TAROT-SAM3: TRAINING-FREE SAM3 FOR ANY REFERRING EXPRESSION SEGMENTATION

Anonymous authors
Paper under double-blind review

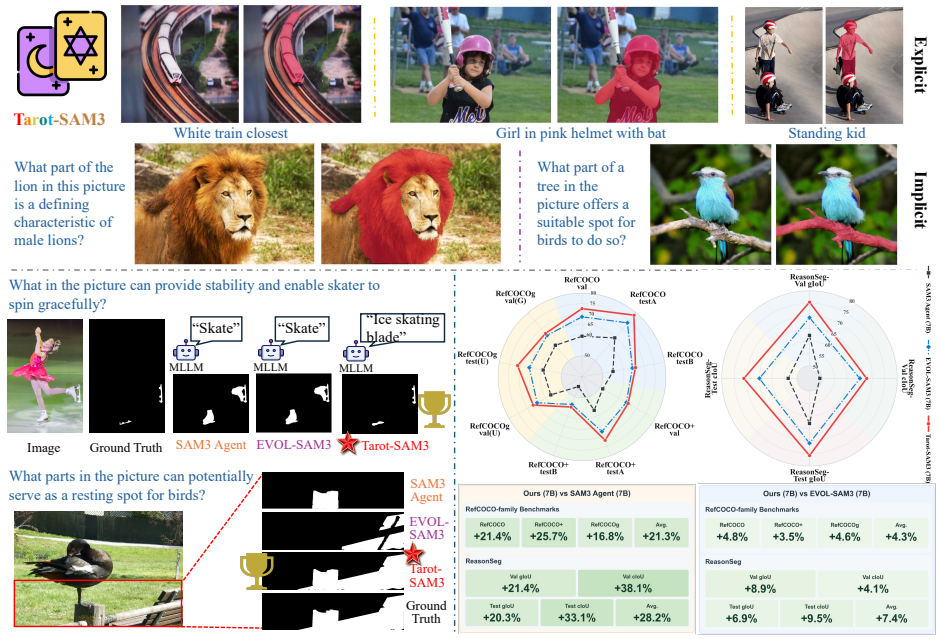

Figure 1: Tarot-SAM3 accurately segments targets from explicit (top) and implicit (middle) referring expressions. Bottom-left example comparisons demonstrate our method's advantages in MLLM parsing and SAM3 mask quality over existing training-free MLLM-SAM3 methods Carion et al. (2025); Ye et al. (2025). Bottom-right radar charts and tables highlight Tarot-SAM3's consistent gains on RefCOCO-family and ReasonSeg with the same 7B backbone.

ABSTRACT

Referring Expression Segmentation (RES) bridges vision and language but traditionally relies on extensive annotated datasets and struggles with implicit queries. While the Segment Anything Model 3 (SAM3) excels in promptable segmentation, its application to RES exposes **two critical challenges**: 1) SAM3 lacks the intrinsic capacity to decode lengthy, implicit expressions, and 2) it is highly vulnerable to both erroneous grounding priors and biased evaluation preferences from Multimodal Large Language Models (MLLMs), leaving its mask predictions unrefined. To address this, we introduce **Tarot-SAM3**, a novel training-free framework driven by the **key idea** of dual-stage refinement over MLLM-derived grounding cues and SAM3 mask predictions. Our **technical contribution** is two-fold. First, an Expression Reasoning Interpreter (**ERI**) is designed to guide the MLLM through structured parsing and evaluation-aware rephrasing, translating arbitrary queries into error-tolerant, heterogeneous prompts for SAM3 mask generation and performing consensus-driven mask selection within each prompt type. Second, the Mask Self-Refining (**MSR**) phase establishes a DINOv3-guided feedback loop, which identifies the most reliable mask and adaptively reconfigures point prompts to automatically resolve over- and under-segmentation. Extensive experiments demonstrate that Tarot-SAM3 achieves state-of-the-art results among MLLM-SAM3 architectures across MLLM scales, while under the Qwen2.5-VL 7B setting, it delivers **21.3%** and **28.2%** average relative improvements over the SAM3 Agent on explicit and implicit benchmarks, respectively. Open-world evaluations confirm our framework's superior generalization and robustness.

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.

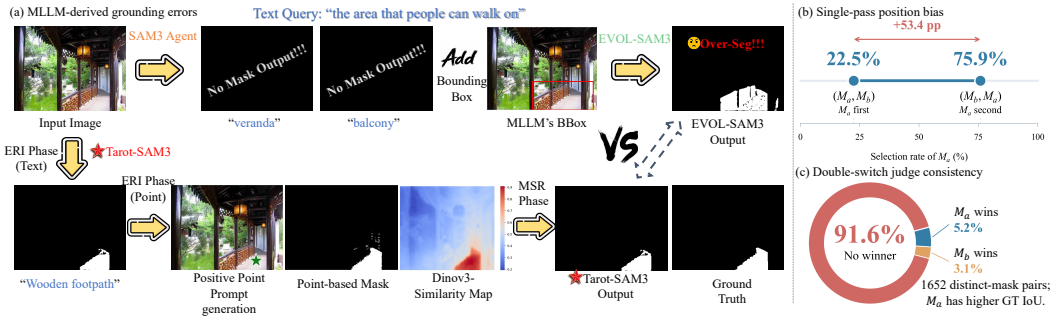

Figure 2: Motivation of Tarot-SAM3. (a) Erroneous MLLM-derived concepts and localization cues can cause mask-generation failure or over-segmentation. (b) Changing the position of M_a (higher GT-IoU) shifts its MLLM single-pass selection rate from **22.5%** to **75.9%**, revealing strong MLLM position bias.(c) Existing double-swith method Ye et al. (2025) yields no consistent winner in **91.6%** of 1,652 distinct-mask pairs, failing to ensure reliable mask selection by the MLLM.

1 Introduction

Referring Expression Segmentation (RES) Yu et al. (2018); Huang et al. (2020); Yang et al. (2022); Wang et al. (2022) aims to segment image regions described by natural-language expressions, serving as a key benchmark for fine-grained vision–language grounding and supporting applications such as multimodal reasoning Lu et al. (2024); Yang et al. (2023b), embodied navigation Driess et al. (2023); Xiao et al. (2026), and vision-based dialogue Song et al. (2025); Yi et al. (2024). Specifically, textual queries can be broadly divided into two types (see Fig. 1): *explicit expressions*, which directly indicate the target object, and *implicit expressions*, which require reasoning over attributes, relations, or commonsense knowledge to infer the target. Despite recent progress, existing methods Jing et al. (2021); Liang et al. (2023a); Kim et al. (2022); Xu et al. (2023) still rely on large-scale annotated data and are often tailored to either *explicit* or *implicit* expressions. This limits their scalability and generalization, leaving *any referring expression* segmentation largely underexplored.

Recently, the Segment Anything Model 3 (SAM3) has demonstrated strong robustness and generalization in text-prompted segmentation Carion et al. (2025); Li et al. (2025b); Jiang et al. (2026). However, extending SAM3 to RES remains challenging. First, SAM3 is not inherently designed to interpret long, relational, or implicit referring expressions. Second, existing MLLM–SAM3 pipelines Carion et al. (2025); Ye et al. (2025) primarily mitigate grounding errors through iterative candidate comparison, but neither correct erroneous target concepts or localization cues (Fig. 2(a)) before SAM3 inference nor account for MLLM-induced spatial preference bias (Fig. 2(b, c), detailed in Appendix Tab. 6) during candidate selection or further refine the resulting SAM3 masks. These challenges motivate a novel question: *How can a training-free MLLM–SAM3 framework jointly mitigate MLLM-induced grounding errors and inconsistent judgments, while automatically refine SAM3's mask predictions for any referring expressions?*

To this end, we propose **Tarot-SAM3**, a novel training-free framework featuring a unified **Reasoning-Calibrated Grounding and Closed-Loop Mask Self-Refinement** mechanism that realizes dual-stage refinement over MLLM-derived grounding cues and SAM3 mask predictions through structured expression reasoning and DINOv3-guided feature coherence Sim'eoni et al. (2025). Our contributions are two-fold: 1) The **Expression Reasoning Interpreter (ERI) phase** addresses the vulnerability of existing MLLM-SAM3 couplings to erroneous MLLM-derived grounding cues, mitigating the error propagation caused by misparsed target concepts or inaccurate localization cues (see the output visualizations in the top row of Fig. 2(a). Specifically, ERI utilizes dedicated reasoning-assisted prompt options to enhance the controllability and precision of MLLM outputs, enabling structured expression parsing and evaluation-aware rephrasing to generate reliable heterogeneous prompts, while incorporating consensus-driven mask selection within each prompt type. 2) The **Mask Self-Refining (MSR) phase** further improves segmentation quality by selecting the most MLLM-consensus-supported mask and refining it through DINOv3's object-aware feature coherence. By modeling feature relationships within discriminative regions, MSR identifies and corrects over- and under-segmentation. Extensive experiments show that Tarot-SAM3 achieves state-of-the-art performance among training-free MLLM-SAM3 methods across explicit and implicit RES benchmarks. Under the same MLLM setting, it surpasses SAM3 Agent Carion et al.

(2025) by **+14.2** cIoU on RefCOCO testA and **+17.7** cIoU on ReasonSeg, while still improving over EVOL-SAM3 Ye et al. (2025) by **+4.8** cIoU and **+6.2** cIoU, respectively. Fig. 9 further demonstrates the strong open-world generalization of Tarot-SAM3, validating the effectiveness of our framework.

2 Related Work

Referring Expression Segmentation. Existing RES methods Li et al. (2018); Yu et al. (2018); Ye et al. (2019); Ding et al. (2022); Hu et al. (2023) have evolved from early multimodal fusion pipelines to recent foundation-model-based frameworks. Early CNN–LSTM-based RES methods Yu et al. (2018); Jing et al. (2021); Hu et al. (2016); Liu et al. (2017); Yang et al. (2021) relied on feature concatenation for multimodal fusion, while subsequent attention-based approaches Huang et al. (2020); Liu et al. (2021); Liang et al. (2023a) improved cross-modal alignment via graph and parsing reasoning. Recent Transformer and CLIP-based frameworks Yang et al. (2022); Kim et al. (2022); Wang et al. (2022); Xu et al. (2023) further enhance vision–language correspondence and generalization. Despite substantial progress in explicit RES, implicit RES introduces additional reasoning requirements involving attributes, relations, and commonsense knowledge. Introduced with LISA Lai et al. (2024), ReasonSeg established the reasoning-before-segmentation paradigm, inspiring subsequent methods Yang et al. (2023a); Chen et al. (2024); Sapkota & Karkee (2025); Liu et al. (2025b;a) to integrate MLLMs and chain-of-thought reasoning for stronger logical inference and segmentation. However, these approaches Hu et al. (2020); Wang et al. (2024) remain annotation-intensive, limiting generalization across diverse expressions. *In contrast, Tarot-SAM3 introduces a training-free paradigm that combines reasoning-calibrated MLLM grounding with DINOv3-guided mask self-refinement, enabling arbitrary RES without task-specific fine-tuning.*

SAM3 and Its Adaptations for RES. The Segment Anything Model 3 Carion et al. (2025) exhibits strong robustness and open-vocabulary generalization through large-scale training on billions of masks and diverse vision–language data. However, its concept-level prompting interface remains ill-suited to long, compositional, or implicit referring expressions Carion et al. (2025); Li et al. (2025b). To bridge this gap, SAM3 Agent Carion et al. (2025) leverages an MLLM to translate complex expressions into simplified prompts for SAM3; SAM3-I Li et al. (2025b) introduces a training-based instruction-aware adaptor to enhance structured language understanding; and EVOL-SAM3 Ye et al. (2025) adopts a training-free evolutionary prompting strategy to iteratively refine MLLM-generated prompts. However, existing methods remain sensitive to erroneous MLLM-derived grounding cues and potential position bias in pairwise mask evaluation, while lacking explicit self-refinement to correct SAM3 masks. *Therefore, Tarot-SAM3 introduces ERI for reasoning-calibrated grounding of MLLM outputs and MSR for feature-coherent mask refinement of SAM3 predictions.*

3 Method

3.1 Overview

As shown in Fig. 3, **Tarot-SAM3** is a *training-free* framework that empowers SAM3 to handle both explicit and implicit referring expressions. Therefore, Given an image X and a text query T as inputs, we treat all backbone models—$\mathcal{E}_{\text{MLLM}}$, $\mathcal{E}_{\text{DINO}}$, and $\mathcal{E}_{\text{SAM}}$—as frozen. However, a straightforward integration of these pretrained models is insufficient: (1) A naïve coupling of an MLLM with SAM3 directly relies on MLLM-derived grounding cues; errors in expression interpretation or localization can therefore propagate to SAM3 and yield inaccurate masks. (2) SAM3 remains a one-shot mask predictor, lacking a feedback mechanism to identify and correct segmentation errors. Although DINOv3 provides rich feature-level semantic coherence, its similarity maps cannot be directly translated into actionable corrective signals for mask refinement. To address these limitations, we introduce two novel phases: the Expression Reasoning Interpreter (ERI)(Sec. 3.2) and the Mask Self-Refining (MSR)(Sec. 3.3) phase. Our code is available anonymously.[1]

3.2 Expression Reasoning Interpreter (ERI) Phase

The ERI phase aims to transform expressions into reliable heterogeneous prompts compatible with SAM3 for accurate mask generation. To reduce grounding ambiguity and improve the controllability of MLLM outputs, we first introduce **Reasoning-assisted Prompt Options** T_R, which structure

[1] https://anonymous.4open.science/r/Tarot-SAM3_Code

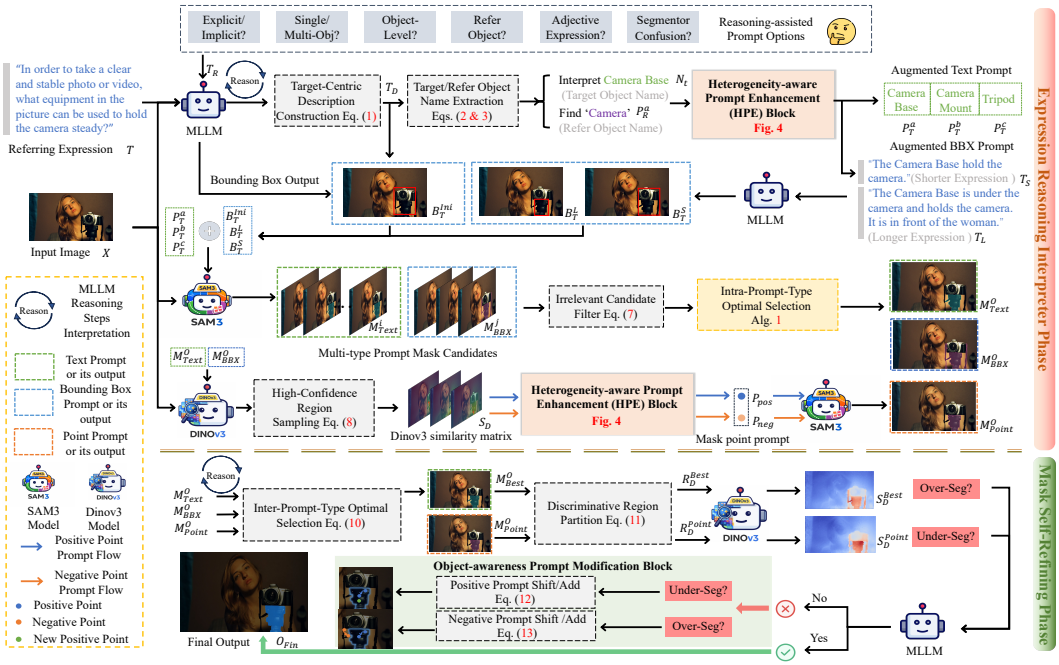

Figure 3: Overview of our Tarot-SAM3 Framework.

the reasoning process along six dimensions: *explicit/implicit identification*, *single/multi-object determination*, *refer-object analysis*, *adjective-expression identification*, *object-level reasoning*, and *segmentor-confusion awareness*. Collectively, these dimensions establish a task-oriented reasoning scaffold that transforms unconstrained MLLM inference into a structured decomposition of expression semantics and segmentation-aware ambiguities, thereby enabling more reliable target grounding; *further details are provided in Sec. B.1.1 of the Appendix.* Conditioned on T, X, and T_R, ERI first employs the MLLM to reformulate the original expression into a target-centric description T_D, making target-specific constraints and relational cues explicit for subsequent grounding:

$$T_D = \text{Describe}_{\text{MLLM}}\left(X, T; T_R\right). \tag{1}$$

Based on the constructed target-centric description T_D, ERI then identifies the target entity and associated referential entities:

$$N_t = \text{Target}_{\text{MLLM}}(T_D), \tag{2}$$

where $\text{Target}_{\text{MLLM}}(\cdot)$ identifies the target object name to be segmented from T_D. Meanwhile, the refer-object set is extracted as:

$$\mathcal{N}_r = \{n \in \text{Noun}(T_D) \setminus \{N_t\} \mid \text{Rel}_{\text{MLLM}}(n, N_t, T_D) = 1\}, \tag{3}$$

where $\text{Noun}(\cdot)$ extracts noun candidates from T_D, and $\text{Rel}_{\text{MLLM}}(\cdot)$ determines whether an entity provides contextual evidence for identifying the target object. After identifying the N_t, ERI invokes the **Heterogeneity-aware Prompt Enhancement (HPE)** block (Fig. 4 and *Algs. 2 & 4 in the Appendix*) to construct complementary reliable SAM3-compatible prompts. ERI first performs text prompt augmentation by leveraging the MLLM to generate synonymous and conceptually related variants. This enriches the target representation and improves its alignment with SAM3's concept-level segmentation space. We define:

$$\mathcal{P}_T = \text{Aug}_{\text{MLLM}}(N_t) = \left\{P_T^a, P_T^b, P_T^c\right\}, \quad P_T^a = N_t. \tag{4}$$

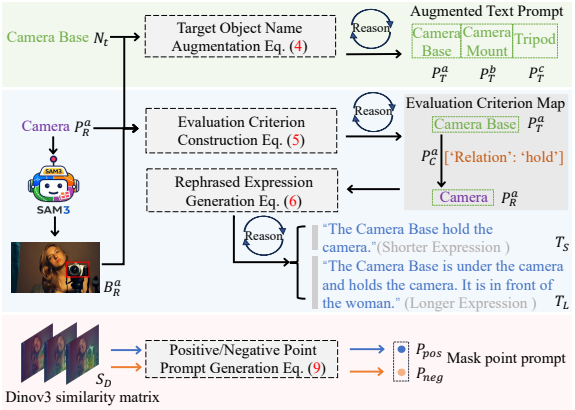

Figure 4: Overview of the HPE block.

where $\mathcal{P}_T$ denotes a set of augmented target text prompts.

Algorithm 1 Intra-Prompt-Type Optimal Selection

Input Filtered masks $\mathcal{M}_{\text{Text}} = \{M^i_{\text{Text}}\}$, $\mathcal{M}_{\text{BBX}} = \{M^j_{\text{BBX}}\}$; SAM3 confidence q; mask stability s
Output $M^O_{\text{Text}}, M^O_{\text{BBX}}$

Dual-branch consensus Top-3 selection

for $t \in \{\text{Text}, \text{BBX}\}$ **do**
  For each candidate M:
  $a_p(M) \leftarrow \max_{N \in \mathcal{M}_{t,p}} \text{IoU}(M, N)$
  $h(M) \leftarrow \sum_p \mathbf{1}[a_p(M) \geq \tau_{\text{con}}]$; $r(M) \leftarrow \text{mean}_p\, a_p(M)$
  $\mathcal{L}_t \leftarrow \text{Sort}^{\text{lex}}_{\downarrow}(\mathcal{M}_t;\, h, r, q)$
  $\mathcal{M}^{(3)}_t \leftarrow \textsc{DiverseTop3}(\mathcal{L}_t, \tau_{\text{con}})$
end for

Text-to-BBox best-mask voting

$\mathbf{w} \leftarrow \text{Reliability}(\mathcal{M}_{\text{Text}}, \mathcal{M}_{\text{BBX}}, q)$
for $B \in \mathcal{M}_{\text{BBX}}$ **do**
  $S(B) \leftarrow \dfrac{\sum_{M \in \mathcal{M}_{\text{Text}}} w_M \,\text{IoU}(B, M)}{\sum_{M \in \mathcal{M}_{\text{Text}}} w_M}$
end for
$\mathcal{W} \leftarrow \{B \in \mathcal{M}_{\text{BBX}} \mid S(B) \geq S(B_F)\}$
$M^O_{\text{BBX}} \leftarrow \arg\max_{B \in \mathcal{W}} \left[S(B) + \frac{1}{2} q(B)\right]$

BBox-to-Text best-mask voting

$\mathcal{G} \leftarrow \text{Group}_{\tau_{\text{con}}}(\mathcal{M}^{(3)}_{\text{Text}})$; $\mathbf{V} \leftarrow \mathbf{0}$
for $B \in \mathcal{M}^{(3)}_{\text{BBX}}$ **do**
  $w_B \leftarrow q(B)s(B)r(B)$
  $a_g \leftarrow \max_{M \in G_g} \text{IoU}(B, M)$
  if a_g has a unique maximum at g^* and $a_{g^*} > \tau_{\text{vote}}$ **then**
    $\Delta_B \leftarrow a_{g^*} - a_{\text{runner-up}}$
    $V_{g^*} \leftarrow V_{g^*} + w_B \Delta_B$
  end if
end for
$M_0 \leftarrow \textsc{Majority}(\mathcal{G}, \mathbf{I}, \mathbf{V}, \mathbf{Q})$
$M^O_{\text{Text}} \leftarrow \textsc{GuardedUpdate}(M_0, \mathcal{M}^{(3)}_{\text{Text}}, \mathcal{E})$

For each refer object in $\mathcal{N}_r$, ERI combines SAM3-based grounding with MLLM-based relation reasoning to construct criteria for subsequent mask evaluation. Specifically, ERI queries SAM3 with the refer-object prompt P^a_R to obtain its bounding box B^a_R, and then conditions the MLLM on (N_t, P^a_R, B^a_R) to derive an *evaluation criterion map* encoding the expected relationship between the target and refer object:

$$P^a_C = \text{Crit}_{\text{MLLM}}(N_t, N_r, B^a_R), \tag{5}$$

where P^a_C denotes a structured evaluation criterion for subsequent mask assessment. Guided by P^a_C, ERI prompts the MLLM to reformulate the original complex query into two complementary target-centric expressions: *shorter expression* T_S and *longer expression* T_L that incorporates relational context between the target and refer objects:

$$(T_S, T_L) = \text{Rephrase}_{\text{MLLM}}(T, N_t, \mathcal{N}_r, P^a_C). \tag{6}$$

We then feed the (T_S, T_L) together with the T_D into the MLLM to obtain target bounding box prompts $(\{B^{Ini}_T, B^S_T, B^L_T\})$. Using $\mathcal{P}_T$ and $\{B^{Ini}_T, B^S_T, B^L_T\}$, ERI queries SAM3 to obtain text- and box-prompt mask candidates, denoted by $\{M^i_{\text{Text}}\}$ and $\{M^j_{\text{BBX}}\}$, respectively. Rather than performing exhaustive pairwise MLLM comparisons among all candidates, as in EVOL-SAM3 Ye et al. (2025) and SAM3 Agent Carion et al. (2025), ERI first applies a **lightweight cross-prompt consistency filter** to eliminate spatially implausible masks. Such pre-filtering also reduces reliance on direct MLLM pairwise judgments, which may suffer from position bias with respect to the ordering of visual candidates, as observed in EVOL-SAM3. This prevents noisy candidates and comparison bias from propagating into subsequent refinement while reducing reasoning overhead:

$$M^i_{\text{Text}} \leftarrow \left\{ M^i_{\text{Text}} \mid \text{IoU}(BBox(M^i_{\text{Text}}), B^k_T) > \tau_{ov} \right\}, \tag{7}$$

where τ_{ov} is an overlap threshold. Subsequently, rather than relying on additional MLLM-based mask comparison, which may inherit erroneous semantic cues or comparison bias and also introduce extra reasoning calls, we propose a **SAM3-consensus-driven Intra-Prompt-Type Optimal Selection** strategy (Alg. 1). The **key idea** is to exploit both *intra-branch consensus* and *cross-branch voting* to identify reliable representatives from the two branches. Within each branch, candidates are ranked lexicographically by the number of supporting prompt routes, mean agreement across routes, and SAM3 confidence, retaining a diverse Top-3 set $\mathcal{M}^{(3)}_t$. In Text-to-BBox voting, all filtered text masks provide reliability-weighted overlap support for the BBox candidates. Candidates with support no lower than that of the T_D BBox mask remain eligible, and the additive score of voting support and SAM3 confidence determines M^O_{BBX}. Conversely, the Top-3 BBox masks vote for consensus clusters formed from the Top-3 text masks. Each vote requires a uniquely preferred cluster with sufficient overlap and is weighted by the voter's reliability and its matching margin over the runner-up. The resulting majority hypothesis is conservatively updated to obtain M^O_{Text}. *Further details are provided in Appendix Sec. B.1.3.* After obtaining the optimal masks M^O_{Text} and M^O_{BBX}, ERI further exploits the point-prompt capability of SAM3 for refinement. ERI first defines a high-confidence region as the overlap of the two masks and samples anchor pixels within this region. We aggregate the anchor-based DINOv3 similarity maps using the pixel-wise harmonic mean:

$$\Omega = M^O_{\text{Text}} \cap M^O_{\text{BBX}}, \quad \mathcal{P} = \text{Sample}(\Omega), \quad S_p(\mathbf{x}) = \text{Sim}_{\text{DINOv3}}(p, X)(\mathbf{x}), \quad S_D(\mathbf{x}) = \frac{|\mathcal{P}|}{\sum_{p \in \mathcal{P}} \frac{1}{S_p(\mathbf{x})}} \tag{8}$$

where $\mathcal{P}$ is the sampled anchor set, and S_D is the aggregated DINOv3 similarity map. Based on S_D, we generate positive and negative point prompts as:

$$p_{\text{pos}} = \arg\max_{x \in \Omega} S_D(x), \quad p_{\text{neg}} = \arg\max_{x \in \overline{\Omega}} \left(-\nabla S_D(x) \cdot \mathbf{d}_{\text{pos}(x)}\right), \tag{9}$$

where $\nabla S_D(x)$ denotes the spatial gradient at candidate location x, and $\mathbf{d}_{\text{pos}}(x)$ is the unit direction from p_{pos} to x. We select p_{pos} as the peak-response pixel in Ω, and p_{neg} as the nearest exterior pixel along the steepest similarity-decay direction, subject to $S_D(p_{\text{neg}}) < s_{\text{neg}}$. The resulting point prompts are fed into SAM3 to obtain M^O_{Point}.

3.3 Mask Self-Refining (MSR) Phase

The MSR phase aims to realize the closed-loop mask-to-prompt refinement stage by coupling DINOv3 feature coherence with MLLM-based region-affiliation reasoning to diagnose and correct over- and under-segmentation. MSR begins with a **Qwen-consensus-based** inter-prompt-type best-mask selection. To reduce selection bias, we repeatedly query Qwen with ERI's three candidates ($\mathcal{C}^O$) in shuffled orders, accepting a mask only upon consensus:

$$\begin{aligned} \widehat{M}_k &= \text{Select}_{\text{MLLM}}\Big(X, T; \text{Shuffle}_k(\mathcal{C}^O)\Big), \quad k = 1, 2, 3, \\ M^O_{\text{Best}} &= \begin{cases} \widehat{M}_1, & \widehat{M}_1 = \widehat{M}_2 = \widehat{M}_3, \\ M^O_{\text{Text}}, & \text{otherwise.} \end{cases} \end{aligned} \tag{10}$$

where $\text{Shuffle}_k(\cdot)$ randomly changes the display order of the three unlabeled masks in the k-th round, and $\widehat{M}_k$ denotes the underlying mask selected by the MLLM. A mask is accepted only when all three rounds reach the same selection; otherwise, we conservatively fall back to M^O_{Text}. Despite being preferred across prompt types, M^O_{Best} may still exhibit over- or under-segmentation. MSR therefore contrasts it with the complementary mask M^O_{Point}, which integrates textual and spatial cues, to derive discriminative regions for subsequent region-affiliation analysis:

$$R_D^{\text{Best}} = M^O_{\text{Best}} \setminus M^O_{\text{Point}}, \qquad R_D^{\text{Point}} = M^O_{\text{Point}} \setminus M^O_{\text{Best}}. \tag{11}$$

Here, R_D^{Best} and R_D^{Point} denote regions exclusive to M^O_{Best} and M^O_{Point}, respectively. MSR verifies masks through region-level target-affiliation reasoning. Using a central pixel from each region as a DINOv3 anchor, we compute S_D^{Best} and S_D^{Point} and derive adaptive thresholds from their respective within-region minima to localize potentially extraneous and missing target regions, respectively. These filtered regions, together with the overlapping high-confidence region, are then provided to the MLLM to determine whether they belong to the same target object. Inconsistency between R_D^{Best} and the high-confidence overlap indicates over-segmentation, whereas consistency between R_D^{Point} and the overlap indicates under-segmentation. Upon detecting over- or under-segmentation, MSR activates the **Object-Awareness Prompt Modification (OPM)** block to update point prompts and re-query SAM3. For under-segmentation, OPM shifts the positive point to the center of R_D^{Point} or adds an additional positive point within it:

$$\mathcal{P}'_{\text{pos}} = \begin{cases} \{\, p_{\text{pos}} \leftarrow c(R_D^{\text{Point}}) \,\}, & \text{(shift)} \\ \mathcal{P}_{\text{pos}} \cup \{\, c(R_D^{\text{Point}}) \,\}, & \text{(add)} \end{cases} \tag{12}$$

where $c(\cdot)$ denotes the region center. For over-segmentation, OPM either shifts the negative prompt to the over-claimed region or adds an additional negative point within it:

$$\mathcal{P}'_{\text{neg}} = \begin{cases} \{\, p_{\text{neg}} \leftarrow c(R_D^{\text{Best}}) \,\}, & \text{(shift)} \\ \mathcal{P}_{\text{neg}} \cup \{\, c(R_D^{\text{Best}}) \,\}, & \text{(add)} \end{cases} \tag{13}$$

Finally, the updated point prompts, together with M^O_{Best}, are fed back to SAM3 to produce the refined mask, enforcing feature coherence and object consistency to correct over- and under-segmentation. *Further algorithmic details are provided in Appendix Alg. 5 and Sec. B.2 .*

4 Experiment

4.1 Datasets and Implementation Details

Datasets. We evaluate Tarot-SAM3 on four explicit and implicit RES benchmarks, assessing both direct visual grounding and reasoning-based target identification. For explicit RES, we use RefCOCO Yu et al. (2016), RefCOCO+ Yu et al. (2016), and RefCOCOg Mao et al. (2016), all derived from MS COCO with natural-language referring expressions. For implicit RES, we evaluate on the ReasonSeg benchmark introduced with LISA Lai et al. (2024), which requires commonsense or logical reasoning. We follow the established benchmark evaluation protocols and report gIoU and cIoU as percentages. *More details are provided in the Appendix Secs. C.1.1 & C.1.2.*

Table 1: Comparison with prior methods on explicit RES benchmarks. **Bold** indicates the overall best cIoU result, and underline denotes the best zero-shot cIoU result. Results with * were obtained from our evaluation.

Method	VLM Backbone	Training	RefCOCO			RefCOCO+			RefCOCOg		
		RES	val	testA	testB	val	testA	testB	val (U)	test (U)	val (G)
LISA	LLaVA 7B	✓	74.9	79.1	72.3	65.1	70.8	58.1	67.9	70.6	–
SAM4MLLM	LLaVA1.6 8B	✓	79.8	82.7	74.7	74.6	80.0	67.2	75.5	76.4	–
PSALM	Phi1.5 1.3B	✓	83.6	84.7	81.6	72.9	75.5	70.1	73.8	74.4	–
EVF-SAM	Extra Data	✓	82.4	84.2	80.2	76.5	80.0	71.9	78.2	78.3	–
RICE	Qwen2.5-7B	✓	83.5	85.3	81.7	**79.4**	82.8	75.4	79.8	80.4	–
MLCD-seg	Qwen2.5-7B	✓	83.6	85.3	81.5	**79.4**	82.9	**75.6**	79.7	80.5	–
HyperSeg	Phi2 2.7B	✓	84.8	85.7	**83.4**	79.0	**83.5**	75.2	79.4	78.9	–
LENS	Qwen2.5-VL 3B	✓	84.2	85.3	81.0	**79.4**	82.8	74.3	81.2	81.0	–
SAM-Veteran	Qwen2.5-VL 7B	✓	–	80.8	–	–	76.6	–	–	73.4	–
SAM-Veteran	Qwen2.5-VL 32B	✓	–	80.4	–	–	77.4	–	–	73.4	–
SegAgent	Qwen-VL 7B	✓	79.2	81.4	75.7	71.5	76.7	65.4	74.8	74.6	–
X-SAM	Phi3 3.8B	✓	**85.1**	**87.1**	**83.4**	78.0	81.0	74.4	**83.8**	**83.9**	–
Ref-Diff	VAE	×	37.2	38.4	37.2	37.3	40.5	33.0	44.0	44.5	44.3
Pseudo-RIS	ETRIS	×	41.1	48.2	33.5	44.3	51.4	35.1	46.0	46.7	46.8
LGD+DINO	ViT-B/32	×	49.5	54.7	41.0	49.6	58.4	38.6	50.3	51.1	52.5
SAM 3 Agent	Qwen2.5-VL 7B	×	59.4	64.3	55.0	51.4	57.0	44.9	57.2	58.8	59.7
SAM 3 Agent	Qwen2.5-VL 72B	×	70.5	74.3	64.6	64.9	70.7	56.6	69.2	69.5	69.8
EVOL-SAM3	Qwen2.5-VL 3B	×	66.8	70.9	59.9	56.1	63.1	49.3	63.0	63.5	64.0
EVOL-SAM3	Qwen2.5-VL 7B	×	68.7	73.7	64.4	64.4	67.8	54.0	64.7	65.5	65.9
EVOL-SAM3*	Qwen2.5-VL 72B	×	71.2	76.6	66.2	67.6	73.8	58.5	70.1	72.2	70.6
Tarot-SAM3 (ours)	Qwen2.5-VL 3B	×	67.3	71.4	58.7	53.6	62.7	48.1	64.5	66.5	64.7
Tarot-SAM3 (ours)	Qwen2.5-VL 7B	×	72.6	78.5	65.9	65.4	72.2	55.3	66.9	71.2	67.1
vs. SAM 3 Agent (7B)			+13.2	+14.2	+10.9	+14.0	+15.2	+10.4	+9.7	+12.4	+7.4
vs. EVOL-SAM3 (7B)			+3.9	+4.8	+1.5	+1.0	+4.4	+1.3	+2.2	+5.7	+1.2
Tarot-SAM3 (ours)	Qwen2.5-VL 72B	×	75.5	79.5	66.7	68.0	75.1	56.7	71.2	74.5	**72.1**
vs. SAM 3 Agent (72B)			+5.0	+5.2	+2.1	+3.1	+4.4	+0.1	+2.0	+5.0	+2.3
vs. EVOL-SAM3 (72B)			+4.3	+2.9	+0.5	+0.4	+1.3	-1.8	+1.1	+2.3	+1.5

Table 2: Comparison with prior methods on ReasonSeg. Training Data indicates whether RES data are used for fine-tuning. **Bold** indicates the overall best result. Results with * were obtained from our evaluation.

Method	VLM Backbone	Training Data		Val Set		Test Set		Test (Short)		Test (Long)	
		RES	ReasonSeg	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU
SEEM	–	×	×	25.5	21.2	24.3	18.7	20.1	11.5	25.6	20.8
Grounded SAM	–	×	×	26.0	14.5	21.3	16.4	17.8	10.8	22.4	18.6
OVSeg	–	×	×	28.5	18.6	26.1	20.8	18.0	15.5	28.7	22.5
SAM4MLLM	Qwen-VL 7B	✓	×	46.7	48.1	–	–	–	–	–	–
Seg-Zero	Qwen2.5-VL 3B	✓	×	58.2	53.1	56.1	48.6	–	–	–	–
Seg-Zero	Qwen2.5-VL 7B	✓	×	62.6	62.0	57.5	52.0	–	–	–	–
X-SAM	Phi3 3.8B	✓	✓	56.6	32.9	57.8	41.0	47.7	48.1	56.0	40.8
LISA	LLaVA1.5 7B	✓	✓	61.3	62.9	55.6	56.9	48.3	46.3	57.9	59.7
LISA	LLaVA1.5 13B	✓	✓	65.0	**72.9**	61.3	62.2	55.4	50.6	63.2	65.3
SAM-Veteran	Qwen2.5-VL 7B	✓	✓	68.2	67.3	62.6	56.1	–	–	–	–
SAM-Veteran	Qwen2.5-VL 32B	✓	✓	72.3	70.0	62.9	58.2	–	–	–	–
SegAgent	Qwen2.5-VL 7B	✓	✓	33.0	25.4	33.5	31.3	–	–	–	–
RSVP	Qwen2-VL 7B	×	×	58.6	48.5	56.1	51.6	48.5	44.3	57.1	53.0
SAM 3 Agent	Qwen2.5-VL 7B	×	×	62.2	49.1	63.0	53.5	59.4	43.5	64.1	56.2
SAM 3 Agent	Qwen2.5-VL 72B	×	×	74.6	65.1	70.8	64.0	70.3	55.7	71.0	66.3
SAM 3 Agent	Llama4 Maverick	×	×	68.5	61.5	67.1	60.9	66.8	**59.4**	67.2	61.3
EVOL-SAM3*	Qwen2.5-VL 3B	×	×	66.8	62.6	66.1	58.0	63.1	49.1	67.1	60.7
EVOL-SAM3*	Qwen2.5-VL 7B	×	×	69.3	65.1	70.9	65.0	65.9	53.4	72.6	67.9
EVOL-SAM3*	Qwen2.5-VL 72B	×	×	73.6	67.2	71.1	64.9	67.9	50.5	72.1	68.7
Tarot-SAM3 (ours)	Qwen2.5-VL 3B	×	×	67.1	55.6	67.0	59.0	64.3	50.7	67.9	61.2
Tarot-SAM3 (ours)	Qwen2.5-VL 7B	×	×	75.5	67.8	75.8	71.2	**71.7**	56.8	77.1	**76.5**
vs. SAM 3 Agent (7B)				+13.3	+18.7	+12.8	+17.7	+12.3	+13.3	+13.0	+20.3
vs. EVOL-SAM3 (7B)				+6.2	+2.7	+4.9	+6.2	+5.8	+3.4	+4.5	+8.6
Tarot-SAM3 (ours)	Qwen2.5-VL 72B	×	×	**75.6**	69.6	**76.0**	**71.8**	71.3	54.5	**77.4**	75.8
vs. SAM 3 Agent (72B)				+1.0	+4.5	+5.2	+7.8	+1.0	-1.2	+6.4	+9.5
vs. EVOL-SAM3 (72B)				+2.0	+2.4	+4.9	+6.9	+3.4	+4.0	+5.3	+7.1

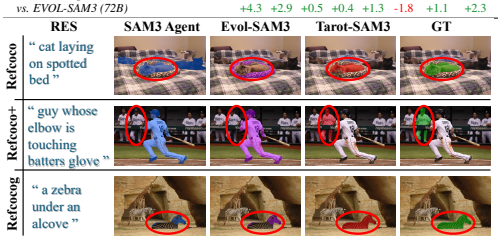

Figure 5: Example visualizations on explicit RES benchmarks, *zoom in for better view*.

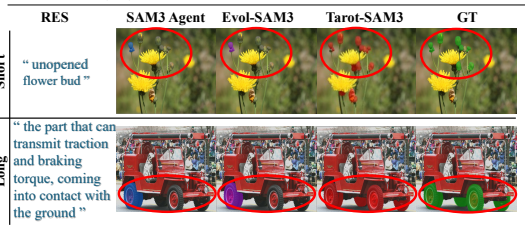

Figure 6: Example visualizations on implicit RES benchmarks, *zoom in for better view*.

Implementation Details. For the backbones, the MLLM is Qwen2.5-VL 3B/7B/72B, and the DINOv3 adopts the ViT-B backbones. All experiments are conducted on 4 NVIDIA A800-80G GPUs. *More implementation details are provided in the Appendix Sec. C.1.3*

4.2 Comparisons with Existing Works

Comparisons with Existing Explicit RES Works. Tab. 1 compares Tarot-SAM3 with supervised and zero-shot methods on RefCOCO, RefCOCO+, and RefCOCOg (also *Appendix* Tab. 10). Although fully supervised models such as X-SAM Wang et al. (2025a) achieve the overall best results, they rely on task-specific training and large-scale annotations. In contrast, Tarot-SAM3 delivers the strongest zero-shot performance across all datasets. Notably, under the 7B MLLM, Tarot-SAM3 achieves **78.5** cIoU on RefCOCO testA, surpassing EVOL-SAM3 Ye et al. (2025) and SAM3 Agent Carion et al. (2025) by **+4.8** and **+14.2** points, respectively. Remarkably, despite using a substantially smaller MLLM, the 7B variant already surpasses SAM3 Agent (72B) and EVOL-SAM3 (72B) on RefCOCO testA by **+4.2** and **+1.9** points, respectively. Qualitative results in Fig. 5 (also Fig. 11 in *Appendix*) further demonstrate its effectiveness.

Comparisons with Existing Implicit RES Works. Tab. 2 compares Tarot-SAM3 with fine-tuned and training-free methods on ReasonSeg, with more comprehensive results in *Appendix* Tab. 11. Our 72B variant ranks best overall on the test set (**76.0/71.8** gIoU/cIoU). Among training-free methods using Qwen2.5-VL 7B, Tarot-SAM3 achieves the best test-set performance of **75.8** gIoU and **71.2** cIoU, exceeding EVOL-SAM3 and SAM3 Agent by **4.9** and **12.8** gIoU points, respectively. Our 7B model also surpasses both 72B baselines and outperforms the fine-tuned SAM-Veteran (7B) Du et al. (2026) by **13.2** gIoU points. Further comparisons are provided in Fig. 6 and *Appendix* Fig. 12.

4.3 Ablation Studies

Effectiveness of ERI phase. In the ERI phase, we analyze the effectiveness of three key designs: Reasoning-assisted Prompt Options (RPO), Text Prompt Augmentation (TA)(Eq. 4), and BBox

Table 3: Ablation study on two key phases of Tarot-SAM3. TA and BA denote Text Prompt Augmentation and BBox Augmentation.

ERI Phase			MSR Phase		RefCOCO (testA)		ReasonSeg (Test Set)	
RPO (T_R)	TA	BA	IPS (Eq. 10)	OPM	gIoU	cIoU	gIoU	cIoU
×	×	×	×	×	35.9	40.4	43.5	38.7
✓	×	×	×	×	75.9	72.7	70.3	60.8
✓	✓	×	×	×	77.1	73.4	70.6	63.6
✓	✓	✓	×	×	77.8	74.2	72.1	65.7
✓	✓	✓	✓	×	79.6	77.4	75.3	70.9
✓	✓	✓	✓	✓	**80.4**	**78.5**	**75.8**	**71.2**

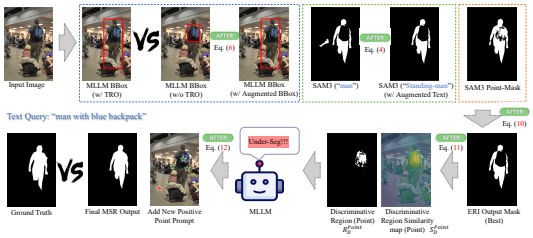

Figure 7: Ablation visualizations of the sequential application of the two effective phases. Zoom in for better review.

Table 4: Sensitivity analysis of key hyperparameters on the RefCOCO testA split.

τ_{ov}	gIoU	cIoU
0.20	79.7	77.4
0.30	**80.4**	**78.5**
0.40	80.1	78.2

(a)

τ_{con}	gIoU	cIoU
0.85	79.6	77.3
0.90	**80.4**	**78.5**
0.95	79.2	77.6

(b)

τ_{vote}	gIoU	cIoU
0.40	77.6	74.3
0.50	**80.4**	**78.5**
0.60	80.0	77.6

(c)

s_{neg}	gIoU	cIoU
0.20	79.9	77.8
0.30	**80.4**	**78.5**
0.40	79.7	78.2

(d)

Table 5: Impact of Tarot-SAM3 designs on Evol-SAM3 performance.

Evol-SAM3 baseline			RefCOCO (testA)		ReasonSeg (Test Set)	
+TA	+BA	+MSR	gIoU	cIoU	gIoU	cIoU
–	–	–	76.3	73.7	70.9	65.0
✓	–	–	76.6 (+0.3)	74.4 (+0.7)	71.8 (+0.9)	65.2 (+0.2)
–	✓	–	78.5 (+2.2)	76.3 (+2.6)	70.2 (-0.7)	64.0 (-1.0)
✓	✓	–	79.2 (+2.9)	76.8 (+3.1)	72.0 (+1.1)	66.5 (+1.5)
–	–	✓	76.8 (+0.5)	74.1 (+0.4)	71.2 (+0.3)	65.3 (+0.3)

Augmentation (BA)(Eq. 6). Tab. 3 shows that T_R stabilizes MLLM parsing, improving RefCOCO testA by **+40.0/+32.3** gIoU/cIoU and ReaonSeg by **+26.8/+22.1**. As illustrated in Fig. 7, it enables the MLLM to correctly ground the *man* rather than the *blue backpack*. Target-name augmentation further yields gains of **+1.2**/**+0.7** gIoU/cIoU on RefCOCO testA and **+0.3**/**+2.8** on ReasonSeg, revealing SAM3's sensitivity to semantically similar noun-phrase variants. Moreover, evaluation-guided rephrasing generates more reliable bounding boxes, yielding gains of **+0.7**/**+0.8** gIoU/cIoU on RefCOCO testA and **+1.5**/**+2.1** on ReasonSeg. *These gains demonstrate that ERI effectively calibrates MLLM-derived grounding cues. Reasoning steps and their detailed contents are analyzed in Appendix Sec. C.4.2.* To further examine the generality of ERI's prompt augmentation design, we incorporate TA and BA into EVOL-SAM3 (Tab. 5). Their joint integration yields additional gains of **+2.9/+3.1** gIoU/cIoU on RefCOCO testA and **+1.1/+1.5** on ReasonSeg. These consistent improvements indicate that the proposed prompt augmentation is **not tailored to Tarot-SAM3**, but captures a generally effective design principle for improving MLLM–SAM3 frameworks.

Effectiveness of MSR phase. In the MSR phase, we analyze the effectiveness of the Inter-Prompt-Type Optimal Selection (IPS, Eq. 10) and the Object-awareness Prompt Modification (OPM) block. IPS improves cross-prompt mask selection over direct MLLM selection by **+1.8/+3.2** gIoU/cIoU on RefCOCO testA and **+3.2/+5.2** on ReasonSeg. ***Notably, even before OPM-based mask refinement, Tarot-SAM3 already surpasses EVOL-SAM3 (77.4 vs. 73.7), validating the effectiveness of ERI and IPS in generating and selecting reliable SAM3 masks.*** OPM further improves RefCOCO by **+0.8/+1.1** gIoU/cIoU and ReasonSeg by **+0.5/+0.3**. Together with Fig. 7 (*also see Sec. C.4.5 and Fig. 15 of Appendix*), these gains validate the effectiveness of converting detected mask errors into corrective point prompts, enabling closed-loop mask self-refinement. Tab. 5 further shows that applying MSR to EVOL-SAM3 consistently improves RefCOCO testA by **+0.5/+0.4** gIoU/cIoU and ReasonSeg by **+0.3/+0.3**, further supporting the generality of MSR beyond the Tarot-SAM3 framework and its complementarity to prompt optimization.

Ablation study of hyperparameters. As shown in Tab. 4, Tarot-SAM3 achieves the best performance with $\tau_{ov} = 0.30$, $\tau_{con} = 0.90$, $\tau_{vote} = 0.50$, and $s_{neg} = 0.30$. These settings consistently yield the best performance across both RES benchmarks.

4.4 Other Analysis

Open-World robustness comparisons. We further compare Tarot-SAM3 with existing methods on diverse open-world cases spanning animal cartoons, artistic paintings, 360° imagery, medical, and remote sensing with both explicit and implicit expressions. As shown in Fig. 9 (*see also Appendix Fig. 13 & Fig. 14*), Tarot-SAM3 consistently produces accurate masks across diverse visual domains.

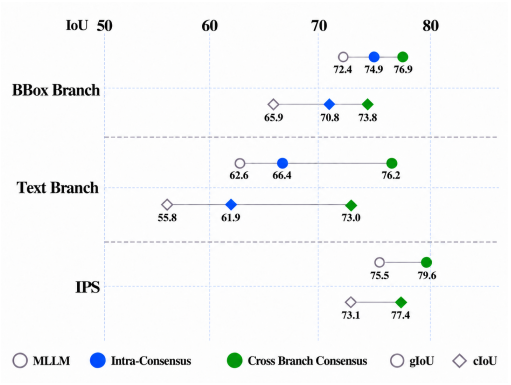

Figure 8: Comparison of mask selection strategies across three branches on RefCOCO testA.

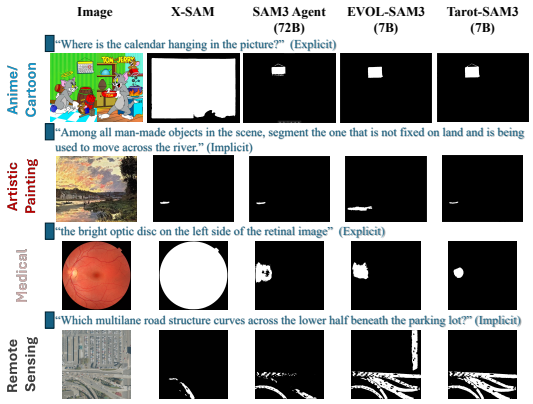

Figure 9: Open-world visual comparisons, *zoom in for better view*.

Validation of RPO Design. To empirically validate the RPO design, we conduct a user study over candidate reasoning steps (*Appendix* Tab. 9), where removing retained dimensions causes rating drops of up to **4.5** in correctness and **6.8** in SAM-friendliness. Consistent downstream ablations in Tab. 14 further show drops of up to **12.7/14.2** cIoU on RefCOCO/ReasonSeg, demonstrating that RPO is empirically grounded rather than heuristically designed.

Beyond Naïve Integration and MLLM Scaling. As shown in Tab. 1 and Tab. 11, Tarot-SAM3 (7B) already surpasses 72B EVOL-SAM3 and SAM3 Agent on RefCOCO testA (**+1.9/+4.2** cIoU) and ReasonSeg (**+4.7/+6.3** and **+5.0/+7.2** gIoU/cIoU, respectively). Together with *Appendix* Sec. C.4.7, these results show that ERI and MSR reduce reliance on MLLM scaling by mitigating MLLM-side parsing, grounding, and evaluation errors while refining residual SAM3 mask errors.

Effectiveness of DINOv3. Using RefCOCO as a representative example, Fig. 10 and Tab. 13 show that DINOv3 consistently improves both ERI and MSR across all splits. These improvements validate the effectiveness of DINOv3 feature coherence for reliable point generation and discrepancy-aware mask diagnosis and refinement. *Qualitative results are provided in Appendix Fig. 16.*

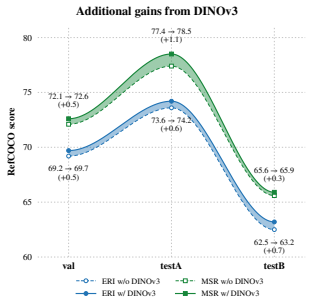

Figure 10: DINOv3 ablation in ERI and MSR.

Effectiveness of Consensus-Based Selection. As shown in Fig. 8, we replace direct MLLM mask comparison with consensus-based selection to mitigate position bias. Intra-Branch Consensus, which selects the best mask using only candidates within the same branch, improves BBox/Text selection by **+2.5/+4.9** and **+3.8/+6.1** gIoU/cIoU, respectively. Cross-Branch Consensus further yields **+2.0/+3.0** and **+9.8/+11.1** gIoU/cIoU gains for BBox and Text selection, while IPS gains **+4.1/+4.3**.

Failure cases. Tarot-SAM3 still remains limited by spatial ambiguity and target-granularity confusion, which may cause incorrect reference grounding or whole-object segmentation instead of the intended local region. *Appendix Fig. 17 illustrates both representative cases.*

5 Conclusion and Future Work

In this paper, we presented Tarot-SAM3, a training-free framework built on dual-stage refinement of MLLM-derived grounding cues and SAM3 mask predictions. The Expression Reasoning Interpreter (ERI) calibrated unreliable grounding before SAM3 inference through structured expression reasoning and heterogeneous prompt construction, while Mask Self-Refining (MSR) exploited feature coherence and object consistency to diagnose and correct over- and under-segmentation. Extensive experiments demonstrated strong performance across multiple RES benchmarks, with ablations validating the effectiveness of each component.

Future Work. Future work will focus on improving inference efficiency and further strengthening ERI and MSR for enhanced robustness and generalization. Extending Tarot-SAM3 from image-based RES to video tasks, particularly referring object tracking, is another promising direction.

AI USE STATEMENT

In this work, we used generative AI tools solely to improve the grammar, wording, and readability of the manuscript. All AI-assisted edits were reviewed and verified by the authors. We take full responsibility for the final content of this work.

ETHICS STATEMENT

This work uses publicly available datasets. Human evaluation was conducted to assess the quality of generated prompts. No personal or sensitive information was collected. We are not aware of any additional ethical concerns arising from this work.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide detailed descriptions of the Tarot-SAM3 pipeline, including the ERI and MSR phases, prompt designs, mask-selection procedures, and inference settings in the main paper and supplementary material. We also report the model backbones, datasets, evaluation protocols, and key hyperparameters used in all experiments. Additional implementation details and ablation analyses are provided in the appendix. To further facilitate reproducibility, the source code and implementation details will be made publicly available upon acceptance.

REFERENCES

Xiang An, Kaicheng Yang, Xiangzi Dai, Ziyong Feng, and Jiankang Deng. Multi-label cluster discrimination for visual representation learning. In *European Conference on Computer Vision*, pp. 428–444. 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. *arXiv preprint arXiv:2511.16719*, 2025.

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

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*, 45(6):7900–7916, 2022.

Danny Driess, F. Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Ho Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Peter R. Florence. Palm-e: An embodied multimodal language model. In *International Conference on Machine Learning*, 2023.

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.

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

Yutao Hu, Qixiong Wang, Wenqi Shao, Enze Xie, Zhenguo Li, Jungong Han, and Ping Luo. Beyond one-to-one: Rethinking the referring image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4067–4077, October 2023.

Zhiwei Hu, Guang Feng, Jiayu Sun, Lihe Zhang, and Huchuan Lu. Bi-directional relationship inferring network for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4424–4433, 2020.

Jiaqi Huang, Zunnan Xu, Ting Liu, Yong Liu, Haonan Han, Kehong Yuan, and Xiu Li. Densely connected parameter-efficient tuning for referring image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3653–3661, 2025.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10488–10497, 2020.

Chongcong Jiang, Tianxingjian Ding, Chuhan Song, Jiachen Tu, Ziyang Yan, Yihua Shao, Zhenyi Wang, Yuzhang Shang, Tianyu Han, and Yu Tian. Medical sam3: A foundation model for universal prompt-driven medical image segmentation. *ArXiv*, abs/2601.10880, 2026.

Ya Jing, Tao Kong, Wei Wang, Liang Wang, Lei Li, and Tieniu Tan. Locate then segment: A strong pipeline for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9858–9867, 2021.

Namyup Kim, Dongwon Kim, Cuiling Lan, Wenjun Zeng, and Suha Kwak. Restr: Convolution-free referring image segmentation using transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18145–18154, 2022.

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

Jiachen Li, Qing Xie, Renshu Gu, Jinyu Xu, Yongjian Liu, and Xiaohan Yu. Lgd: Leveraging generative descriptions for zero-shot referring image segmentation. *Pattern Recognition*, pp. 112549, 2025a.

Jingjing Li, Yue Feng, Yuchen Guo, Jincai Huang, Wei Ji, Qi Bi, Yongri Piao, Miao Zhang, Xiaoqi Zhao, Qiang Chen, et al. Sam3-i: Segment anything with instructions. *arXiv preprint arXiv:2512.04585*, 2025b.

Ruiyu Li, Kaican Li, Yi-Chun Kuo, Michelle Shu, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Referring image segmentation via recurrent refinement networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5745–5753, 2018.

Chen Liang, Wenguan Wang, Tianfei Zhou, Jiaxu Miao, Yawei Luo, and Yi Yang. Local-global context aware transformer for language-guided video segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10055–10069, 2023a.

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, 2023b.

Weihuang Lin, Yiwei Ma, Xiaoshuai Sun, Shuting He, Jiayi Ji, Liujuan Cao, and Rongrong Ji. Hrseg: High-resolution visual perception and enhancement for reasoning segmentation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 3202–3211, 2025.

Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan Yuille. Recurrent multimodal interaction for referring image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pp. 1271–1280, 2017.

Si Liu, Tianrui Hui, Shaofei Huang, Yunchao Wei, Bo Li, and Guanbin Li. Cross-modal progressive comprehension for referring segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4761–4775, 2021.

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.

Yang Liu, Ming Ma, Xiaomin Yu, Pengxiang Ding, Han Zhao, Mingyang Sun, Siteng Huang, and Donglin Wang. Ssr: Enhancing depth perception in vision-language models via rationale-guided spatial reasoning. *arXiv preprint arXiv:2505.12448*, 2025a.

Yong Liu, Cairong Zhang, Yitong Wang, Jiahao Wang, Yujiu Yang, and Yansong Tang. Universal segmentation at arbitrary granularity with language instruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3459–3469, 2024.

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

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, 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.

Minheng Ni, Yabo Zhang, Kailai Feng, Xiaoming Li, Yiwen Guo, and Wangmeng Zuo. Ref-diff: Zero-shot referring image segmentation with generative models. *arXiv preprint arXiv:2308.16777*, 2023.

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.

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

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*, 2024.

Ranjan Sapkota and Manoj Karkee. Object detection with multimodal large vision-language models: An in-depth review. *Available at SSRN 5233953*, 2025.

Oriane Sim'eoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michael Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Herv'e J'egou, Patrick Labatut, and Piotr Bojanowski. Dinov3. 2025.

Jinwang Song, Zhongtian Hua, Hongying Zan, Yingjie Han, and Min Peng. Optimizing discriminative vision-language models for efficient multimodal intent recognition. *Companion Proceedings of the ACM on Web Conference 2025*, 2025.

Shuyang Sun, Runjia Li, Philip Torr, Xiuye Gu, and Siyang Li. Clip as rnn: Segment countless visual concepts without training endeavor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13171–13182, 2024.

Yucheng Suo, Linchao Zhu, and Yi Yang. Text augmented spatial aware zero-shot referring image segmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1032–1043, 2023.

Hao Wang, Limeng Qiao, Zequn Jie, Zhijian Huang, Chengjian Feng, Qingfang Zheng, Lin Ma, Xiangyuan Lan, and Xiaodan Liang. X-sam: From segment anything to any segmentation. *arXiv preprint arXiv:2508.04655*, 2025a.

Junchi Wang and Lei Ke. Llm-seg: Bridging image segmentation and large language model reasoning. *arXiv preprint arXiv:2404.08767*, 2024.

Shijie Wang, Dahun Kim, Ali Taalimi, Chen Sun, and Weicheng Kuo. Learning visual grounding from generative vision and language model. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 8046–8056. IEEE, 2025b.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12998–13008, 2024.

Yuji Wang, Jingchen Ni, Yong Liu, Chun Yuan, and Yansong Tang. Iterprime: Zero-shot referring image segmentation with iterative grad-cam refinement and primary word emphasis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8159–8168, 2025c.

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.

Cong Wei, Yujie Zhong, Haoxian Tan, Yong Liu, Zheng Zhao, Jie Hu, and Yujiu Yang. Hyperseg: Towards universal visual segmentation with large language model. *arXiv preprint arXiv:2411.17606*, 2024.

Junfeng Wu, Yi Jiang, Qihao Liu, Zehuan Yuan, Xiang Bai, and Song Bai. General object foundation model for images and videos at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3783–3795, 2024.

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.

Dingwen Xiao, Weiming Zhang, Shiqi Wen, and Lin Wang. Panosam2: Lightweight distortion- and memory-aware adaptions of sam2 for 360 video object segmentation. *arXiv preprint arXiv:2604.07901*, 2026.

Yin Xie, Kaicheng Yang, Xiang An, Kun Wu, Yongle Zhao, Weimo Deng, Zimin Ran, Yumeng Wang, Ziyong Feng, Roy Miles, et al. Region-based cluster discrimination for visual representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1793–1803, 2025.

Zunnan Xu, Zhihong Chen, Yong Zhang, Yibing Song, Xiang Wan, and Guanbin Li. Bridging vision and language encoders: Parameter-efficient tuning for referring image segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 17503–17512, 2023.

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*, 2023a.

Shu Yang, Yaowei Wang, Ke Chen, Wei Zeng, and Zesong Fei. Attribute-aware feature encoding for object recognition and segmentation. *IEEE Transactions on Multimedia*, 24:3611–3623, 2021.

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.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *ArXiv*, abs/2303.11381, 2023b.

Kai Ye, Xiaotong You, Jianghang Lin, Jiayi Ji, Pingyang Dai, and Liujuan Cao. Evolving, not training: Zero-shot reasoning segmentation via evolutionary prompting. *ArXiv*, abs/2512.24702, 2025.

Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10502–10511, 2019.

Zihao Yi, Jiarui Ouyang, Zhe Xu, Yuwen Liu, Tianhao Liao, Haohao Luo, and Ying Shen. A survey on recent advances in llm-based multi-turn dialogue systems. *ACM Computing Surveys*, 2024.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European conference on computer vision*, pp. 69–85. Springer, 2016.

Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1307–1315, 2018.

Seonghoon Yu, Paul Hongsuck Seo, and Jeany Son. Zero-shot referring image segmentation with global-local context features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19456–19465, 2023.

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, 2024.

Yuxuan Zhang, Tianheng Cheng, Lianghui Zhu, Rui Hu, Lei Liu, Heng Liu, Longjin Ran, Xiaoxin Chen, Wenyu Liu, and Xinggang Wang. Evf-sam: Early vision-language fusion for text-prompted segment anything model. *arXiv preprint arXiv:2406.20076*, 2024a.

Zheng Zhang, Yeyao Ma, Enming Zhang, and Xiang Bai. Psalm: Pixelwise segmentation with large multi-modal model. In *European Conference on Computer Vision*, pp. 74–91. Springer, 2024b.

Lianghui Zhu, Bin Ouyang, Yuxuan Zhang, Tianheng Cheng, Rui Hu, Haocheng Shen, Longjin Ran, Xiaoxin Chen, Li Yu, Wenyu Liu, et al. Lens: Learning to segment anything with unified reinforced reasoning. *arXiv preprint arXiv:2508.14153*, 2025a.

Muzhi Zhu, Yuzhuo Tian, Hao Chen, Chunluan Zhou, Qingpei Guo, Yang Liu, Ming Yang, and Chunhua Shen. Segagent: Exploring pixel understanding capabilities in mllms by imitating human annotator trajectories. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3686–3696, 2025b.

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

Table 7: Prompt types supported by different methods.

Prompt	SAM3 Agent Carion et al. (2025)	EVOL-SAM3 Ye et al. (2025)	☆ **Tarot-SAM3 (ours)**
Text	✓	✓	✓
BBox	×	✓	✓
Point	×	×	✓

A More Details of Motivation

This section further clarifies the motivation of Tarot-SAM3 through comparisons with existing training-free MLLM–SAM3 pipelines and closely related trained MLLM–SAM methods.

A.1 Comparison with Existing Training-Free MLLM–SAM3 Methods.

Existing training-free methods, including SAM3 Agent Carion et al. (2025) and EVOL-SAM3 Ye et al. (2025), coordinate an MLLM with SAM3 to convert referring expressions into segmentation prompts. SAM3 Agent iteratively searches for noun phrases compatible with SAM3, while EVOL-SAM3 further evaluates the parsed descriptions and introduces MLLM-generated bounding boxes when textual prompts remain unreliable. Despite these improvements, two levels of uncertainty remain unresolved. First, these methods ultimately rely on MLLM-derived target concepts or localization cues; erroneous interpretations can therefore propagate to SAM3 and cause missing or inaccurate masks. As illustrated in Fig. 2, the concepts "veranda" and "balcony" fail to identify the intended "wooden footpath," resulting in no valid text-based mask. Beyond grounding errors, both SAM3 Agent and EVOL-SAM3 rely on MLLM judgments for candidate evaluation, making mask selection susceptible to presentation-order bias. Our double-switch diagnostic in Tab. 6 reveals inconsistent selections in **91.6%** of comparisons after swapping the two masks, highlighting the instability of single-order MLLM judgments. Second, even when a plausible prompt is obtained, the resulting SAM3 mask may still exhibit over- or under-segmentation. In the same example, the MLLM-generated bounding box introduces irrelevant fence regions, showing that improved grounding alone does not guarantee reliable mask quality.

Table 6: Order sensitivity and dual-order consistency of MLLM mask selection. M_a denotes the oracle-preferred mask with higher GT-IoU. Single-order results report the probability of selecting M_a at each presentation position.

Protocol	**Position / Outcome**	**Rate (%)**
Single-pass	M_a shown first	22.5
	M_a shown second	75.9
Double-switch judgement	Agree on M_a	5.2
	Agree on M_b	3.1
	Disagree	91.6

Tarot-SAM3 addresses these uncertainties through two complementary phases. The **Expression Reasoning Interpreter (ERI)** calibrates MLLM-derived grounding cues through structured expression reasoning, evaluation-aware rephrasing, and heterogeneous prompt construction. As summarized in Tab. 7, ERI combines augmented text and bounding-box prompts with DINOv3-guided point prompts, enabling more reliable target identification and SAM3 mask generation. ERI further combines intra-prompt consensus with cross-prompt voting to select the best mask for each prompt type (Alg. 1). By grounding selection in agreement among heterogeneous predictions, this design reduces reliance on the order-sensitive MLLM pairwise judgments exposed in Tab. 6. The **Mask Self-Refining (MSR)** phase subsequently diagnoses residual mask discrepancies through DINOv3 feature coherence and MLLM-based region-affiliation reasoning, and converts the diagnosed errors into corrective point prompts for SAM3. Therefore, ERI and MSR form a synergistic process: ERI resolves **target-level uncertainty** by identifying and localizing the correct target, whereas MSR resolves **mask-level uncertainty** by diagnosing and refining the selected SAM3 mask. *To the best of our knowledge, Tarot-SAM3 is the first framework that jointly leverages three prompt types—text, bounding box, and point—together with prompt augmentation, significantly improving prompt robustness and adaptability.* Through this self-refining decision process, Tarot-SAM3 effectively mitigates segmentation errors caused by unreliable single prompt and achieves more precise segmentation results, surpassing the mask quality of prior approaches.

Table 8: Focused comparison with closely related MLLM-SAM methods. ZS denotes zero-shot evaluation.

Method	High-level Principle	RefCOCO Setting	ReasonSeg Setting	ReasonSeg-val		RefCOCO	RefCOCO+	RefCOCOg
				gIoU	cIoU	testA	testA	test(U)
SAM4MLLM Chen et al. (2024)	train MLLM as SAM prompt generator	trained	**ZS**	46.7↑28.8	48.1↑19.7	**80.9**	**76.8**	**75.2**
SAM-VETERAN Du et al. (2026)	train MLLM as interaction/refinement agent	trained	trained	68.2↑7.3	67.3↑0.5	80.8↓0.1	76.6↓0.2	73.4↓1.8
Tarot-SAM3	**training-free closed-loop reasoning/refinement**	**ZS**	**ZS**	**75.5**	**67.8**	78.5↓2.4	72.2↓4.6	71.2↓4.0

Table 9: User study of candidate reasoning steps across junior, average, and senior participant groups. Act. Freq. denotes the proportion of queries activating each step. Δ is the change in the corresponding 10-point rating after its removal, where a larger negative value indicates greater importance. Corr., Conc., and SAM-F. denote correctness, conciseness, and SAM-friendliness, respectively; Keep and Drop indicate the final design decisions.

Candidate Reasoning Step	Act. Freq.	Junior (Δ)Corr.	Average (Δ)Conc.	Senior (Δ)SAM-F.	Decision
1. Explicit/Implicit?	100.0%	**-4.5**	-2.1	-3.2	✓ Keep
2. Single/Multi-Obj.?	100.0%	-3.8	-1.5	-5.1	✓ Keep
3. Refer Object?	38.7%	-4.2	-1.8	-4.5	✓ Keep
4. Adjective Exp.?	65.1%	-3.1	**-2.8**	-2.2	✓ Keep
5. Object-Level?	29.4%	-1.2	-1.5	**-6.8**	✓ Keep
6. Seg. Confusion?	100.0%	-2.5	-1.2	-4.9	✓ Keep
7. Absolute Location?	12.4%	-0.6	-0.5	-1.3	× Drop
8. Material/Texture?	15.6%	-1.1	-0.4	-1.1	× Drop
9. Action/State?	4.2%	-0.9	-0.5	-1.6	× Drop

A.2 Comparison with Closely Related Trained MLLM–SAM Methods.

SAM4MLLM Chen et al. (2024) and SAM-VETERAN Du et al. (2026) explore technically related directions by improving MLLM-based prompt generation or introducing interaction and refinement between MLLMs and SAM. However, both methods rely on task-specific training: SAM4MLLM trains the MLLM as a SAM prompt generator, while SAM-VETERAN trains it as an interaction and refinement agent. In contrast, Tarot-SAM3 operates under a more restrictive training-free setting, where the MLLM, DINOv3, and SAM3 remain entirely frozen, and no RES-specific adaptation is permitted. Consequently, the framework cannot learn dataset-specific prompt translation or mask-quality estimation, and must instead coordinate the pretrained models through inference-time structured reasoning, feature evidence, and corrective prompting. As shown in Tab. 8, Tarot-SAM3 requires no RES-specific training for either explicit or implicit RES. Although it trails task-specifically trained methods on explicit benchmarks, it substantially outperforms both SAM4MLLM and SAM-VETERAN on ReasonSeg, demonstrating stronger generalization without RES-specific training. *This comparison highlights that our contribution lies not merely in employing additional prompts or refinement, but in establishing reliable closed-loop grounding and mask correction without task-specific training.*

B More Details of Methodology

B.1 More Details about ERI

B.1.1 User study of Reasoning-assisted Prompt Options

To rigorously determine the optimal reasoning chain for our expression reasoning interpreter, we conducted a user study to filter a candidate pool of 9 potential reasoning steps down to the most essential 6. We employed a leave-one-out (LOO) ablation strategy: for each test sample, we generated a baseline refined object name using the full 9-step pipeline, and then generated ablated versions by omitting one reasoning step at a time.

Following a similar evaluation framework to previous studies, we invited three groups of human evaluators with varying levels of expertise to assess the quality drop (Δ) of the refined names on a 10-point scale:

- **Junior Testers (Non-CS background):** Evaluated *Correctness*. A larger score drop indicates the generated prompt failed to semantically match the target object in the image when the step was removed.
- **Average Testers (CS Researchers):** Evaluated *Conciseness*. A larger score drop indicates the resulting prompt became overly verbose or lost visual descriptiveness.
- **Senior Testers (Segmentation/SAM Experts):** Evaluated *SAM-Friendliness*. A larger score drop indicates the generated prompt became sub-optimal for SAM3's input space (e.g., failing to specify part vs. whole).

The results, including the activation frequency of each step across the ReasonSeg subset, are detailed in Tab. 9. As illustrated in Tab. 9, the first 6 core steps exhibit significant importance across different evaluation metrics. For instance, removing the *Explicit/Implicit* step heavily impacts basic target correctness (Junior $\Delta = -4.5$), while omitting *Object-Level* reasoning causes a severe drop in SAM-friendliness (Senior $\Delta = -6.8$), as the model loses the ability to distinguish between parts and wholes.

Conversely, candidate steps 7 through 9 (*Absolute Location, Material/Texture, Action/State*) demonstrated low activation frequencies ($\leq 15.6\%$) and negligible performance drops upon removal. This suggests that these dimensions are either overly sparse in referring image segmentation tasks or their semantic roles are inherently covered by other steps (e.g., *Adjective Exp.* often encapsulates material properties). Consequently, these three steps were discarded, finalizing our robust 6-step multi-turn reasoning chain.

B.1.2 HPE-Based Text Augmentation and Multi-Granularity Expression Rephrasing

Alg. 2 details the first HPE procedure, whose principal inputs, N_t and P_R^a, are obtained from the preceding target/refer-object extraction stage. HPE retains the extracted target name as P_T^a and generates alternative semantic or functional names conditioned on the target-centric description T_D and the refer object P_R^a. Candidates that introduce ambiguity or change the target identity are discarded, and two complementary candidates are selected as P_T^b and P_T^c. HPE then localizes the refer object and identifies its relation to the target, together with the target's discriminative attributes. The shorter expression T_S preserves the explicit target name and the most decisive relational clue while removing redundant clauses. In contrast, the longer expression T_L combines target attributes, spatial or functional relations, and refer-object context to support more precise localization. Finally, consistency checks ensure that both expressions preserve the original target identity and contain only information supported by T_D.

B.1.3 More Details about Intra-Prompt-Type Optimal Selection.

Alg. 3 details how ERI selects one representative mask from each prompt type without additional MLLM pairwise comparisons. First, candidates within the Text and BBox branches are evaluated by intra-branch consensus agreement. For each mask, a_p records its best IoU match under prompt

Algorithm 2 HPE: Text Augmentation and Multi-Granularity Expression Rephrasing

Input Image X; target-centric description T_D; target object name N_t; refer-object prompt P_R^a

Output Augmented text prompts P_T^a, P_T^b, P_T^c; shorter expression T_S; longer expression T_L

$P_T^a \leftarrow N_t$ ▷ Preserve the original target name

$\mathcal{C}_T \leftarrow \mathrm{NameAug}_{\mathrm{MLLM}}(N_t, T_D, P_R^a)$ ▷ Generate semantic and functional alternatives

$\mathcal{C}_T \leftarrow \mathrm{TargetFilter}(\mathcal{C}_T, N_t, T_D)$ ▷ Remove ambiguous or target-shifting candidates

$(P_T^b, P_T^c) \leftarrow \mathrm{DiverseBestTwo}(\mathcal{C}_T, N_t)$ ▷ Select complementary target names

$B_R^a \leftarrow \mathrm{Ground}_{\mathrm{MLLM}}(X, P_R^a)$ ▷ Localize the refer object

$(R_T^a, D_T^a) \leftarrow \mathrm{Relation}_{\mathrm{MLLM}}(X, T_D, N_t, P_R^a, B_R^a)$ ▷ Extract the relation and discriminative cue

$A_T \leftarrow \mathrm{Attribute}_{\mathrm{MLLM}}(T_D, N_t)$ ▷ Collect target-specific visual attributes

$C_S \leftarrow \mathrm{Core}(N_t, R_T^a, P_R^a)$ ▷ Keep the most decisive relational clue

$T_S \leftarrow \mathrm{Shorten}_{\mathrm{MLLM}}(C_S)$ ▷ Remove redundant and non-discriminative clauses

$C_L \leftarrow \mathrm{Compose}(N_t, A_T, R_T^a, D_T^a, P_R^a)$ ▷ Combine attributes and relational context

$T_L \leftarrow \mathrm{Elaborate}_{\mathrm{MLLM}}(C_L, T_D)$ ▷ Generate a context-rich expression

if $\mathrm{Target}(T_S) \neq N_t$ **then**

    $T_S \leftarrow \mathrm{CanonicalShort}(N_t, R_T^a, P_R^a)$

end if

if $\mathrm{Target}(T_L) \neq N_t$ **or** $\neg\,\mathrm{Supported}(T_L, T_D)$ **then**

    $T_L \leftarrow \mathrm{CanonicalLong}(N_t, A_T, R_T^a, P_R^a)$

end if

return $(P_T^a, P_T^b, P_T^c, T_S, T_L)$

route p; h counts the routes that support it above τ_{con}, while r measures its mean agreement. Lexicographic ranking by (h, r, q) prioritizes broadly supported candidates before using SAM3 confidence for tie-breaking, and DIVERSETOP3 removes redundant hypotheses. For BBox selection, every filtered text mask contributes reliability-weighted overlap support. Only BBox candidates whose support is no lower than that of the full-description mask B_F remain eligible. Among these candidates, M^O_{BBX} is selected by maximizing the additive score $S(B) + \frac{1}{2}q(B)$. For Text selection, the Top-3 text masks form consensus clusters. Each Top-3 BBox mask votes only when it uniquely matches one cluster above τ_{vote}; its vote is weighted by confidence, stability, branch agreement, and the margin over the runner-up. MAJORITY combines intra-cluster evidence, cross-branch votes, and confidence, whereas GUARDEDUPDATE changes the initial hypothesis only when the collected evidence is sufficiently consistent. The resulting masks define ERI's high-confidence overlap region and seed the subsequent SAM3 point-prompt refinement.

B.1.4 MORE DETAILS ABOUT POINT MASK EXTRACTION.

As in Alg. 4, ERI utilizes the overlapping region between M^O_{Text} and M^O_{BBX} to generate point prompts. If either side is an empty set, the other side is considered the intersection. Concretely, ERI first samples n candidate points from the intersection region of these two masks. Then, leveraging the pixel-level relational modeling capability of DINOv3, we compute the pixel-wise harmonic mean of the n DINO feature similarity maps. The position with the highest similarity value is selected as the positive point P_{pos}, while the position corresponding to the maximum gradient

Algorithm 3 Detailed Intra-Prompt-Type Optimal Selection for ERI

Input Route-specific filtered masks $\{\mathcal{M}_{t,p}\}_{t\in\{\text{Text},\text{BBX}\},p\in\mathcal{P}_t}$; SAM3 confidence $q(\cdot)$; mask stability $s(\cdot)$; thresholds τ_{con} and τ_{vote}; full-description BBox mask $B_F \in \mathcal{M}_{\text{BBX}}$
Output Optimal masks M^O_{Text} and M^O_{BBX}

Stage 1: Intra-branch consensus and diverse Top-3 selection
for $t \in \{\text{Text}, \text{BBX}\}$ **do**
  $\mathcal{M}_t \leftarrow \bigcup_{p\in\mathcal{P}_t} \mathcal{M}_{t,p}$
  for all $M \in \mathcal{M}_t$ **do**
    for all $p \in \mathcal{P}_t$ **do**
      $a_p(M) \leftarrow \max_{N\in\mathcal{M}_{t,p}} \text{IoU}(M, N)$ ▷ Best agreement with route p
    end for
    $h(M) \leftarrow \sum_{p\in\mathcal{P}_t} \mathbf{1}[a_p(M) \geq \tau_{\text{con}}]$ ▷ Number of supporting routes
    $r(M) \leftarrow \frac{1}{|\mathcal{P}_t|} \sum_{p\in\mathcal{P}_t} a_p(M)$ ▷ Mean cross-route agreement
  end for
  $\mathcal{L}_t \leftarrow \text{Sort}^{\text{lex}}_{\downarrow}(\mathcal{M}_t; h, r, q)$ ▷ Prioritize support, agreement, then confidence
  $\mathcal{M}^{(3)}_t \leftarrow \textsc{DiverseTop3}(\mathcal{L}_t, \tau_{\text{con}})$ ▷ Retain up to three non-redundant candidates
end for

Stage 2: Text-to-BBox reliability-weighted voting
$\{w_M\}_{M\in\mathcal{M}_{\text{Text}}} \leftarrow \textsc{Reliability}(\mathcal{M}_{\text{Text}}, \mathcal{M}_{\text{BBX}}, q)$
$Z_{\text{Text}} \leftarrow \sum_{M\in\mathcal{M}_{\text{Text}}} w_M$ ▷ Normalization factor
for all $B \in \mathcal{M}_{\text{BBX}}$ **do**
  $S(B) \leftarrow \frac{1}{Z_{\text{Text}}} \sum_{M\in\mathcal{M}_{\text{Text}}} w_M \ \text{IoU}(B, M)$ ▷ Support received from all Text masks
end for
$\mathcal{W} \leftarrow \{B \in \mathcal{M}_{\text{BBX}} \mid S(B) \geq S(B_F)\}$ ▷ Guard against degrading the full-description baseline
$M^O_{\text{BBX}} \leftarrow \arg\max_{B\in\mathcal{W}} \left[S(B) + \frac{1}{2}q(B)\right]$ ▷ Break support ties using SAM3 confidence

Stage 3: BBox-to-Text margin-aware voting
$\langle \mathcal{G}, \mathbf{I}, \mathbf{Q}\rangle \leftarrow \textsc{GroupStats}(\mathcal{M}^{(3)}_{\text{Text}}, \tau_{\text{con}})$ ▷ $\mathbf{I}$ and $\mathbf{Q}$ store cluster support and confidence
$\mathbf{V} \leftarrow \mathbf{0}$ ▷ Initialize cross-branch votes
for all $B \in \mathcal{M}^{(3)}_{\text{BBX}}$ **do**
  $w_B \leftarrow q(B)s(B)r(B)$ ▷ Reliability of the BBox voter
  for all $G_g \in \mathcal{G}$ **do**
    $a_g(B) \leftarrow \max_{M\in G_g} \text{IoU}(B, M)$ ▷ BBox-to-cluster agreement
  end for
  $g_1 \leftarrow \arg\max_g a_g(B)$
  $g_2 \leftarrow \arg\max_{g\neq g_1} a_g(B)$ ▷ Best and runner-up clusters
  if g_1 is unique **and** $a_{g_1}(B) > \tau_{\text{vote}}$ **then**
    $\Delta_B \leftarrow a_{g_1}(B) - a_{g_2}(B)$ ▷ Preference margin over the runner-up
    $V_{g_1} \leftarrow V_{g_1} + w_B\Delta_B$ ▷ Cast a reliability- and margin-weighted vote
  end if
end for
$M_0 \leftarrow \textsc{Majority}(\mathcal{G}, \mathbf{I}, \mathbf{V}, \mathbf{Q})$ ▷ Fuse intra-branch and cross-branch evidence
$\mathcal{E} \leftarrow \{\mathbf{I}, \mathbf{V}, \mathbf{Q}, h, r, q\}$ ▷ Evidence used by the conservative update
$M^O_{\text{Text}} \leftarrow \textsc{GuardedUpdate}(M_0, \mathcal{M}^{(3)}_{\text{Text}}, \mathcal{E})$ ▷ Update only when evidence is sufficiently consistent
return $\left(M^O_{\text{Text}}, M^O_{\text{BBX}}\right)$

Algorithm 4 HPE: Point Mask Extraction

Input Best Text Mask M^O_{Text}; Best BBox Mask M^O_{BBX}; number of points n **Output** M^O_{Point}

1: $\mathcal{P} \leftarrow$ Cluster$(M^O_{Text} \cap M^O_{BBX}, n)$

2: $S_D \leftarrow$ pixel-wise harmonic mean of DINO's feature similarity maps of $\mathcal{P}$

3: Obtain P_{pos} as the position with maximum similarity and P_{neg} as position with the maximum gradient descent

4: $M^O_{Point} \leftarrow SAM3(P_{pos}, P_{neg})$

5: **return** M^O_{Point}

descent is chosen as the negative point P_{neg}. Finally, these point prompts are fed into SAM3 to generate M^O_{Point}, which is retained only if its IoU with at least one of M^O_{Text} and M^O_{BBX} is at least 0.8, helping reject gross segmentation errors induced by unreliable point prompts. The three types of prompts provide complementary cues for robust mask generation. Text-based masks (M^O_{Text}) ensure semantic alignment with the referring expression, bounding-box-based masks (M^O_{BBX}) impose strong spatial constraints for stable object localization, and point-based masks (M^O_{Point}) enable fine-grained boundary refinement to correct potential over- or under-segmentation errors. By jointly leveraging these complementary cues with prompt augmentation, Tarot-SAM3 achieves more accurate and robust segmentation than methods that rely on a single prompt modality or limited prompt combinations without augmentation.

B.2 More Details about MSR

To further refine SAM3's segmentation results, we introduce the Mask Self-Refining (MSR) phase, as described in Alg. 5. Given the three masks obtained from different prompt types, namely the text-based mask M^O_{Text}, bounding-box-based mask M^O_{BBX}, and point-based mask M^O_{Point}, we first employ the MLLM to select the most reliable mask among them, denoted as M^O_{Best}.

After obtaining the optimal mask M^O_{Best}, the Mask Self-Refinement (MSR) module further analyzes the segmentation discrepancy between M^O_{Best} and M^O_{Point} using DINOv3, enabling the MLLM to more clearly determine whether under-segmentation or over-segmentation occurs.

Specifically, we first compute the discrepancy regions between the two masks, *i.e.*, the regions exclusively owned by each mask (see Line 2 in Alg. 5). We then obtain the centroid p^{Point}_D of the region R^{Point}_D. Using this centroid as the reference point, we leverage DINOv3 to compute the corresponding feature similarity map S^{Point}_D. MSR then identifies the minimum similarity value within R^{Point}_D, denoted as s^{Point}_D, and uses it as a threshold to extract the regions in M^O_{Best} whose similarity values exceed this threshold. These regions are merged with R^{Point}_D to form an additional candidate region. The candidate region is then queried by the MLLM to determine whether it belongs to the target object. If the region is confirmed to be part of the target, a new positive point p^{New}_P is defined as the centroid of R^{Point}_D, and this point is provided to SAM3 together with the current mask M^O_{Best} to refine the segmentation result.

For the over-segmentation case, MSR instead focuses on the discriminative region R^{Best}_D that is exclusively owned by M^O_{Best}. The centroid p^{Best}_D of this region is first computed and fed into DINOv3 to generate the corresponding similarity map S^{Best}_D. Similarly, we identify the minimum similarity value within R^{Best}_D as the threshold s^{Best}_D, and select the regions in the similarity map whose similarity values exceed this threshold. These regions are further intersected with the area outside the overlapping region between M^O_{Best} and M^O_{Point}, producing the candidate region A^{Best}. If the MLLM judges this region to not belong to the target object, the centroid of this region is used as a new negative point p^{New}_N, which is then fed into SAM3 to suppress the over-segmented region and refine the current mask M^O_{Best}. Through this similarity-guided discrepancy analysis, MSR enables Tarot-SAM3 to effectively detect and correct both under-segmentation and over-segmentation, leading to more accurate and reliable segmentation results.

Algorithm 5 Mask Self-Refinement

Input Masks M^O_{Text}, M^O_{BBX}, M^O_{Point}

Output Refined masks

$M^O_{\text{Best}} \leftarrow \text{MLLM}_{best}(M^O_{\text{Text}}, M^O_{\text{BBX}})$

$R_D^{\text{Best}} \leftarrow M^O_{\text{Best}} - M^O_{\text{Point}}$, $R_D^{\text{Point}} \leftarrow M^O_{\text{Point}} - M^O_{\text{Best}}$

$p_D^{\text{Point}} \leftarrow \text{Centroid}(R_D^{\text{Point}})$

$S_D^{\text{Point}} \leftarrow$ DINO's simi map of p_D^{Point}

$s_D^{\text{Point}} \leftarrow \min(S_D^{\text{Point}}[R_D^{\text{Point}}])$

$A^{Point} \leftarrow R_D^{\text{Point}} \cup (S_D^{\text{Point}}[M^O_{\text{Best}}] > s_D^{\text{Point}})$

if A^{Point} is part of the target judged by MLLM **then**

    New Positive Point $p_P^{New} \leftarrow \text{Centroid}(R_D^{\text{Point}})$

    return $M^O_{\text{Best}} \leftarrow SAM3(p_P^{New}, M^O_{\text{Best}})$

end if

$p_D^{\text{Best}} \leftarrow \text{Centroid}(R_D^{\text{Best}})$

$S_D^{\text{Best}} \leftarrow$ DINO's simi map of p_D^{Best}

$s_D^{\text{Best}} \leftarrow \min(S_D^{\text{Best}}[R_D^{\text{Best}}])$

$A^{Best} \leftarrow (\overline{M^O_{\text{Point}} \cap M^O_{\text{Best}}}) \cap (S_D^{\text{Best}} > s_D^{\text{Best}})$

if A^{Best} is not part of the target judged by MLLM **then**

    New Negative Point $p_N^{New} \leftarrow \text{Centroid}(A^{\text{Best}})$

    return $M^O_{\text{Best}} \leftarrow SAM3(p_N^{New}, M^O_{\text{Best}})$

end if

C More Details of Experiment

C.1 More Details of Datasets and Implementation Details

C.1.1 Datasets.

We evaluate Tarot-SAM3 on four referring expression segmentation benchmarks: RefCOCO Yu et al. (2016), RefCOCO+ Yu et al. (2016), and RefCOCOg Mao et al. (2016) for explicit text-guided segmentation, and LISA Lai et al. (2024) for implicit reasoning segmentation. The RefCOCO series extends MS COCO with natural-language referring expressions, forming image–expression–mask triplets. RefCOCO mainly contains concise, location-oriented expressions; RefCOCO+ places greater emphasis on appearance and attributes with fewer spatial cues; and RefCOCOg provides longer, linguistically more complex descriptions. RefCOCO and RefCOCO+ contain val, testA, and testB splits, whereas RefCOCOg is divided into val and test splits. LISA contains 1,218 image–instruction–mask triplets whose expressions require commonsense or logical reasoning beyond direct visual grounding. It is divided into val and test splits, with the test set further categorized into Short non-sentence queries and Long interrogative expressions.

C.1.2 Evaluation Metrics.

Following prior RES protocols, we report global intersection-over-union (gIoU) and cumulative intersection-over-union (cIoU). gIoU averages the sample-wise IoU scores, whereas cIoU is com-

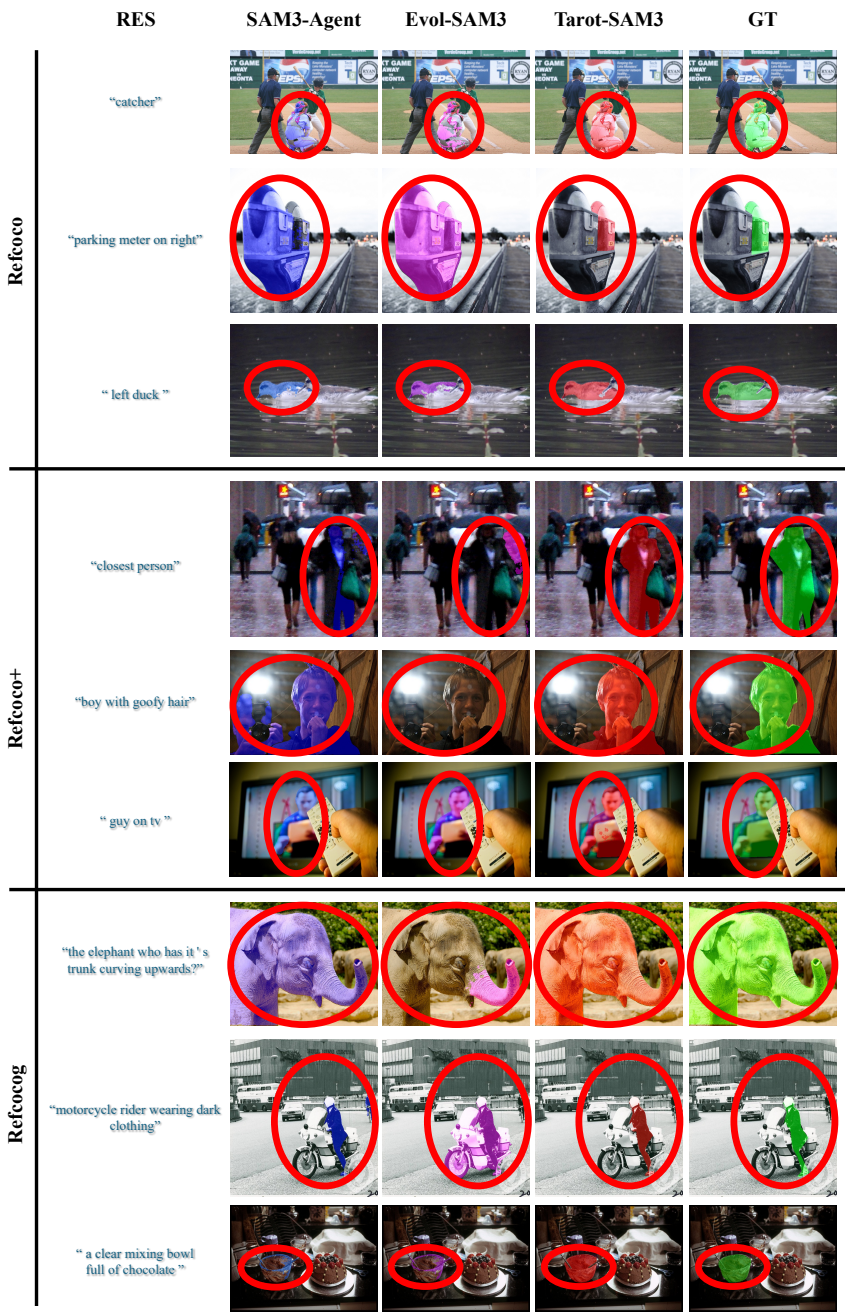

Figure 11: Additional visual comparisons on explicit RES datasets (RefCOCO, RefCOCO+, and RefCOCOg). *Zoom in for a better view.*

puted by dividing the total intersection by the total union over the entire evaluation set. All results are reported as percentages.

C.1.3 Additional Implementation Details.

The multiple DINOv3 similarity maps are aggregated using the harmonic mean. During MSR, mask refinement is performed iteratively for at most three rounds.

Table 10: Full comparison with prior methods on the explicit RES benchmarks RefCOCO, RefCOCO+, and RefCOCOg, covering all official validation and test splits. Results are reported in cIoU and organized by VLM backbone and training setting. **Bold** marks the overall best score, underlining marks the best zero-shot score, and * denotes results obtained from our evaluation, with with results of other methods from Evol-SAM3 Ye et al. (2025).

Method	VLM Backbone	Training	RefCOCO			RefCOCO+			RefCOCOg		
		RES	val	testA	testB	val	testA	testB	val (U)	test (U)	val (G)
LISA Lai et al. (2024)	LLaVA 7B	✓	74.9	79.1	72.3	65.1	70.8	58.1	67.9	70.6	–
GSVA Xia et al. (2024)	13B	✓	79.2	81.7	77.1	70.3	73.8	63.6	75.7	77.0	–
GLaMM Rasheed et al. (2024)	Vicuna 7B	✓	79.5	83.2	76.9	72.6	78.7	64.6	74.2	74.9	–
SAM4MLLM Chen et al. (2024)	LLaVA1.6 7B	✓	79.6	82.8	76.1	73.5	77.8	65.8	74.5	75.6	–
SAM4MLLM Chen et al. (2024)	LLaVA1.6 8B	✓	79.8	82.7	74.7	74.6	80.0	67.2	75.5	76.4	–
GLEE Wu et al. (2024)	Plus	✓	79.5	–	–	68.3	–	–	70.6	–	–
GLEE Wu et al. (2024)	Pro	✓	80.0	–	–	69.6	–	–	72.9	–	–
DETRIS Huang et al. (2025)	DETRIS-L	✓	81.0	81.9	79.0	75.2	78.6	70.2	74.6	75.3	–
UniLSeg Liu et al. (2024)	UniLSeg-20	✓	80.5	81.8	78.4	72.7	77.0	67.0	78.4	79.5	–
UniLSeg Liu et al. (2024)	UniLSeg-100	✓	81.7	83.2	79.9	73.2	78.3	68.2	79.3	80.5	–
PSALM Zhang et al. (2024b)	Phi1.5 1.3B	✓	83.6	84.7	81.6	72.9	75.5	70.1	73.8	74.4	–
EVF-SAM Zhang et al. (2024a)	RC	✓	82.1	83.7	80.0	75.2	78.3	70.1	76.8	77.4	–
EVF-SAM Zhang et al. (2024a)	Extra Data	✓	82.4	84.2	80.2	76.5	80.0	71.9	78.2	78.3	–
RICE Xie et al. (2025)	Qwen2.5-7B	✓	83.5	85.3	81.7	**79.4**	82.8	75.4	79.8	80.4	–
MLCD-seg An et al. (2024)	Qwen2.5-7B	✓	83.6	85.3	81.5	**79.4**	82.9	**75.6**	79.7	80.5	–
HyperSeg Wei et al. (2024)	Phi2 2.7B	✓	84.8	85.7	**83.4**	79.0	**83.5**	75.2	79.4	78.9	–
LENS Zhu et al. (2025a)	Qwen2.5-VL 3B	✓	84.2	85.3	81.0	**79.4**	82.8	74.3	81.2	81.0	–
SAM-Veteran Du et al. (2026)	Qwen2.5-VL 7B	✓	–	80.8	–	–	76.6	–	–	73.4	–
SAM-Veteran Du et al. (2026)	Qwen2.5-VL 32B	✓	–	80.4	–	–	77.4	–	–	73.4	–
READ Qian et al. (2025)	LLaVA1.5 7B	✓	78.1	80.2	73.2	68.4	73.7	60.4	70.1	71.4	–
SegAgent Zhu et al. (2025b)	Qwen-VL 7B	✓	79.2	81.4	75.7	71.5	76.7	65.4	74.8	74.6	–
X-SAM Wang et al. (2025a)	Phi3 3.8B	✓	**85.1**	**87.1**	**83.4**	78.0	81.0	74.4	**83.8**	**83.9**	–
GL-CLIP Yu et al. (2023)	ResNet-50	×	32.7	35.3	30.1	37.7	40.7	34.9	41.6	42.9	44.0
GL-CLIP Yu et al. (2023)	ViT-B/32	×	32.9	34.9	30.1	38.4	42.1	32.7	42.0	42.0	42.7
CaR Sun et al. (2024)	ViT-B/16	×	33.6	35.4	30.5	34.2	36.0	31.0	36.7	36.6	36.6
Ref-Diff Ni et al. (2023)	VAE	×	37.2	38.4	37.2	37.3	40.5	33.0	44.0	44.5	44.3
TAS Suo et al. (2023)	ResNet-50	×	39.9	42.9	35.9	44.0	50.6	36.4	47.7	47.4	48.7
TAS Suo et al. (2023)	ViT-B/32	×	39.8	41.1	36.2	43.6	49.1	36.5	46.6	46.8	48.1
IterRPrimeE Wang et al. (2025c)	–	×	40.2	46.5	33.9	44.2	51.6	35.3	46.0	45.1	45.8
Pseudo-RIS Yu et al. (2024)	CRIS	×	39.8	44.8	33.0	42.2	46.3	34.5	43.7	43.4	43.8
Pseudo-RIS Yu et al. (2024)	ETRIS	×	41.1	48.2	33.5	44.3	51.4	35.1	46.0	46.7	46.8
LGD+DINO Li et al. (2025a)	ViT-B/32	×	49.5	54.7	41.0	49.6	58.4	38.6	50.3	51.1	52.5
VLM-VG Wang et al. (2025b)	ResNet-50	×	47.7	51.8	44.7	41.2	45.9	34.7	46.6	47.1	–
VLM-VG Wang et al. (2025b)	ResNet-101	×	49.9	53.1	46.7	42.7	47.3	36.2	48.0	48.5	–
HybridGL Liu & Li (2025)	ViT-B/32	×	49.5	53.4	45.2	43.4	49.1	37.2	51.3	51.6	–
SAM 3 Agent Carion et al. (2025)	Qwen2.5-VL 7B	×	59.4	64.3	55.0	51.4	57.0	44.9	57.2	58.8	59.7
SAM 3 Agent Carion et al. (2025)	Qwen2.5-VL 72B	×	70.5	74.3	64.6	64.9	70.7	56.6	69.2	69.5	69.8
EVOL-SAM3Ye et al. (2025)	Qwen2.5-VL 3B	×	66.8	70.9	59.9	56.1	63.1	49.3	63.0	63.5	64.0
EVOL-SAM3Ye et al. (2025)	Qwen2.5-VL 7B	×	68.7	73.7	64.4	64.4	67.8	54.0	64.7	65.5	65.9
EVOL-SAM3*Ye et al. (2025)	Qwen2.5-VL 72B	×	71.2	76.6	66.2	67.6	73.8	58.5	70.1	72.2	70.6
Tarot-SAM3 (ours)	Qwen2.5-VL 3B	×	67.3	71.4	58.7	53.6	62.7	48.1	64.5	66.5	64.7
Tarot-SAM3 (ours)	Qwen2.5-VL 7B	×	72.6	78.5	65.9	65.4	72.2	55.3	66.9	71.2	67.1
Tarot-SAM3 (ours)	Qwen2.5-VL 72B	×	75.5	79.5	66.7	68.0	75.1	56.7	71.2	74.5	**72.1**

Table 11: Full comparison with prior methods on ReasonSeg, reporting gIoU and cIoU on the validation set, full test set, and short- and long-query test subsets. Training Data indicates whether RES annotations are used for fine-tuning, thereby distinguishing fine-tuned from zero-shot methods. Results with * were obtained from our evaluation, with results of other methods from Evol-SAM3 Ye et al. (2025).

Method	VLM Backbone	Training Data		Val Set		Test Set		Test (Short)		Test (Long)	
		RES	ReasonSeg	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU
SEEM Zou et al. (2023)	–	×	×	25.5	21.2	24.3	18.7	20.1	11.5	25.6	20.8
Grounded SAM Ren et al. (2024)	–	×	×	26.0	14.5	21.3	16.4	17.8	10.8	22.4	18.6
OVSeg Liang et al. (2023b)	–	×	×	28.5	18.6	26.1	20.8	18.0	15.5	28.7	22.5
GLaMM Rasheed et al. (2024)	Vicuna 7B	✓	×	47.4	47.2	–	–	–	–	–	–
SAM4MLLM Chen et al. (2024)	Qwen-VL 7B	✓	×	46.7	48.1	–	–	–	–	–	–
SAM4MLLM Chen et al. (2024)	LLaVA1.6 8B	✓	×	58.4	60.4	–	–	–	–	–	–
Seg-Zero Liu et al. (2025b)	Qwen2.5-VL 3B	✓	×	58.2	53.1	56.1	48.6	–	–	–	–
Seg-Zero Liu et al. (2025b)	Qwen2.5-VL 7B	✓	×	62.6	62.0	57.5	52.0	–	–	–	–
X-SAM Wang et al. (2025a)	Phi3 3.8B	✓	✓	56.6	32.9	57.8	41.0	47.7	48.1	56.0	40.8
HyperSeg Wei et al. (2024)	Phi2 3B	✓	✓	59.2	56.7	–	–	–	–	–	–
Ref-Diff Ni et al. (2023)	LLaVA1.5 7B	×	×	–	52.4	–	48.7	–	48.0	–	49.1
Ref-Diff Ni et al. (2023)	LLaVA1.5 13B	×	×	–	60.5	–	49.9	–	48.7	–	51.0
LISA Lai et al. (2024)	LLaVA 7B	✓	×	44.4	46.0	36.8	34.1	37.6	34.4	36.6	34.7
LISA Lai et al. (2024)	LLaVA 7B	✓	✓	52.9	54.0	47.3	34.1	40.6	40.6	49.4	51.0
LISA Lai et al. (2024)	LLaVA 13B	✓	×	48.9	46.9	44.8	45.8	39.9	43.3	46.4	46.5
LISA Lai et al. (2024)	LLaVA 13B	✓	✓	56.2	62.9	51.7	51.1	44.3	42.0	54.0	54.3
LISA Lai et al. (2024)	Llama2 13B	✓	✓	60.0	67.8	51.5	51.3	43.9	45.8	54.0	53.8
LISA Lai et al. (2024)	LLaVA1.5 7B	✓	×	53.6	52.3	48.8	47.1	48.3	48.8	49.2	48.9
LISA Lai et al. (2024)	LLaVA1.5 7B	✓	✓	61.3	62.9	55.6	56.9	48.3	46.3	57.9	59.7
LISA Lai et al. (2024)	LLaVA1.5 13B	×	✓	57.7	60.3	53.8	50.8	50.8	50.0	54.7	50.9
LISA Lai et al. (2024)	LLaVA1.5 13B	✓	✓	65.0	**72.9**	61.3	62.2	55.4	50.6	63.2	65.3
LENS Zhu et al. (2025a)	Qwen2.5-VL-3B	✓	✓	62.1	64.9	57.2	58.0	–	–	–	–
HRSeg Lin et al. (2025)	LLaVA1.1 7B	✓	✓	57.4	54.7	43.4	41.7	54.4	54.5	51.6	50.5
HRSeg Lin et al. (2025)	LLaVA1.6 7B	✓	✓	64.9	63.1	49.5	48.7	59.2	58.9	57.0	57.2
SAM-Veteran Du et al. (2026)	Qwen2.5-VL 7B	✓	✓	68.2	67.3	62.6	56.1	–	–	–	–
SAM-Veteran Du et al. (2026)	Qwen2.5-VL 32B	✓	✓	72.3	70.0	62.9	58.2	–	–	–	–
READ Qian et al. (2025)	LLaVA1.5 7B	✓	✓	59.8	67.6	58.5	58.6	52.6	49.5	60.4	61.0
SegAgent Zhu et al. (2025b)	Qwen2.5-VL 7B	✓	✓	33.0	25.4	33.5	31.3	–	–	–	–
LLM-Seg Wang & Ke (2024)	LLaVA1.1 7B	✓	✓	52.3	47.5	47.9	46.2	41.1	40.2	49.8	49.1
RSVP Lu et al. (2025)	LLaVA1.6 7B	×	×	59.2	56.7	56.9	50.7	47.9	42.0	58.4	53.0
RSVP Lu et al. (2025)	Qwen2-VL 7B	×	×	58.6	48.5	56.1	51.6	48.5	44.3	57.1	53.0
RSVP Lu et al. (2025)	Gemini1.5-Flash	×	×	56.9	49.2	57.1	59.2	47.3	40.2	60.2	65.6
RSVP Lu et al. (2025)	GPT-4o	×	×	64.7	63.1	60.3	60.0	55.4	50.4	61.9	62.5
Gemini Seg	Gemini2.5 Flash	?	?	28.3	13.3	30.6	9.2	16.5	8.0	35.0	9.5
SAM 3 Agent Carion et al. (2025)	Qwen2.5-VL 7B	×	×	62.2	49.1	63.0	53.5	59.4	43.5	64.1	56.2
SAM 3 Agent Carion et al. (2025)	Qwen2.5-VL 72B	×	×	74.6	65.1	70.8	64.0	70.3	55.7	71.0	66.3
SAM 3 Agent Carion et al. (2025)	Llama4 Maverick	×	×	68.5	61.5	67.1	60.9	66.8	**59.4**	67.2	61.3
EVOL-SAM3* Ye et al. (2025)	Qwen2.5-VL 3B	×	×	66.8	62.6	66.1	58.0	63.1	49.1	67.1	60.7
EVOL-SAM3* Ye et al. (2025)	Qwen2.5-VL 7B	×	×	69.3	65.1	70.9	65.0	65.9	53.4	72.6	67.9
EVOL-SAM3* Ye et al. (2025)	Qwen2.5-VL 72B	×	×	73.6	67.2	71.1	64.9	67.9	50.5	72.1	68.7
Tarot-SAM3 (ours)	Qwen2.5-VL 3B	×	×	67.1	55.6	67.0	59.0	64.3	50.7	67.9	61.2
Tarot-SAM3 (ours)	Qwen2.5-VL 7B	×	×	75.5	67.8	75.8	71.2	**71.7**	56.8	77.1	**76.5**
Tarot-SAM3 (ours)	Qwen2.5-VL 72B	×	×	**75.6**	69.6	**76.0**	**71.8**	71.3	54.5	**77.4**	75.8

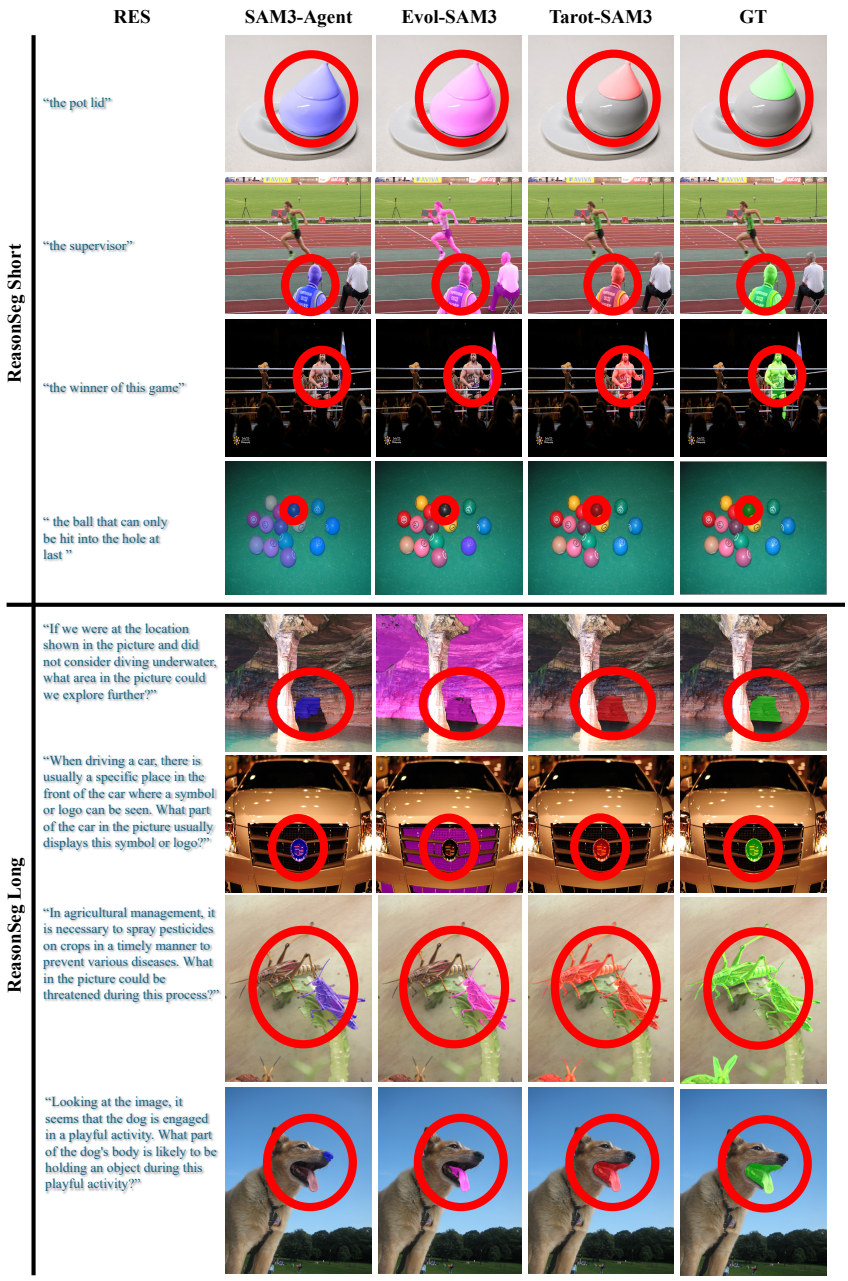

Figure 12: Additional visual comparisons on the implicit RES dataset. *Zoom in for a better view.*

C.2 More Details about Qualitative Comparisons

C.2.1 Detailed Comparisons on Explicit/Implicit Referring Segmentation

Tab. 10 and Tab. 11 highlight the advantages of Tarot-SAM3 over existing training-free MLLM–SAM3 pipelines. With the same Qwen2.5-VL 7B backbone, Tarot-SAM3 outperforms SAM3 Agent and EVOL-SAM3 across all reported RefCOCO-family and ReasonSeg metrics. More notably, our 7B model also surpasses their 72B variants on RefCOCO testA and ReasonSeg test. On RefCOCO testA, Tarot-SAM3 achieves **78.5** cIoU, compared with 74.3 for SAM3 Agent (72B) and 76.6 for EVOL-SAM3 (72B). On ReasonSeg test, it reaches **75.8** gIoU and **71.2** cIoU, exceeding these two 72B baselines by **5.0/7.2** and **4.7/6.3** points, respectively. It also outperforms SAM3 Agent equipped with Llama4 Maverick on both ReasonSeg test metrics.

Table 12: Comparison of MLLM-based and consensus-based mask selection on the RefCOCO fCOCO. testA split. Results are reported as gIoU/cIoU.

Method	BBox Branch	Text Branch	IPS
MLLM	72.4/65.9	62.6/55.8	75.5/73.1
Intra-Consensus	74.9/70.8	66.4/61.9	–
Cross Branch Consensus	**76.9/73.8**	**76.2/73.0**	**79.6/77.4**

Table 13: DINOv3 ablation analysis on RefCOCO.

Phase	DINO	RefCOCO val	RefCOCO A	RefCOCO B
ERI	w/o	69.2	73.6	62.5
	w/	69.7	74.2	63.2
MSR	w/o	72.1	77.4	65.6
	w/	**72.6**	**78.5**	**65.9**

These results support our central design choice of jointly refining MLLM-derived grounding cues and SAM3 mask predictions. ERI improves target grounding through structured reasoning and heterogeneous prompt construction, while MSR converts feature-guided mask diagnosis into corrective prompts. This coordinated design enables a 7B model to outperform substantially larger MLLM-based pipelines on the evaluated tasks. Its effectiveness across both explicit referring segmentation and reasoning-intensive segmentation further supports the transferability of the framework, with all components frozen and no task-specific fine-tuning.

C.2.2 Analysis of MLLM Preference in Mask Comparison

Fig. 2 (b,c) and Tab. 6 reveal a pronounced position bias in MLLM mask comparison. Let M_a denote the oracle-preferred mask, i.e., the mask with higher GT-IoU. In single-order evaluation, the MLLM selects M_a in **22.5%** of comparisons when shown first, but in **75.9%** when shown second. Since the mask pair and its quality ranking are unchanged, this **53.4**-point reversal shows that presentation order can override quality differences. We further test the dual-order (double-switch) protocol on **1,652** distinct mask pairs by querying both orders and accepting only consistent judgments. The predictions agree on M_a in **5.2%** of cases, on M_b in **3.1%**, and conflict in **91.6%**. Thus, order swapping yields a consistent winner in only **8.3%** of cases and fails to neutralize the bias. These results expose direct pairwise comparison as unreliable and motivate our consensus-based strategy, which aggregates support across candidates and prompt branches rather than trusting a single order-sensitive judgment.

C.3 More Details of Visual Comparisons

C.3.1 More visual comparisons on explicit RES datasets.

To further evaluate our method under explicit RES settings, we provide additional qualitative comparisons on explicit RES datasets, including RefCOCO, RefCOCO+, and RefCOCOg, as shown in Fig. 11. The results demonstrate the effectiveness of Tarot-SAM3 in handling challenging referring expressions. In the first row ("catcher"), both SAM3 Agent Carion et al. (2025) and EVOL-SAM3 Ye et al. (2025) directly adopt the segmentation masks produced by SAM3. Although the MLLM correctly interprets the noun phrase and provides the appropriate prompt, the masks generated by SAM3 may still suffer from fragmentation, leading to under-segmentation of the target object. A similar issue can also be observed in the second row from the bottom. In contrast, Tarot-SAM3 effectively alleviates this problem by leveraging the Mask Self-Refining (MSR) phase, *which performs self-refinement on the SAM3 masks and produces more complete object segmentation.*

Moreover, in the second row and the last row, Tarot-SAM3 successfully segments only the correct target object, while avoiding the inclusion of irrelevant objects that would otherwise lead to over-segmentation. Finally, in the fourth row, EVOL-SAM3 fails to produce any valid mask, indicating that both the MLLM-based expression interpretation and the bounding-box prediction fail to identify the target region. *In contrast, Tarot-SAM3 still produces accurate segmentation results, highlighting the effectiveness of the ERI phase and demonstrating the importance of prompt augmentation for both text and bounding box prompts.*

Table 14: Ablation study of different ERI design components.

Design Variant	RefCOCO testA		ReasonSeg test	
	gIoU	cIoU	gIoU	cIoU
Reasoning-Assisted Prompt Options				
No Single/Multi-Object Determination	78.8 ↓1.6	76.1 ↓2.4	64.3 ↓11.5	57.0 ↓14.2
No Object-Level Reasoning	79.6 ↓0.8	77.9 ↓0.6	61.3 ↓14.5	57.9 ↓13.3
No Refer-Object Analysis	76.4 ↓4.0	72.8 ↓5.7	67.8 ↓8.0	62.5 ↓8.7
No Adjective Expression Identification	71.7 ↓8.7	67.2 ↓11.3	65.9 ↓9.9	62.0 ↓9.2
No Segmentor Confusion Awareness	69.4 ↓11.0	65.8 ↓12.7	66.2 ↓9.6	62.8 ↓8.4
Reasoning Steps				
No Augmented Text Prompt	77.2 ↓3.2	73.8 ↓4.7	65.5 ↓10.3	60.3 ↓10.9
No Shorter & Longer Expression	75.7 ↓4.7	70.1 ↓8.4	70.6 ↓5.2	65.8 ↓5.4
No BBox Mask	71.4 ↓9.0	69.1 ↓9.4	73.3 ↓2.5	68.2 ↓3.0
No Text Mask	74.5 ↓5.9	70.4 ↓8.1	47.4 ↓28.4	40.0 ↓31.2
Tarot-SAM3	**80.4**	**78.5**	**75.8**	**71.2**

C.3.2 More visual comparisons on implicit RES datasets.

To further evaluate our method under implicit RES settings, we present additional qualitative comparisons on the ReasonSeg dataset, including both ReasonSeg Short and ReasonSeg Long, as shown in Fig. 12. For ReasonSeg Short, previous approaches often rely on MLLM-based expression parsing, which may lead to misinterpretation of the referring expression or produce prompts that cannot be effectively utilized by SAM3. As illustrated in the first and second examples, both SAM3 Agent and EVOL-SAM3 fail to accurately capture the target object or properly filter irrelevant regions. In the last example of the ReasonSeg Short subset, the MLLM outputs from SAM3 Agent and EVOL-SAM3 demonstrate incorrect semantic understanding, resulting in segmentation errors.

When the referring expressions become longer and more complex, as in ReasonSeg Long, the advantage of Tarot-SAM3 becomes more evident. In the last two examples, Tarot-SAM3 produces more complete and accurate segmentation results, while previous methods either miss parts of the target object or introduce incorrect regions. *This improvement demonstrates the effectiveness of the ERI phase, which enables more accurate interpretation of complex referring expressions, as well as the MSR phase, which further refines the segmentation masks to correct potential errors.* Overall, these results show that Tarot-SAM3 provides more robust and reliable segmentation under implicit RES scenarios, particularly when handling complex or reasoning-intensive referring expressions.

C.3.3 More visual comparisons in open-world scenarios.

To evaluate the generalization ability of Tarot-SAM3 compared with previous training-free methods, we further present additional qualitative comparisons in open-world scenarios, as shown in Fig. 13. Specifically, we consider four representative open-world domains: medical-related scenarios, remote sensing, cityscape scenes, and cartoon/comic imagery. Across the ten examples shown in Fig. 13, Tarot-SAM3 consistently demonstrates superior performance compared with previous approaches. In particular, for the remote sensing and cityscape examples, Tarot-SAM3 successfully identifies and segments the correct target objects while avoiding the introduction of irrelevant regions. In contrast to prior methods, it neither introduces segmentation noise nor misses the target objects described by the referring expressions. Moreover, in the more challenging cartoon and comic scenarios, where visual styles differ significantly from natural images, Tarot-SAM3 still produces accurate and complete segmentation masks. The results show that our method effectively avoids both over-segmentation and under-segmentation, demonstrating strong robustness across diverse visual domains. Overall, these results highlight the strong open-world generalization capability of Tarot-SAM3 and its effectiveness in handling diverse scene types beyond the standard RES benchmarks.

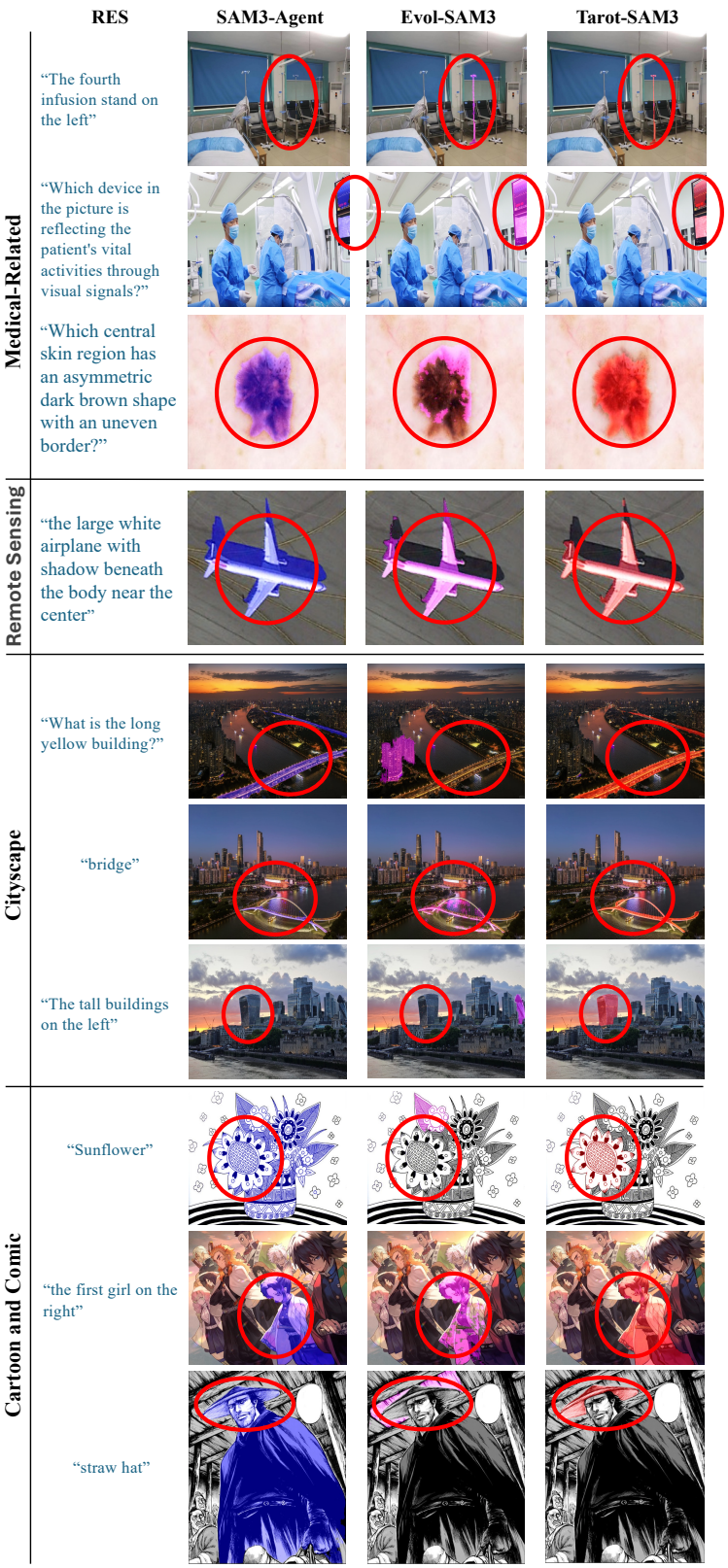

Figure 13: Additional visual comparisons in open-world scenarios. *Zoom in for a better view.*

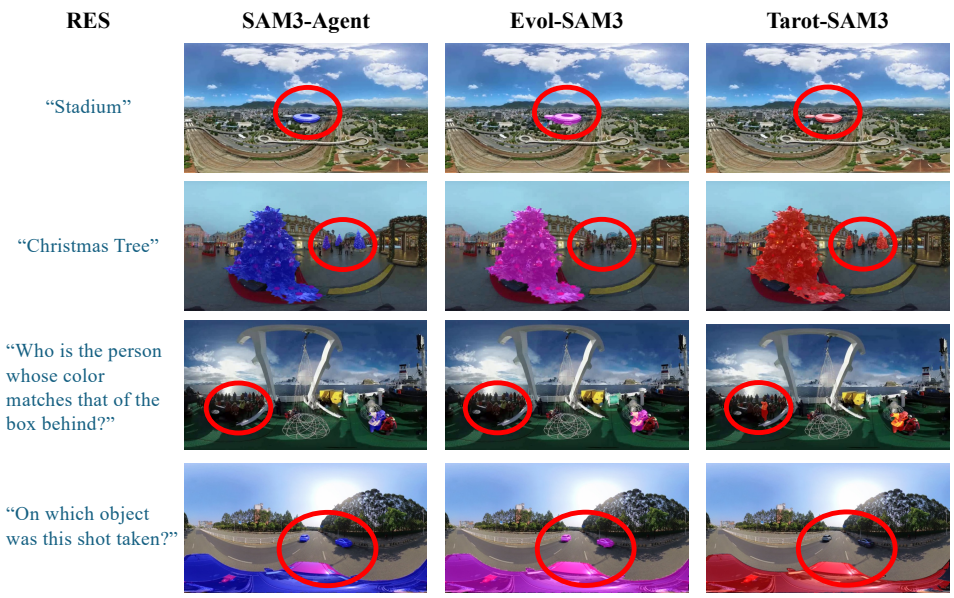

Figure 14: Visual comparisons in 360 scenarios. *Zoom in for a better view.*



Figure 15: Visual comparisons illustrating the effectiveness of the MSR phase.

C.4 More Details of Other Discussions

C.4.1 Comparisons in 360 scenarios.

To further evaluate the generalization capability of Tarot-SAM3, we conduct additional qualitative comparisons in 360° panoramic scenarios, as shown in Fig. 14. Compared with conventional images, 360° scenes introduce additional challenges due to the large field of view (FoV) and severe geometric distortions, which make accurate object localization and segmentation more diffi-

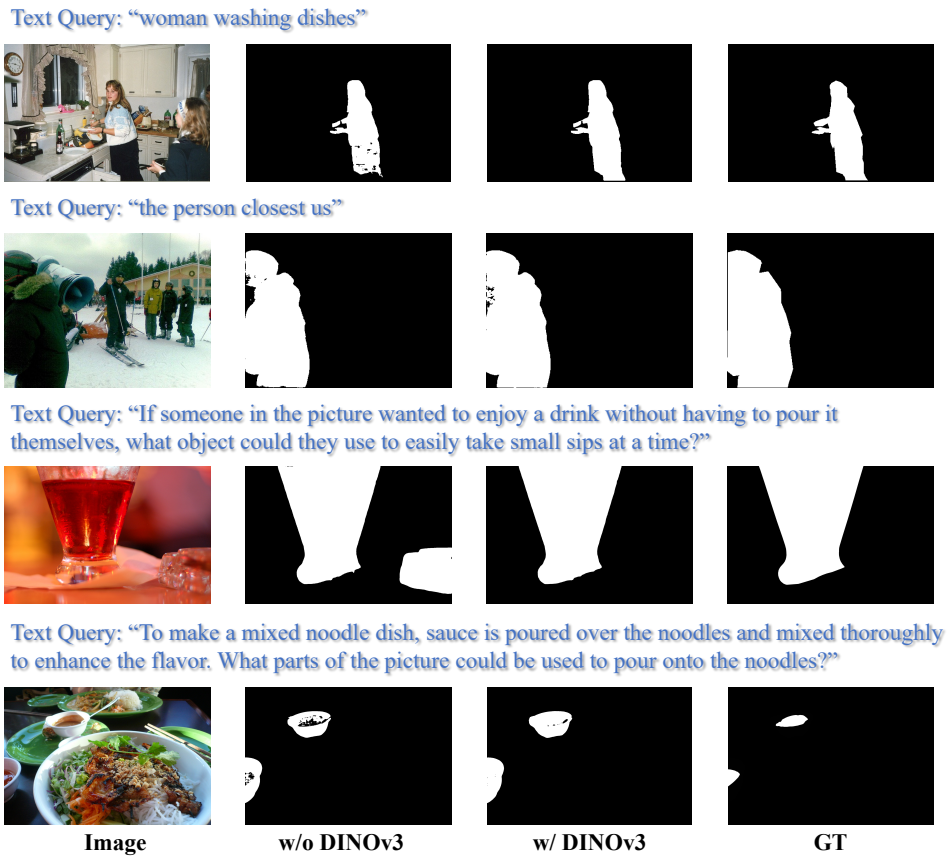

Figure 16: Visual comparisons illustrating the effectiveness of DINOv3.

cult for existing methods. As illustrated in Fig. 14, Tarot-SAM3 consistently outperforms previous training-free methods in panoramic scenes. The results highlight the complementary roles of ERI and MSR: reliable grounding cues help resolve distortion-induced localization ambiguity, while closed-loop mask refinement corrects missing or spurious regions in the initial SAM3 predictions. Consequently, Tarot-SAM3 maintains accurate and complete segmentation despite severe geometric distortion. Overall, these results indicate that Tarot-SAM3 maintains stable and accurate segmentation performance in challenging 360° environments, further validating its strong generalization capability across diverse visual settings.

C.4.2 Analysis of Reasoning-Assisted Prompt Options.

The upper part of Tab. 14 shows that five reasoning-assisted prompt options contribute to reliable expression interpretation. Removing any individual dimension consistently degrades performance on both explicit and implicit RES, confirming that these options provide complementary rather than redundant reasoning cues. Among them, segmentor-confusion awareness causes the largest degradation, reducing RefCOCO testA by **11.0/12.7** gIoU/cIoU and ReasonSeg by **9.6/8.4**. Adjective-expression identification and refer-object analysis are also particularly important on ReasonSeg, where their removal decreases gIoU by **9.9** and **8.0**, respectively. These results indicate that anticipating SAM3-sensitive expressions, resolving object attributes, and identifying relational reference objects are especially critical for reasoning-intensive queries.

C.4.3 Analysis of Reasoning Steps in ERI.

The lower part of Tab. 14 further validates the complementary roles of prompt augmentation and heterogeneous prompt generation. Removing augmented text prompts or short/long rephrasing causes larger degradation on ReasonSeg than on RefCOCO, demonstrating their importance for converting

complex expressions into SAM3-compatible grounding cues. The two prompt modalities exhibit distinct dataset-dependent roles: removing the BBox mask severely reduces RefCOCO testA by **9.0/9.4** gIoU/cIoU, confirming the importance of explicit spatial constraints for concise referring expressions. In contrast, removing the text mask causes the largest degradation on ReasonSeg, with drops of **28.4/31.2**, showing that semantic text cues are indispensable for implicit reasoning. Together, these results demonstrate that text and bounding-box prompts provide complementary semantic and spatial evidence, while prompt augmentation improves their robustness before SAM3 mask generation.

C.4.4 Analysis of Mask Selection Strategies

Tarot-SAM3 generates multiple candidate masks from heterogeneous prompt branches, making reliable selection essential for Inter-Prompt-Type Optimal Selection (IPS). Tab. 12 provides the detailed numerical breakdown of the selection comparison summarized in the main text. Replacing direct MLLM selection with Intra-Branch Consensus improves the BBox branch from **72.4/65.9** to **74.9/70.8** gIoU/cIoU and the Text branch from **62.6/55.8** to **66.4/61.9**. Cross-Branch Consensus further improves both branches and achieves the best results throughout: **76.9/73.8** on BBox, **76.2/73.0** on Text, and **79.6/77.4** after IPS. Relative to direct MLLM selection, these scores correspond to gains of **4.5/7.9**, **13.6/17.2**, and **4.1/4.3** points, respectively. The especially large improvement on the Text branch indicates that cross-branch agreement effectively filters ambiguous or noisy text-derived candidates by exploiting complementary evidence from heterogeneous prompts. Overall, consensus-based aggregation provides a more reliable selection signal than a direct, order-sensitive MLLM judgment, thereby supplying IPS with higher-quality masks for final prediction.

C.4.5 Importance of MSR.

Although ERI substantially improves target grounding and provides strong initial SAM3 masks, accurate grounding does not necessarily guarantee complete and precise segmentation. The mask selected from heterogeneous prompt types may still contain residual mask-level errors, such as missing object parts, fragmented predictions, or extraneous regions. Since MSR operates on top of these already strong ERI outputs, its numerical contribution is naturally smaller than that of ERI; nevertheless, it addresses a complementary source of error that cannot be resolved through expression interpretation or prompt augmentation alone. This reflects the greater difficulty of refining masks in a near-saturated regime rather than a lower methodological value. As shown in Fig. 15, directly selecting the best cross-prompt mask without MSR can still preserve visually significant over- and under-segmentation errors. MSR analyzes the discrepancy between the selected mask and the point-based mask using DINOv3 feature coherence, verifies the target affiliation of uncertain regions, and converts the resulting diagnosis into corrective positive or negative point prompts. This closed-loop process effectively recovers missing target regions while suppressing irrelevant predictions, producing cleaner, more complete, and more object-consistent masks.

More importantly, the effectiveness of MSR is not limited to the ERI pipeline. As shown in Tab. 15, incorporating MSR into the naïve Qwen2.5-VL 72B + SAM3 (BBox) baseline improves RefCOCO testA/testB by **0.7/1.4** points and ReasonSeg gIoU/cIoU by **0.5/0.4** points. Applying the same MSR mechanism to EVOL-SAM3 also yields consistent gains of **0.4/0.3** points on RefCOCO testA/testB and **0.3/0.3** points on ReasonSeg gIoU/cIoU. The consistent improvements across different MLLM-based grounding pipelines, together with the qualitative corrections in Fig. 15, demonstrate that MSR is a generalizable closed-loop mask-refinement mechanism rather than an ERI-dependent post-processing component.

C.4.6 Importance of DINOv3

DINOv3 is employed at multiple stages of Tarot-SAM3 to provide spatially coherent visual representations. Tab. 13 demonstrates its consistent benefits for both ERI and MSR. Specifically, incorporating DINOv3 into ERI produces gains of **+0.5/+0.6/+0.7** on val/testA/testB, respectively. Its contribution remains consistent in MSR, yielding further improvements of **+0.5/+1.1/+0.3** across the same splits. Notably, DINOv3 improves every evaluated split in both stages, indicating that its effect is not limited to a particular data partition. In ERI, the feature coherence provided by DINOv3 supports more reliable point generation by preserving spatial correspondence between referring semantics and candidate regions. In MSR, it enables more accurate discrepancy-aware mask diagnosis

Table 15: Comparison with alternative baselines and MSR-enhanced variants. For baseline methods, differences are computed against Tarot-SAM3; for MSR-enhanced variants, differences denote gains over their corresponding baselines.

Method	RefCOCO (cIoU)		ReasonSeg test	
	testA	testB	gIoU	cIoU
Qwen2.5-VL 72B + SAM3 (BBox)	67.4 ↓ −11.1	58.4 ↓ −7.5	45.8 ↓ −30.0	37.9 ↓ −33.3
SAM3 Agent (Qwen2.5-VL 72B)	74.3 ↓ −4.2	64.6 ↓ −1.3	70.8 ↓ −5.0	64.0 ↓ −7.2
SAM3 Agent (Gemini 2.5 Pro)	77.8 ↓ −0.7	69.9 ↑ +4.0	74.0 ↓ −1.8	72.1 ↑ +0.9
Qwen2.5-VL 72B + SAM3 (BBox) + MSR (ours)	68.1 ↑ +0.7	59.8 ↑ +1.4	46.3 ↑ +0.5	38.3 ↑ +0.4
EVOL-SAM3 + MSR (ours)	74.1 ↑ +0.4	64.7 ↑ +0.3	71.2 ↑ +0.3	65.3 ↑ +0.3
Tarot-SAM3 (Qwen2.5-VL 7B)	**78.5**	**65.9**	**75.8**	**71.2**

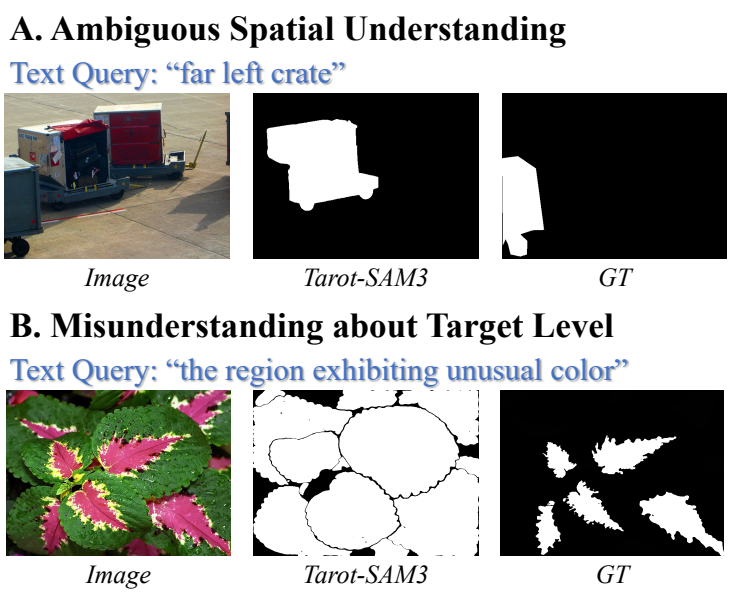

Figure 17: Visualization of failure cases.

and targeted refinement, with the largest gain observed on testA. Similar improvements across the evaluated RES benchmarks further confirm its general effectiveness. Together with the qualitative examples in Fig. 16, DINOv3 is proved as an important component of Tarot-SAM3.

C.4.7 Beyond Naïve Integration and MLLM Scaling.

Tab. 15 shows that naïve MLLM–SAM3 integration remains insufficient, particularly on ReasonSeg. Increasing MLLM capacity substantially improves SAM3 Agent, yet does not consistently dominate Tarot-SAM3: Gemini 2.5 Pro achieves higher RefCOCO scores and ReasonSeg cIoU, whereas Tarot-SAM3 retains the best ReasonSeg gIoU using only a 7B backbone. Moreover, the gap over the naïve Qwen2.5-VL 72B + SAM3 baseline indicates that model scale alone cannot reliably resolve grounding and mask-level errors. *These results highlight the value of structured grounding calibration and closed-loop mask refinement beyond straightforward integration or MLLM scaling.*

C.4.8 Failure Cases.

As shown in Fig. 17, Tarot-SAM3 still exhibits two representative limitations. First, ambiguous spatial expressions may lead to incorrect target grounding when multiple objects admit plausible relative interpretations. For the query "far left crate," the framework selects an incorrect crate because the reference frame of "far left" is not sufficiently constrained. Second, region-level expressions may

cause target-granularity confusion. For "the region exhibiting unusual color," Tarot-SAM3 segments broader plant regions rather than only the intended abnormally colored leaves. These cases motivate future improvements in viewpoint-aware spatial disambiguation and granularity-aware region reasoning.