RAR + Collaborative	✓	✓	**24.4**

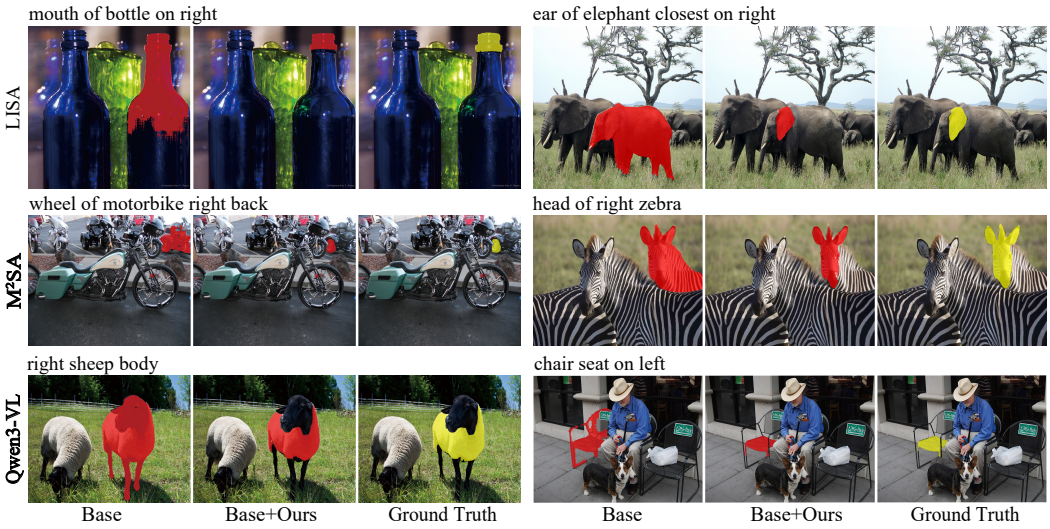


Figure 4: Qualitative comparisons across different vision-language segmentation paradigms. From left to right: image, ground truth, baseline, and baseline enhanced by our method.

Guidance Ablation. We further study the roles of target-part support and structural affinity in Table 6. Compared with RAR only, using structural affinity alone (see "RAR + Stru.") enables score propagation across structurally related patches. However, structural affinity by itself does not indicate which regions are relevant to the referred target-part and may therefore propagate responses toward structurally similar but irrelevant regions. Using target-part support alone (see "RAR + Support.") provides localization evidence derived from the internal attention between target-part text tokens and image tokens, but lacks explicit structural correspondence between image regions. When both guidance signals are combined (see "RAR + Both"), the support map guides rectification toward regions relevant to the referred part, while structural affinity captures correspondence among related image patches.

Table 6: Guidance ablation on the val-Part subset of RefCOCOm.

Setting	Supp.	Stru.	mIoU
RAR only	×	×	22.7
RAR + Stru.	×	✓	22.9
RAR + Support	✓	×	23.2
RAR + Both	✓	✓	**24.4**

Refinement Ablation. We further compare BCR with SAMRefiner Lin et al. (2025), a recent representative mask refinement method, in Table 7. Applying SAMRefiner directly to LISA (see "LISA + SAMRefiner") brings only limited improvement. Combining RAR with SAMRefiner (see "RAR + SAMRefiner") further improves the performance, showing that generic mask refinement can benefit from the improved target-part localization provided by RAR. Under the same RAR setting, our BCR (see "Ours") achieves better performance than SAMRefiner. These results demonstrate the effectiveness of BCR for refining fine-grained target-part predictions.

Table 7: Comparison with mask refinement on the val-Part subset of RefCOCOm.

Setting	RAR	BCR	mIoU
LISA	×	×	21.3
LISA + SAMRefiner	×	×	21.6
RAR + SAMRefiner	✓	×	23.1
Ours	✓	✓	**24.4**

5.5 Qualitative Results

Fig. 4 provides qualitative comparisons with the base models. Our method more accurately localizes referred target-parts, especially when the baseline segments the whole host-object or only a small discriminative region. These results further demonstrate its effectiveness in alleviating referential attribution ambiguity and improving fine-grained predictions.

6 Conclusion

In this paper, we revisit fine-grained referring expression segmentation from the perspective of referential attribution and prediction reliability. We propose a training-free framework that reformulates complex expressions into host-guided target-part localization and rectifies the resulting score maps through bidirectional propagation and suppression using internal target-part support and structural affinity. Experiments across different vision-language segmentation paradigms demonstrate its effectiveness and compatibility without additional fine-grained annotations or retraining.

AI USE STATEMENT

We used generative AI tools to assist with feedback on research methodology and experimental design, code editing, and manuscript writing and revision, including language polishing, translation, and LaTeX formatting. Generative AI tools were also used to assist with the preparation and refinement of scientific figures. All AI-assisted content was reviewed and verified by the authors, and all implementations and experimental results were independently checked for correctness. The authors take full responsibility for the final content of this work, including all methods, results, claims, and AI-assisted materials.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

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.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *International Conference on Learning Representations*, 2025.

Jierun Chen, Fangyun Wei, Jinjing Zhao, Sizhe Song, Bohuai Wu, Zhuoxuan Peng, S-H Gary Chan, and Hongyang Zhang. Revisiting referring expression comprehension evaluation in the era of large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 513–524, 2025.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1290–1299, 2022.

Ming Dai, Lingfeng Yang, Yihao Xu, Zhenhua Feng, and Wankou Yang. Simvg: A simple framework for visual grounding with decoupled multi-modal fusion. *Advances in Neural Information Processing Systems*, 37:121670–121698, 2024.

Ming Dai, Jian Li, Jiedong Zhuang, Xian Zhang, and Wankou Yang. Multi-task visual grounding with coarse-to-fine consistency constraints. In *AAAI Conference on Artificial Intelligence*, volume 39, pp. 2618–2626, 2025.

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

Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Yu-Gang Jiang. Grex: Generalized referring expression segmentation, comprehension, and generation. *International Journal of Computer Vision*, 134(2):79, 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):87–110, 2022.

Ali Hatamizadeh and Jan Kautz. Mambavision: A hybrid mamba-transformer vision backbone. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 25261–25270, 2025.

Joy Hsu, Jiayuan Mao, Joshua B Tenenbaum, Noah D Goodman, and Jiajun Wu. What makes a maze look like a maze? International Conference on Learning Representations, 2025.

Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *European Conference on Computer Vision*, pp. 108–124. Springer, 2016.

Jingwei Huang, Yanfeng Zhang, and Mingwei Sun. Primitivenet: Primitive instance segmentation with local primitive embedding under adversarial metric. In *IEEE/CVF International Conference on Computer Vision*, pp. 15343–15353, 2021.

Shaofei Huang, Tianrui Hui, Si Liu, Guanbin Li, Yunchao Wei, Jizhong Han, Luoqi Liu, and Bo Li. Referring image segmentation via cross-modal progressive comprehension. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10488–10497, 2020.

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. *arXiv preprint arXiv:2503.13881*, 2025.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. Your large vision-language model only needs a few attention heads for visual grounding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9339–9350, 2025.

Shiu-hong Kao, Yu-Wing Tai, and Chi-Keung Tang. Think before you segment: High-quality reasoning segmentation with gpt chain of thoughts. *arXiv preprint arXiv:2503.07503*, 2025.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Conference on Empirical Methods in Natural Language Processing*, pp. 787–798, 2014.

Ron Keuth, Lasse Hansen, Maren Balks, Ronja Jäger, Anne-Nele Schröder, Ludger Tüshaus, and Mattias Heinrich. Sam carries the burden: A semi-supervised approach refining pseudo labels for medical segmentation. *arXiv preprint arXiv:2411.12602*, 2024.

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 *IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in Neural Information Processing Systems*, 24, 2011.

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

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34:9694–9705, 2021.

Xinlong Li, Di Lin, Shaoyiyi Gao, Jiaxin Li, Ruonan Liu, and Qing Guo. Open-vocabulary part segmentation via progressive and boundary-aware strategy. *Advances in Neural Information Processing Systems*, 38:63609–63634, 2026.

Yuqi Lin, Hengjia Li, Wenqi Shao, Zheng Yang, Jun Zhao, Xiaofei He, Ping Luo, and Kaipeng Zhang. Samrefiner: Taming segment anything model for universal mask refinement. *arXiv preprint arXiv:2502.06756*, 2025.

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.

Jiang Liu, Hui Ding, Zhaowei Cai, Yuting Zhang, Ravi Kumar Satzoda, Vijay Mahadevan, and R Manmatha. Polyformer: Referring image segmentation as sequential polygon generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18653–18663, 2023c.

Ting Liu and Siyuan Li. Hybrid global-local representation with augmented spatial guidance for zero-shot referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29634–29643, 2025.

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 *Annual Meeting of the Association for Computational Linguistics*, pp. 14699–14716, 2025.

Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10034–10043, 2020.

Zhuoyan Luo, Yinghao Wu, Yong Liu, Yicheng Xiao, Xiao-Ping Zhang, and Yujiu Yang. Hdc: Hierarchical semantic decoding with counting assistance for generalized referring expression segmentation. *arXiv preprint arXiv:2405.15658*, 1(5):6, 2024.

Zhuoyan Luo, Yinghao Wu, Tianheng Cheng, Yong Liu, Yicheng Xiao, Hongfa Wang, Xiao-Ping Zhang, and Yujiu Yang. Cohd: A counting-aware hierarchical decoding framework for generalized referring expression segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22685–22694, 2025.

Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *European Conference on Computer Vision*, pp. 792–807. Springer, 2016.

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.

Xing Shen, Jirui Yang, Chunbo Wei, Bing Deng, Jianqiang Huang, Xian-Sheng Hua, Xiaoliang Cheng, and Kewei Liang. Dct-mask: Discrete cosine transform mask representation for instance segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8720–8729, 2021.

Chufeng Tang, Hang Chen, Xiao Li, Jianmin Li, Zhaoxiang Zhang, and Xiaolin Hu. Look closer to segment better: Boundary patch refinement for instance segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13926–13935, 2021.

Wenxuan Wang, Tongtian Yue, Yisi Zhang, Longteng Guo, Xingjian He, Xinlong Wang, and Jing Liu. Unveiling parts beyond objects: Towards finer-granularity referring expression segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12998–13008, 2024.

Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11686–11695, 2022.

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

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

Linhui Xiao, Xiaoshan Yang, Fang Peng, Yaowei Wang, and Changsheng Xu. Oneref: Unified one-tower expression grounding and segmentation with mask referring modeling. *Advances in Neural Information Processing Systems*, 37:139854–139885, 2024.

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

Seonghoon Yu, Ilchae Jung, Byeongju Han, Taeoh Kim, Yunho Kim, Dongyoon Wee, and Jeany Son. A simple baseline with single-encoder for referring image segmentation. *arXiv preprint arXiv:2408.15521*, 2024a.

Seonghoon Yu, Paul Hongsuck Seo, and Jeany Son. Pseudo-ris: Distinctive pseudo-supervision generation for referring image segmentation. In *European Conference on Computer Vision*, pp. 18–36. Springer, 2024b.

Yuhui Yuan, Jingyi Xie, Xilin Chen, and Jingdong Wang. Segfix: Model-agnostic boundary refinement for segmentation. In *European Conference on Computer Vision*, pp. 489–506. Springer, 2020.

Shiyi Zheng, Peizhi Zhao, Zhilong Zheng, Peihang He, Haonan Cheng, Yi Cai, and Qingbao Huang. Look qround before locating: Considering content and structure information for visual grounding. In *AAAI Conference on Artificial Intelligence*, volume 39, pp. 1656–1664, 2025.

Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. Seqtr: A simple yet universal network for visual grounding. In *European Conference on Computer Vision*, pp. 598–615. Springer, 2022.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15116–15127, 2023.

A Fine-Grained Failure Categorization

To further clarify the categorization criteria, we note that RefCOCOm does not provide object-level masks for all cases. We therefore define five categories using the host-object score map $\mathbf{S}_o$ as a reference. A prediction is labeled as `CORRECT` when its IoU with the ground-truth target-part mask is at least 0.85. Predictions with no overlap with the ground-truth target-part mask (IoU $= 0$) are labeled as `TARGET_ERROR`.

A prediction is labeled as `AMBIGUITY` when it primarily covers the host-object rather than the target-part. Specifically, we compare the mean responses of $\mathbf{S}_o$ and the initial target-part score map outside the predicted mask. If their difference is below 0.05, the IoU between the thresholded host-object score map and the predicted mask exceeds 0.8, and at least 85% of the ground-truth target-part region is covered by the predicted mask, the prediction is categorized as `AMBIGUITY`.

A prediction is labeled as `OVER_SEG` when at least 85% of the ground-truth target-part region is covered by the predicted mask, but either the IoU between the predicted and ground-truth masks is below 0.8 or the mean score difference outside the predicted mask exceeds 0.05. Such predictions largely cover the target-part but expand into adjacent host-object or background regions.

Finally, predictions that partially overlap with the ground-truth target-part region but cover less than 85% of it (IoU > 0) are labeled as `UNDER_SEG`. The categories are assigned sequentially according to the above criteria, ensuring that each prediction is assigned to exactly one category.

As shown in Table 8, 23.79% of the predictions are categorized as `CORRECT`, indicating that accurate target-part localization remains challenging. The dominant failure type is `AMBIGUITY`, accounting for 51.71% of all cases. This suggests that the model frequently localizes the host-object rather than assigning the referential focus to the target-part, providing empirical evidence for the referential attribution ambiguity discussed above.

Table 8: Fine-grained failure categorization on RefCOCOm.

Category	Count	Percentage
`CORRECT`	4,399	23.79
`AMBIGUITY`	9,563	51.71
`UNDER_SEG`	619	3.35
`OVER_SEG`	2,836	15.34
`TARGET_ERROR`	1,075	5.81
Total	18,492	100.00

Beyond attribution ambiguity, fine-grained score-map reliability remains another limitation. `OVER_SEG` accounts for 15.34% of the predictions, indicating that predictions often expand beyond the target-part into adjacent regions. `UNDER_SEG` accounts for 3.35%, reflecting incomplete target-part localization, while `TARGET_ERROR` accounts for 5.81%, corresponding to complete localization failures. Overall, these results support the two challenges motivating our method: referential attribution ambiguity and unreliable fine-grained target-part predictions.

B Evidence of Referential Attribution Bias in Existing Models

Before performing the failure categorization, we investigate how existing models allocate attention between the host-object and target-part in fine-grained expressions. Specifically, using M^{2}SA as a

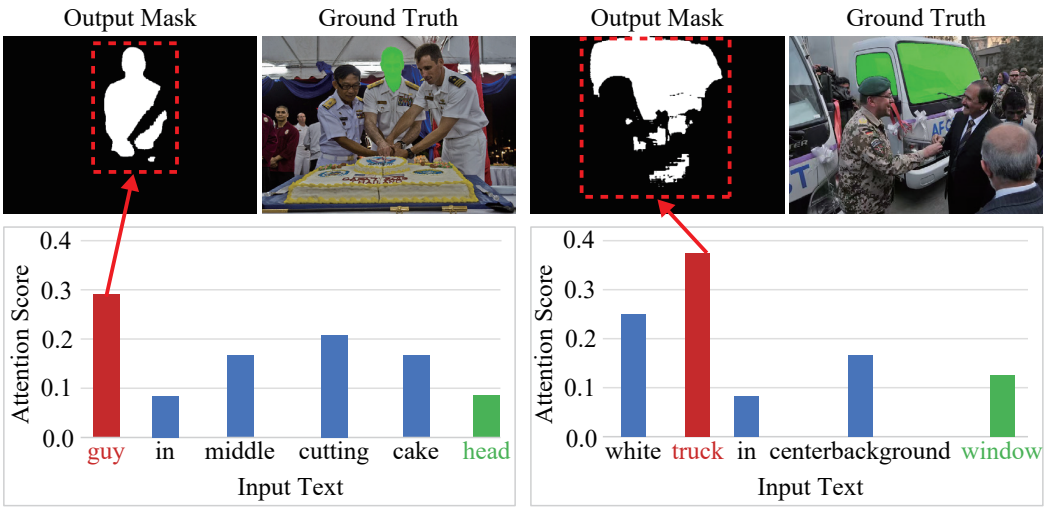


Figure 5: Attention analysis of fine-grained referring expressions. Green denotes the target-part word, while red denotes the host-object word receiving higher attention.

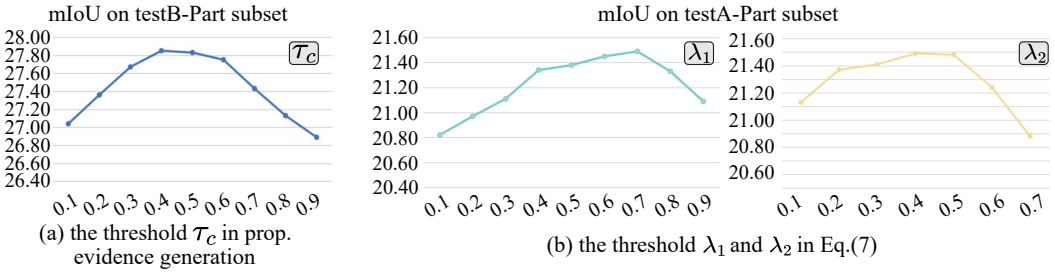
representative fine-grained method, we extract the attention weights from the <SEG> token to the tokenized referring expression during inference and average them across multiple attention heads.

As shown in Fig. 5, the <SEG> token can assign substantially higher attention to the host-object word than to the target-part word, even in a model trained with fine-grained supervision. Such cases reflect a referential attribution bias toward the host-object and are accompanied by object-level predictions rather than the referred target-part masks.

These observations motivate the explicit separation of host-object and target-part descriptions in our RAR module. By reformulating the original expression into separate referential signals, RAR reduces interference from the host-object and provides a clearer target for fine-grained localization.

C Additional Experiments

C.1 Parameter Sensitivity Analysis



(a) the threshold τ_c in prop. evidence generation

(b) the threshold λ_1 and λ_2 in Eq.(7)

Figure 6: **Parameter sensitivity analysis.** We evaluate the effects of the propagation threshold τ_c and the rectification weights λ_1, λ_2 on different subsets of RefCOCOm. The results show that our method is relatively stable across different parameter values.

Effect of τ_c. We study the confidence threshold τ_c used to construct the propagation evidence $\mathbf{E}_p$. As shown in Fig. 6(a), performance remains relatively stable across different threshold values, indicating that our method is not highly sensitive to τ_c. Increasing the threshold initially improves performance by filtering low-confidence responses from propagation evidence. However, an overly high threshold removes useful evidence from weakly activated target-part regions, leading to performance degradation. We therefore set $\tau_c = 0.4$ by default.

Effect of λ_1 and λ_2. We further analyze the rectification weights λ_1 and λ_2 on the testA split of RefCOCOm. As shown in Fig. 6(b), we first fix $\lambda_1 = 0.7$ and vary λ_2 to examine the effect of negative suppression. Performance remains relatively stable, with favorable results around $\lambda_2 = 0.5$. We then fix $\lambda_2 = 0.5$ and vary λ_1 to evaluate the effect of positive propagation. Performance remains stable over a range of values, with favorable results around $\lambda_1 = 0.7$. These results suggest that

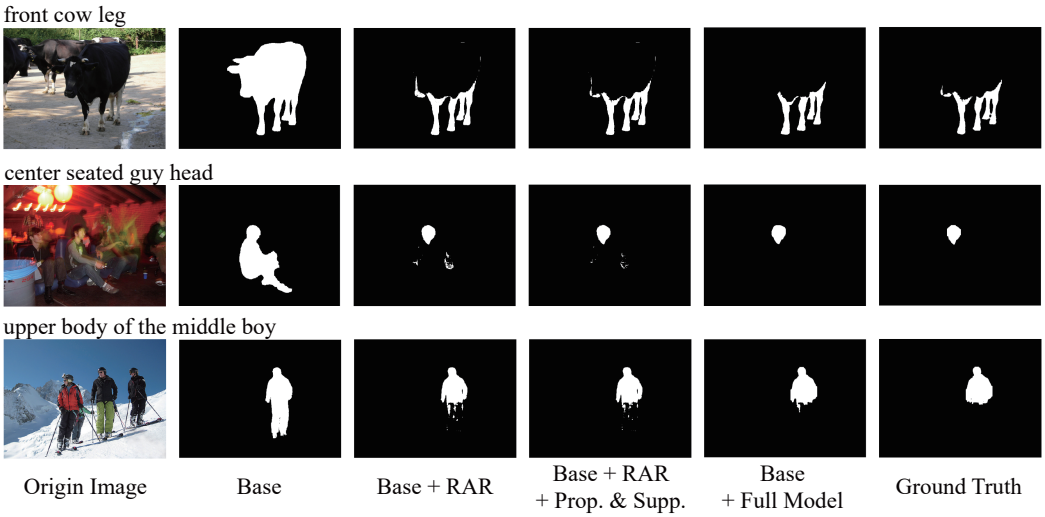


Figure 7: **Qualitative results of stage-level and rectification-branch ablations.** We visualize the predictions of the base model, Base+RAR, Base+RAR+Prop. & Supp., and the full model. RAR constrains the prediction to the correct host-object scope, while the propagation and suppression branches independently recover missing regions and suppress drifted activations. With cross-branch collaboration, the full model better balances positive and negative updates and produces masks that are more consistent with the ground truth.

excessively weak rectification may leave initial prediction errors uncorrected, while overly strong rectification may introduce over-propagation or over-suppression. Therefore, we set $\lambda_1 = 0.7$ and $\lambda_2 = 0.5$ in all experiments.

C.2 Additional Ablation Study

Referential Reformulation Ablation. We further examine the importance of the language-side reformulation in RAR. As shown in Table 9, directly using the original expression (see "Original Expression + BCR") keeps the host-object and target-part entangled in a single holistic description. In this case, the expression still contains strong object-level semantics, and the base model may treat the host-object as the main referential target. Although BCR can refine the resulting score map, it cannot fully correct the attribution error if the initial localization has already been biased toward the whole object. Using only the target-part description (see "target-part Only + BCR") reduces host-object dominance by removing explicit object-level words. However, it also discards the host-object context needed to identify the correct instance, especially when multiple similar objects or similar parts appear in the image. As a result, the model may localize a semantically matched part from the wrong object or produce unstable target-part responses. The full reformulation (see "Full Reformulation + BCR") performs best because it decouples visual scope determination from target-part localization. The host-object description first identifies the correct object instance and provides a constrained search region, while the target-part description focuses the prediction on the intended fine-grained region. This design reduces referential attribution ambiguity before rectification and provides a more reliable initial score map for BCR.

Table 9: Referential reformulation ablation on the val-Part subset of RefCOCOm.

Setting	mIoU
Original Expression + BCR	22.1
target-part Only + BCR	17.2
Full Reformulation + BCR	**24.4**

Text-Split Module Ablation. We further evaluate different tools for decomposing the original expression into host-object and target-part descriptions. As shown in Table 10, the dependency-based parser (see "dependency-based parser") yields substantially lower performance than the two instruction-following language models. Its reliance on syntactic dependencies and part-related keywords makes it difficult to handle expressions in which part words serve

Table 10: Text-split module ablation on the val-Part subset of RefCOCOm.

Setting	mIoU
dependency-based parser	17.93
Qwen3-8B	24.37
Llama-3-8B-Instruct	24.38

as modifiers rather than the actual referential target. In contrast, Llama-3-8B-Instruct (see "Llama-3-8B-Instruct") and Qwen3-8B (see "Qwen3-8B") achieve nearly identical performance under the same prompt setting. These results suggest that the effectiveness of RAR is not tied to a particular language model for text splitting. We use Llama-3-8B-Instruct as the default text-splitter in our experiments.

C.3 Layer-wise Analysis of Image Attention

We analyze the layer-wise image attention ratio of the base vision-language model to determine the layers used for constructing the target-part support map $\mathbf{A}_t$. For each layer, we measure the proportion of prefill-stage attention assigned to image tokens and normalize the resulting ratios by the maximum value across all layers.

As shown in Fig. 8, the first eight layers exhibit relatively high normalized image attention ratios and account for a substantial portion of the visual attention responses. This suggests that the early layers provide strong visual signals during the prefill stage. We therefore aggregate the target-part-to-image attention from layers 1–8 to construct $\mathbf{A}_t$.

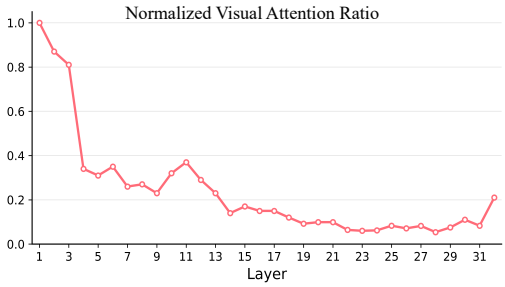


Figure 8: Layer-wise image attention ratios during the prefill stage.

D Limitations

Although our framework improves fine-grained RES in a training-free manner, it still has several limitations. First, RAR relies on language-side decomposition of referring expressions, which may be unreliable for highly complex or grammatically ambiguous expressions. For example, in an expression such as "person in the background in white shorts holding something white hand", the intended target-part is "hand", but the text-split module may incorrectly identify the host-object or target-part due to ambiguous modifiers and implicit host–part relations. Such decomposition errors can propagate to the subsequent host-object and target-part localization stages, limiting the effectiveness of the framework.

Second, our method incurs additional inference cost compared with directly applying the base model. RAR requires sequential host-object and target-part localization, while BCR extracts internal target-part support from the base model and computes patch-to-patch structural affinity for rectification. These operations introduce additional computation and memory overhead, particularly when processing high-resolution images. Improving inference efficiency while preserving fine-grained segmentation quality remains an important direction for future work.

E Prompt for Part Identification

We employ a lightweight instruction prompt to identify object parts and decompose referring expressions into part and whole descriptions. The complete prompt is provided in Fig. 9.

F Qualitative Results

Uniform Prompt

You are a professional assistant for multi-granularity RES. Determine whether the caption describes a specific PART of an object or the WHOLE object.

Rules:
1. A part must be a physical component of an object, e.g., a dog's tail, a car's wheel, or a person's head.
2. Spatial/positional descriptions (e.g., left, right, top, middle, corner, "4th from left") describe the WHOLE object, not a part.
3. Action/verb phrases (e.g., holding, running, crossed) describe the state or behavior of the WHOLE object. Do not extract nouns inside such phrases as parts.
4. Some captions follow a simple "whole + part" pattern, such as "middle dog head" or "right human head". In these cases, identify the physical part and keep positional modifiers with the whole. Preserve all other modifiers with the noun they describe.

Examples:
Input: "white car wheel"
Output: {"ans": "yes", "part": "white wheel", "whole": "car"}

Input: "middle dog head"
Output: {"ans": "yes", "part": "head", "whole": "middle dog"}

Output requirements:
- Output strictly valid JSON with exactly three fields:
  "ans": "yes" or "no",
  "part": part description or "",
  "whole": whole description.
- Preserve all relevant modifiers and assign them to the correct part or whole.
- Output no explanations or extra text.

Now process:
{sent}

Figure 9: Prompt for identifying object parts and decomposing referring expressions into part and whole descriptions.

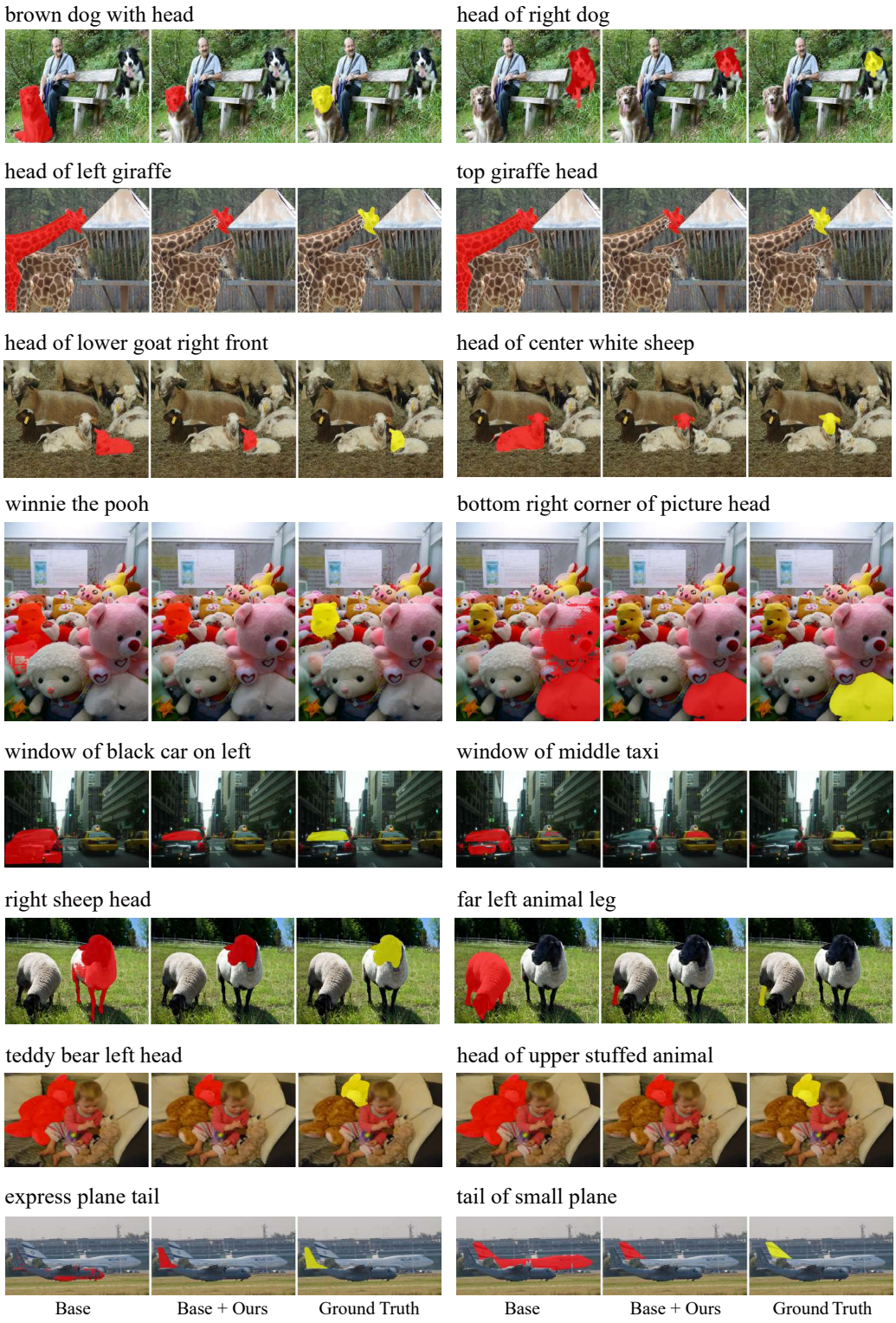


Figure 10: **Additional qualitative comparison between LISA and LISA+Ours.** We provide additional visualization results comparing the original LISA predictions with those refined by our method. Our framework improves fine-grained target-part localization by reducing host-object bias and correcting incomplete or drifted predictions.

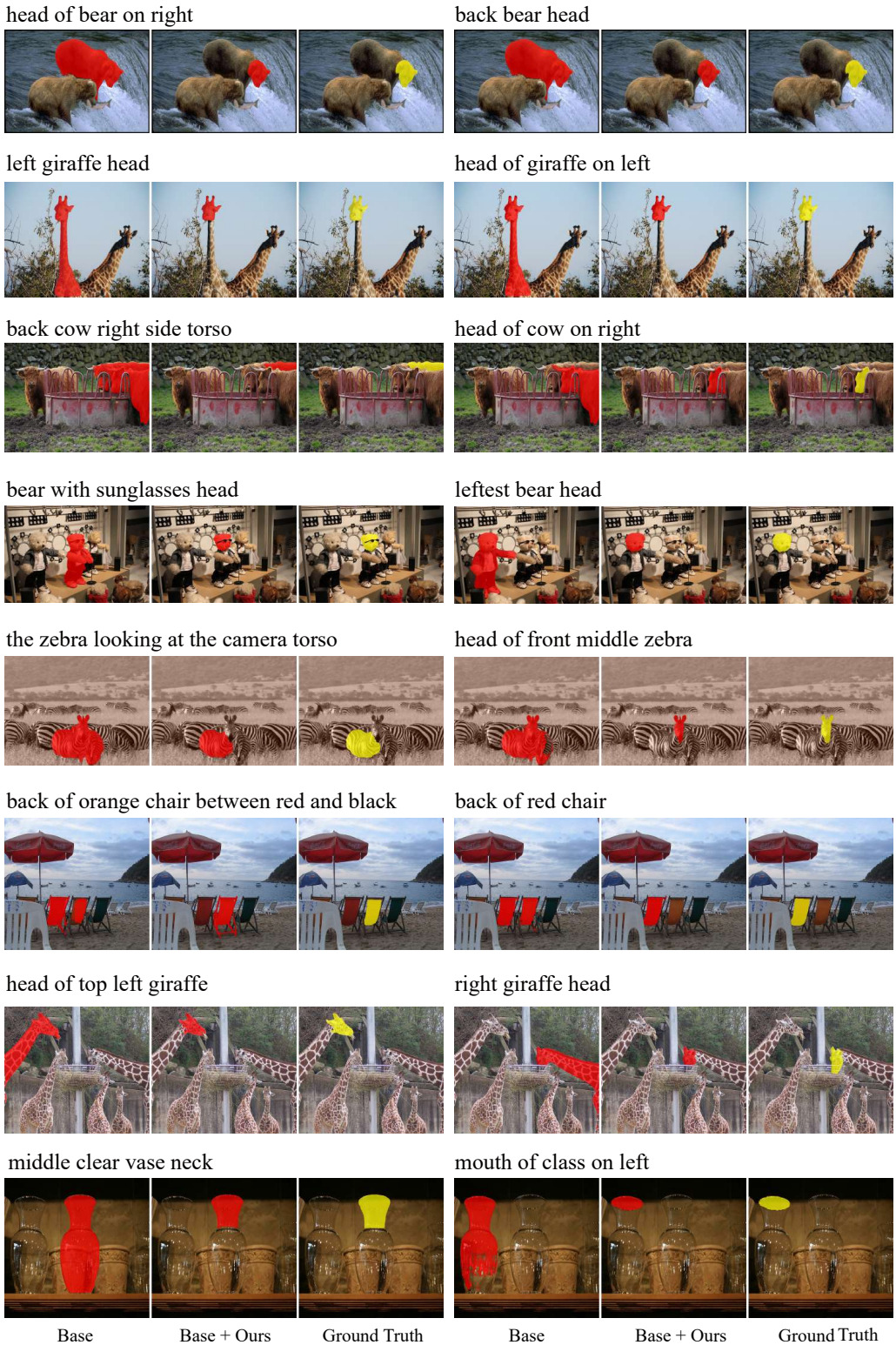


Figure 11: **Additional qualitative comparison between M^{2}SA and M^{2}SA+Ours.** We provide additional visualization results comparing M^{2}SA with M^{2}SA enhanced by our method. Our framework further improves the fine-grained segmentation results of existing fine-grained methods by alleviating referential attribution ambiguity and refining target-part score maps.



Figure 12: **Additional qualitative comparison between Qwen+SAM and Qwen+SAM+Ours.** We provide additional visualization results comparing the Qwen+SAM combination with its variant enhanced by our method. The results show that our framework can also improve MLLM-based segmentation pipelines by guiding predictions toward the referred target-part and suppressing object-level or irrelevant activations.