ERASE-TO-VERIFY: AGENTIC SELF-CORRECTION FOR REFERRING IMAGE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Referring image segmentation aims to identify the object described by a natural-language expression and predict its pixel-level mask. Although frozen vision-language models and promptable segmenters provide a practical training-free solution, their predictions may select the wrong instance or cover only part of the intended target. We propose *Erase-to-Verify*, a training-free framework that treats each predicted mask as a hypothesis to be verified. Given a candidate mask, we locally blur its region and ask whether the referred target remains visible in the resulting image. Visible residual evidence indicates that the current mask is insufficient and should be refined, whereas target absence provides an operational stopping signal. By intervening on the image with the predicted mask, the framework obtains new visual evidence instead of repeatedly judging the prediction within the original observation. For a rejected mask, the framework further diagnoses the failure type. If the prediction corresponds to a wrong instance, it re-localizes the target from the target-suppressed image. If the prediction covers the correct target only partially, it zooms into the residual region and generates point prompts for SAM3 to recover the missing parts. The updated mask is then verified again within the same loop. All components remain frozen, requiring no task-specific fine-tuning, agent training, learned rewards, or learned routing policies. Experiments on all eight RefCOCO-family splits show that our method improves over the strongest reported training-free baseline in our comparison by 6.0 gIoU points on average. Results on ReasonSeg further demonstrate its generalization to implicit reasoning expressions. Our code will be made publicly available soon.

1 INTRODUCTION

Referring image segmentation (RIS) requires a model to segment the object instance specified by a natural-language expression (Hu et al., 2016; Yu et al., 2016; Mao et al., 2016). Expressions such as *the person holding a cup* or *the dog next to the child* cannot be solved by category recognition alone. The model must understand attributes, spatial relations, and scene context to identify the intended instance, and then recover its pixel-level mask. This capability is important for robotics, embodied AI, autonomous driving, and interactive image editing.

Recent foundation-model-based RIS systems have made strong progress by combining language grounding with promptable segmentation. Modular pipelines use a vision-language model (VLM) to generate boxes or points, which are then converted into masks by SAM-family models (Kirillov et al., 2023; Ravi et al., 2025; Carion et al., 2025; Bai et al., 2025b;a). Generative segmentation MLLMs instead integrate mask prediction into the language model through segmentation tokens, mask decoders, or autoregressive mask generation (Lai et al., 2024; Xia et al., 2024; Ren et al., 2024b; Rasheed et al., 2024; Zhang et al., 2024; Lan et al., 2025; 2026; Deng et al., 2025; Ren et al., 2025). More recent agentic systems further allow models to invoke segmentation tools, inspect intermediate outputs, propose new prompts, and iteratively refine their predictions (Zhang et al., 2026; Ye et al., 2025; Liu et al., 2025; Huang et al., 2026; You & Wu, 2025; He et al., 2025; Du et al., 2026).

However, a common limitation remains: many iterative systems ask the model to judge and correct a mask under nearly the same visual evidence that produced the mask in the first place. A mask may look plausible in the original image while selecting a wrong same-category instance, or it may

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.

cover only the most discriminative part of the correct target. In both cases, direct mask inspection, confidence scores, or repeated grounding over the original image may fail to expose whether the current prediction has actually explained the referring expression. Prior studies on LLM self-correction similarly suggest that internal revision alone is often unreliable, whereas feedback from external tools can provide more effective corrective signals (Huang et al., 2024; Gou et al., 2023).

Referring segmentation naturally provides such an external test. A predicted mask defines a region that can be intervened on. If the mask correctly covers the referred target, suppressing recognizable evidence inside that region should make the target disappear from the edited observation. Conversely, if the edited image still contains residual evidence of the referred target, then the current mask has failed to explain the expression: it may have selected a distractor, or it may have covered only part of the correct instance. We therefore treat a mask as an executable visual hypothesis and verify it not by re-inspecting the same image, but by semantically erasing its predicted region through localized target suppression and checking the visual consequence. Figure 1 illustrates this motivation.

Based on this insight, we propose *Erase-to-Verify*, an agentic training-free framework for referring image segmentation. A single frozen Qwen3-VL model (Bai et al., 2025a) is instantiated into three role-specialized agents: a localizer, a verifier, and a refiner. The localizer first predicts a spatial hypothesis, and SAM3 converts it into a candidate mask. The framework then suppresses visual content inside the predicted region to produce an edited observation. The verifier accepts the mask when the referred target is no longer visible. Otherwise, it requests refinement and determines whether the candidate corresponds to a mismatched target or incomplete coverage. A mismatched target is re-localized from the mask-out image, whereas incomplete coverage triggers residual-region zooming and point-guided SAM3 refinement. The repaired mask reenters the same suppress-and-verify loop. All components remain frozen, and the framework uses no task-specific fine-tuning, learned reward, or learned routing policy.

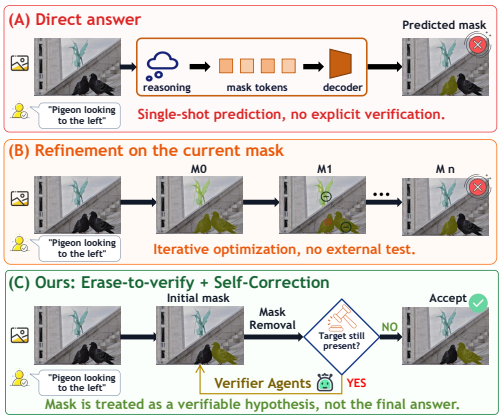

Figure 1: Motivation of *Erase-to-Verify*. Existing refinement loops judge masks within the original evidence space, whereas our framework changes the observation by suppressing visual content inside the predicted region and checking whether residual target evidence remains.

We evaluate *Erase-to-Verify* on the RefCOCO-family benchmarks, including RefCOCO, RefCOCO+, and RefCOCOg, and further assess its generalization on ReasonSeg as an auxiliary test of implicit referring queries. Across eight standard RefCOCO-family splits, *Erase-to-Verify* consistently improves over strong training-free baselines. More importantly, controlled same-stack ablations show that the improvement cannot be explained by iterative grounding alone: the target-suppressed observation provides actionable feedback that direct mask inspection lacks, while statistical analysis shows how the repair branches contribute to refinement gains.

Our contributions are threefold:

- We introduce *erase-to-verify*, a target-suppression strategy for referring image segmentation. We operationalize erasure by locally blurring the predicted region to suppress recognizable target evidence, thereby exposing residual evidence that is difficult to detect from the original image and mask alone.

- We propose an agentic training-free framework that instantiates a frozen MLLM as a localizer, verifier, and refiner. These role-specialized agents collaborate through a deterministic suppress-and-verify controller, requiring no task-specific fine-tuning, learned rewards, or learned routing policies.

- We design a deterministic controller that first decides whether to accept or refine a mask and then selects the appropriate repair for a rejected candidate: re-localization for a mismatched target or residual-region zooming and point-guided refinement for incomplete coverage.

2 Related Work

2.1 Modular and Agentic Referring Segmentation

Promptable segmentation models expose a modular interface between language understanding and pixel prediction. SAM and its successors generate masks from boxes, points, masks, or concept prompts (Kirillov et al., 2023; Ravi et al., 2025; Carion et al., 2025). A VLM can therefore resolve a referring expression into sparse geometry, after which a SAM-family model supplies the object boundary. Grounded-SAM follows this open-world composition pattern (Ren et al., 2024a), and Qwen-family VLMs provide the structured localization needed by similar pipelines (Bai et al., 2025b;a). The modular design is attractive because each component can remain frozen and can be upgraded independently. Its main weakness is the interface between the components: the localizer reasons about the expression but emits sparse geometry, whereas the segmenter predicts pixels without independently establishing that the expression has been satisfied.

Recent systems extend this interface into an iterative or agentic process. Training-free methods search over prompt hypotheses, compare candidate masks, or refine SAM3 predictions at inference time (Ye et al., 2025; Zhang et al., 2026). Other approaches train MLLMs to reason about segmentation or invoke visual tools through supervised fine-tuning and reinforcement learning. Seg-Zero, SAM-R1, and Seg-R1 learn reasoning or prompt-generation policies (Liu et al., 2025; Huang et al., 2026; You & Wu, 2025); RSAgent formulates segmentation as multi-turn tool use (He et al., 2025); and SAM-Veteran predicts initial prompts, refinement points, and stopping decisions (Du et al., 2026).

While these agentic methods demonstrate the benefit of test-time interaction, their verification signals remain internal to the original observation—confidence scores, region affinity, or consistency with the prompt. These statistics are useful heuristics, but they do not constitute a new observation that materially tests the consequence of the mask. A high-confidence mask can still select the wrong instance, and an incomplete mask can retain the target's most discriminative part without triggering a low-confidence alert.

Our work does not seek a more general strategy for search or tool use. Instead, we change the verification signal itself. Rather than estimate mask quality directly, we inspect the mask's visual consequence: after suppressing visual content inside the masked region, is the referred target still visible? This signal is neither an internal confidence estimate from the segmenter nor another statistic of the original prediction. It is an external observation produced by intervening on the image and is therefore separate from the mask-generation process. It directly supports the binary decision of whether the current candidate should be accepted or refined, while the location of residual evidence provides a corrective cue for the next prediction.

The use of target suppression also relates to language-guided object removal and masked image inpainting. Referring object removal aims to delete a language-specified object as the final output, while recent erasing systems study how to preserve context during masked completion (Jiang et al., 2025). In contrast, our goal is not to synthesize a user-facing background. We apply deterministic blur only within the predicted mask to create a diagnostic observation, which is discarded after it provides residual-target evidence. This distinction also clarifies our assumption: the blur should suppress recognizable target evidence inside the mask while leaving image content outside the mask unchanged.

2.2 Generative Segmentation with MLLMs

A complementary line of work integrates mask generation into the multimodal language model. LISA introduces a segmentation token connected to a mask decoder (Lai et al., 2024), and GSVA extends segmentation-token prediction to generalized referring expressions and rejection cases (Xia et al., 2024). PixelLM represents masks with segmentation codebook tokens (Ren et al., 2024b); GLaMM connects grounded language responses with object masks (Rasheed et al., 2024); and

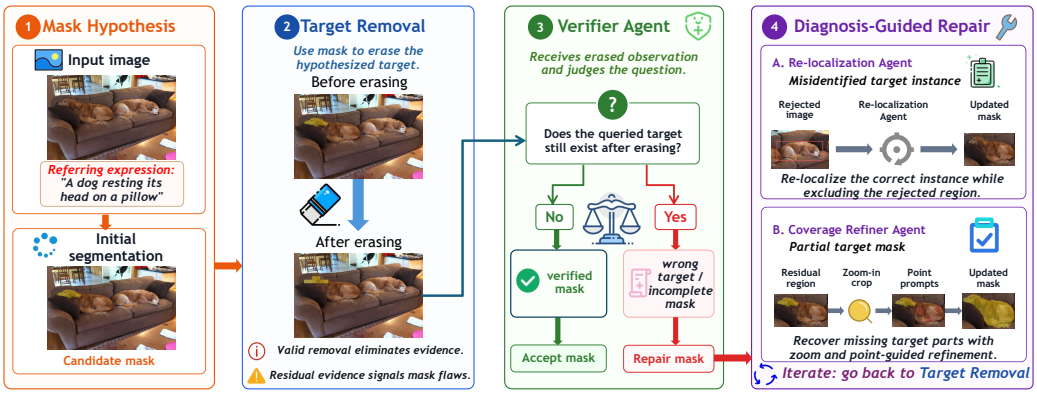

Figure 2: Overview of *Erase-to-Verify*. The framework produces an initial mask, suppresses visual content inside the predicted region to create an edited observation, verifies whether the referred target remains, and routes residual evidence to re-localization or zoom-guided refinement.

PSALM uses mask tokens and a pixel decoder across multiple segmentation tasks (Zhang et al., 2024). X-Decoder and SEEM develop more general pixel–language decoders (Zou et al., 2023a;b). Text4Seg, LlamaSeg, and grounding-token methods further explore text-like or autoregressive mask representations (Lan et al., 2025; 2026; Deng et al., 2025; Ren et al., 2025).

These methods strengthen the coupling between language reasoning and pixel prediction, but most optimize mask production within a forward decoding process and do not introduce an explicit post hoc residual-target test. A mask that selects the wrong instance can remain visually self-consistent, while an incomplete mask can retain the target's most discriminative region. A better decoder or mask representation therefore does not automatically answer the post hoc question of whether the underlying reasoning was correct. Our framework is agnostic to how the mask is generated. It requires no architectural modification, additional mask tokens, or decoder training; instead, it operates directly on any predicted mask and uses an edit-and-verify loop to determine whether residual target evidence remains. It can thus serve as a verification stage in a modular pipeline or be attached to the output of a generative segmentation method.

3 Method

3.1 Overview and Problem Formulation

Referring image segmentation takes an image $I \in \mathbb{R}^{H\times W\times 3}$ and a natural-language expression q as input, and aims to predict a binary mask $M \subseteq \Omega$, where $\Omega = \{1,\dots,H\} \times \{1,\dots,W\}$ is the image domain. Let $G \subseteq \Omega$ denote the unknown ground-truth region of the object referred to by q. At inference time, G is unavailable. We adopt a modular two-stage initial prediction pipeline: a VLM proposes a spatial hypothesis, and SAM3 converts it into a pixel mask. Figure 2 gives an overview of the proposed framework.

Specifically, the Proposal Agent ($\mathcal{A}_P$) receives the original image I and the full referring expression q, and outputs a candidate box b_0 and a compact concept phrase s_0, where b_0 uses normalized image coordinates and s_0 serves as the concept prompt for SAM3:

$$(b_0, s_0) = \mathcal{A}_P(I, q). \tag{1}$$

SAM3 provides a box-prompted interface $\mathcal{S}_{\text{box}}$ that generates a set of candidate masks from (b_0, s_0). We select the one whose tight bounding box best matches b_0 as the initial mask M_0:

$$M_0 = \arg\max_{M\in\mathcal{S}_{\text{box}}(I;b_0,s_0)} \text{IoU}(B(M), b_0), \tag{2}$$

where $B(M)$ denotes the tight bounding box of mask M, and IoU is the intersection-over-union between two boxes. This single forward pass yields an initial prediction, which may suffer from two common failure modes that we focus on in this work: the VLM may localize the wrong instance, or the candidate mask may only partially cover the correct target.

The central object in our verification logic is the unmasked target region

$$U_t = G \setminus M_t, \tag{3}$$

where $U_t \subseteq \Omega$ contains the ground-truth target pixels not covered by the candidate mask at round t. Although U_t is unavailable at inference time, it clarifies what the verification step tests. If $U_t \neq \varnothing$, the candidate has not completely covered the referred target. A wrong-target prediction has little semantic overlap with G and should be replaced, whereas an incomplete prediction covers the correct target only partially and should be augmented. The edited observation described below provides an operational visual proxy for detecting whether U_t remains perceptible.

If subsequent refinements still operate on the original image and the current prediction alone, the model does not obtain new visual evidence that can directly test its current conclusion. To address this, we treat a candidate mask as an executable visual hypothesis and verify it through the visual consequence of suppressing visual content inside the masked region. The framework revolves around two core designs: (1) *Erase-to-Verify*, which constructs a new verification observation via localized target suppression; and (2) *Verifier-Guided Refinement*, which first makes an accept-or-refine decision and then, for a rejected mask, selects re-localization or zoom/point refinement according to the observed failure.

The system consists of a single frozen Qwen3-VL model, SAM3, and a fixed localized blur operator. In *Erase-to-Verify*, *erasure* denotes the functional suppression of recognizable target evidence rather than photorealistic background reconstruction; we implement this operation efficiently with localized blur. Qwen3-VL is prompted in three roles – Proposal Agent, Verification Agent, and Refinement Agent – rather than using three separately trained models. The overall iterative loop is:

$$M_t \rightarrow \widetilde{I}_t \rightarrow z_t \rightarrow a_t \rightarrow M_{t+1}, \tag{4}$$

where $\widetilde{I}_t$ is the target-suppressed image, z_t is the verification result, and a_t is the controller action. The symbol s_t denotes the SAM3 concept phrase associated with M_t; it is regenerated after wrong-target re-localization and otherwise carried into the next round.

3.2 Erase-to-Verify: Residual-Target Verification

For a candidate mask M_t, we suppress recognizable appearance inside the masked region using a fixed blur operator $\mathcal{B}$, while copying pixels outside the mask from the original image:

$$\widetilde{I}_t(x) = \begin{cases} \mathcal{B}(I)(x), & x \in M_t, \\ I(x), & x \in \Omega \setminus M_t, \end{cases} \tag{5}$$

where x denotes a pixel location and $\mathcal{B}(I)(x)$ is the blurred image value at x. Thus, $\widetilde{I}_t$ is a target-suppressed observation that differs from I only inside M_t. If M_t completely covers the referred target and the blur sufficiently suppresses its recognizable appearance, no target evidence should remain outside the masked region. Conversely, visible target evidence in the preserved region indicates that $U_t \neq \varnothing$, so the current mask has failed the coverage test.

The Verification Agent ($\mathcal{A}_V$) receives the overlay of the current mask on the original image, denoted by $O_t = \mathcal{O}(I, M_t)$, together with the edited image $\widetilde{I}_t$ and expression q. It first makes a binary accept-or-refine decision and, when refinement is required, provides an auxiliary repair tag and a localized residual cue:

$$z_t = \mathcal{A}_V(O_t, \widetilde{I}_t, q) = (d_t, c_t, R_t), \tag{6}$$

where $\mathcal{O}$ is the mask-overlay operator and $d_t \in \{\texttt{accept}, \texttt{refine}\}$ is the primary controller decision. If $d_t = \texttt{accept}$, the auxiliary outputs are empty, $c_t = R_t = \varnothing$. If $d_t = \texttt{refine}$, $c_t \in \{\texttt{wrong_target}, \texttt{incomplete_mask}\}$ selects the internal repair operation. For $c_t = \texttt{wrong_target}$, no point or residual box is required and $R_t = \varnothing$; the mask-out image itself is used for re-localization. For $c_t = \texttt{incomplete_mask}$, $R_t \in [0,1]^4$ localizes the remaining target region for subsequent zoom/point refinement. To facilitate parsing, the outputs are emitted in a standardized JSON format.

The verification logic relies on a localized suppression assumption: the blur should remove recognizable target evidence inside M_t, while pixels outside M_t remain unchanged. Under this ideal assumption, visible residual target evidence in $\widetilde{I}_t$ must originate from the unmasked region U_t and is

therefore sufficient to reject M_t. In practice, both the visual suppression and the VLM decision are approximate. Target absence is consequently used as an operational acceptance signal rather than a certificate of pixel-level perfection: an overly large mask can suppress the complete target together with unrelated background, and a subtle residual may be missed by the verifier.

3.3 VERIFIER-GUIDED MASK REFINEMENT

The deterministic controller π maps the verification result to an action a_t. If $d_t = \texttt{accept}$, it sets $a_t = \texttt{stop}$. Otherwise, the auxiliary repair tag selects the internal operation:

$$\begin{aligned} \pi(\texttt{wrong_target}) &= \texttt{relocalize}, \\ \pi(\texttt{incomplete_mask}) &= \texttt{zoom_refine}. \end{aligned} \tag{7}$$

Wrong-Target Relocalization. When the current candidate is judged to correspond to a mismatched target, the Proposal Agent receives the target-suppressed image and the original expression, and generates a new box and concept phrase:

$$(b_{t+1}, s_{t+1}) = \mathcal{A}_P(\widetilde{I}_t, q). \tag{8}$$

Because recognizable content inside M_t has been suppressed, the edited observation discourages the Proposal Agent from selecting the same mismatched target again without maintaining an explicit rejected-region history or using point prompts. The new proposal is then applied to the original image through SAM3, selecting the candidate whose tight bounding box best matches b_{t+1}:

$$M_{t+1} = \arg \max_{M \in \mathcal{S}_{\text{box}}(I; b_{t+1}, s_{t+1})} \text{IoU}(B(M), b_{t+1}). \tag{9}$$

The repaired mask then re-enters the verification loop.

Incomplete-Mask Zoom-Guided Refinement. The current mask is considered semantically aligned with the target but incomplete. The system uses the residual region R_t to construct a corrected mask. We crop a window W_t centered at R_t with a padding factor ρ, obtaining $I_t^z = \mathcal{C}(I, W_t)$ and $\widetilde{I}_t^z = \mathcal{C}(\widetilde{I}_t, W_t)$ from the original and edited images, respectively, where $\mathcal{C}$ denotes cropping and ρ controls the context retained around R_t. Zooming focuses the Refinement Agent on small or thin residual regions that are difficult to localize in the full image, while the edited crop makes the missing target evidence visually explicit.

The Refinement Agent ($\mathcal{A}_R$) compares the two crops and generates signed point prompts:

$$\begin{aligned} P_t &= \mathcal{A}_R(I_t^z, \widetilde{I}_t^z, q, R_t), \\ P_t &= \{(p_k, \ell_k)\}_{k=1}^K, \quad \ell_k \in \{-1, +1\}. \end{aligned} \tag{10}$$

where K is the number of generated prompts, $p_k \in [0,1]^2$ is the normalized crop coordinate of the k-th point, and $\ell_k \in \{-1, +1\}$ is its signed label. Positive points mark target pixels that should be included, whereas negative points mark distractor or background pixels that should be excluded. The coordinate transform Π_{W_t} maps these points from the crop to the original image. SAM3 jointly interprets the signed points and directly predicts the complete corrected mask:

$$M_{t+1} = \mathcal{S}_{\text{point}}(I; \Pi_{W_t}(P_t), s_t). \tag{11}$$

where $\mathcal{S}_{\text{point}}$ denotes the point-prompted SAM3 interface. Because the output is a complete mask rather than an incremental region, the update can both add missing target pixels and remove pixels contradicted by negative prompts. This branch is triggered only when the Verification Agent judges the current mask to correspond to the correct target but to have incomplete coverage.

After either repair, the new mask M_{t+1} and its associated concept phrase re-enter the same suppression-and-verification loop; no history of earlier rejected regions is maintained. If the Verification Agent predicts `accept`, the current mask is returned immediately. We execute at most $T = 3$ refinement rounds. If the verifier still predicts `refine` when this budget is exhausted, the latest mask M_T is returned. All components remain frozen throughout; no parameter updates or learned policies are involved.

Method	RefCOCO			RefCOCO+			RefCOCOg	
	Val	TestA	TestB	Val	TestA	TestB	Val	Test
GL-CLIP (Yu et al., 2023)	32.9	34.9	30.1	38.4	42.1	32.7	42.0	42.0
CaR (Sun et al., 2024)	33.6	35.4	30.5	34.2	36.0	31.0	36.7	36.6
Ref-Diff (Ni et al., 2023)	37.2	38.4	37.2	37.3	40.5	33.0	44.0	44.5
TAS (Suo et al., 2023)	39.8	41.1	36.2	43.6	49.1	36.5	46.6	46.8
IteRPrimE (Wang et al., 2025b)	40.2	46.5	33.9	44.2	51.6	35.3	46.0	45.1
Pseudo-RIS (Yu et al., 2024)	41.1	48.2	33.5	44.3	51.4	35.1	46.0	46.7
LGD (Li et al., 2025)	49.5	54.7	41.0	49.6	58.4	38.6	50.3	51.1
VLM-VG (Wang et al., 2025a)	49.9	53.1	46.7	42.7	47.3	36.2	48.0	48.5
HybridGL (Liu & Li, 2025)	49.5	53.4	45.2	43.4	49.1	37.2	51.3	51.6
SAM3 Agent (Carion et al., 2025)	59.4	64.3	55.0	51.4	57.0	44.9	57.2	58.8
EVOL-SAM3 (Ye et al., 2025)	68.7	73.7	64.4	64.4	67.8	54.0	64.7	65.5
Tarot-SAM3 (Zhang et al., 2026)	71.8	75.5	65.5	64.8	70.7	56.0	67.1	69.4
Erase-to-Verify (Qwen2.5-VL-7B)	73.8	76.9	68.6	67.6	72.7	61.0	70.0	69.8
Erase-to-Verify (Qwen3-VL-4B)	**76.9**	**79.6**	**72.6**	**71.6**	**76.3**	**65.0**	**73.5**	**73.7**

Table 1: End-to-end comparison on the RefCOCO-family benchmarks. Values are reported as gIoU scores (%); higher is better. For the training-free baseline rows, SAM3 Agent, EVOL-SAM3, and Tarot-SAM3 use Qwen2.5-VL-7B as the VLM backbone. The final two rows report *Erase-to-Verify* with Qwen2.5-VL-7B and Qwen3-VL-4B, respectively.

4 EXPERIMENTS

In this section, we first describe the datasets, data usage, evaluation protocol, baselines, and implementation details. We then report the main results on the RefCOCO-family benchmarks (Yu et al., 2016) and the more challenging ReasonSeg benchmark (Lai et al., 2024). Finally, we conduct controlled ablation studies and additional statistical analyses to examine the effectiveness and component contributions of the proposed target-suppression feedback mechanism. These experiments are designed to isolate the contribution of the target-suppressed observation from the benefits of iterative inference and localized repair, ensuring that performance gains can be attributed to the proposed verification signal rather than to increased inference budget alone.

4.1 EXPERIMENTAL SETUP

Datasets and Metrics. We evaluate *Erase-to-Verify* on the RefCOCO family, including RefCOCO, RefCOCO+, and RefCOCOg, using their standard evaluation splits. Specifically, we report Val, TestA, and TestB for RefCOCO and RefCOCO+, and the Val and Test splits for RefCOCOg. We report global intersection over union (gIoU) for the RefCOCO-family main comparison and for the controlled component ablation under the same model stack. We additionally evaluate on ReasonSeg, which contains implicit queries requiring higher-level visual reasoning. The cross-method ReasonSeg comparison reports both cIoU and gIoU, while the stage-wise analysis in Fig. 3 uses gIoU to align with the controlled refinement analysis.

Training-Free Protocol. Our method performs no task-specific training or parameter optimization. The training splits and ground-truth masks of the evaluation benchmarks are not accessed during inference, and annotations are used only for computing evaluation metrics. All foundation models remain frozen, and no segmentation-specific fine-tuning, agent training, reward learning, or learned routing policy is introduced.

Implementation Details. We use a single frozen Qwen3-VL-4B model for the Proposal, Verification, and Refinement Agents through role-specific prompts. SAM3 is used as the segmentation tool. We operationalize erasure with a fixed localized blur that suppresses recognizable target evidence inside each candidate mask while preserving the image outside it. Each sample begins with one localization and one SAM3 prediction. The verifier receives the mask overlay and target-suppressed image and first makes a binary *Accept/Refine* decision. For a rejected candidate, it further selects wrong-target re-localization or incomplete-mask zoom/point refinement. We allow at most three

Method	Val Set		Test Set	
	gIoU	cIoU	gIoU	cIoU
Training-based methods				
LISA-7B (Lai et al., 2024)	52.9	54.0	47.3	34.1
Seg-Zero-7B (Liu et al., 2025)	62.6	62.0	57.5	52.0
SAM-Veteran-7B (Du et al., 2026)	68.2	67.3	62.6	56.1
Training-free methods				
SAM3 Agent-7B (Carion et al., 2025)	62.2	49.1	63.0	53.5
EVOL-SAM3-3B (Ye et al., 2025)	66.9	57.1	65.9	58.9
Tarot-SAM3-3B (Zhang et al., 2026)	67.5	57.3	66.7	59.5
Erase-to-Verify	**73.0**	**68.3**	**67.8**	**60.5**

Table 2: Comparison on ReasonSeg. Values report gIoU and cIoU scores (%) on the validation and test splits. Methods are grouped into training-based and training-free subsets for readability. The closest-scale comparison for our Qwen3-VL-4B configuration is restricted to the 3B variants of EVOL-SAM3 and Tarot-SAM3.

Variant	Iter.	E2V	Zoom/ Point	RefCOCO TestA	RefCOCOg Test	RefCOCO+ TestA	RSeg Val	RSeg Test
Base	×	×	×	72.2	70.6	69.9	69.1	63.3
+ Iteration (Overlay)	✓	×	×	75.9	71.6	72.7	70.7	65.4
+ *Erase-to-Verify*	✓	✓	×	77.2	72.4	74.8	71.1	66.4
+ Zoom/Point Refinement (Ours)	✓	✓	✓	**79.6**	**73.7**	**76.3**	**73.0**	**67.8**

Table 3: Cumulative component ablation across five evaluation splits under the same frozen model stack. Iter. denotes iterative verification, E2V denotes the target-suppressed *Erase-to-Verify* observation, and Zoom/Point denotes residual-region zooming with signed point prompts. Values are gIoU scores (%).

refinement rounds and return the latest mask when the budget is exhausted. All experiments are run on a server equipped with eight NVIDIA L20 GPUs.

4.2 Main Results

Results on RefCOCO-Family Benchmarks. Table 1 compares *Erase-to-Verify* with reported foundation-model-based RIS systems across eight standard RefCOCO-family evaluation splits. The results measure end-to-end system performance; the controlled experiments in Sec. 4.3 further isolate the contribution of target-suppression feedback.

Among the methods compared in Table 1, the Qwen3-VL-4B configuration of *Erase-to-Verify* achieves the highest performance on all eight splits. Relative to Tarot-SAM3, the strongest prior training-free system in the table, the gains are 5.1, 4.1, and 7.1 gIoU points on RefCOCO Val, TestA, and TestB; 6.8, 5.6, and 9.0 points on RefCOCO+; and 6.4 and 4.3 points on RefCOCOg, for an average improvement of 6.0 gIoU points. Under the Qwen2.5-VL-7B backbone, the additional *Erase-to-Verify* configuration improves over Tarot-SAM3 by 2.5 gIoU points on average across the same eight splits.

The gains are consistent across all eight evaluation splits, with the largest improvements appearing on RefCOCO+ TestB and RefCOCOg Val. The improvements on RefCOCO+ and RefCOCOg indicate that target-suppression feedback remains useful for expressions that depend on attributes, relations, and longer context rather than explicit location words.

Generalization to ReasonSeg. We further evaluate the same frozen framework on ReasonSeg, which contains implicit referring queries requiring higher-level visual reasoning. We treat ReasonSeg as an auxiliary generalization study rather than the primary state-of-the-art comparison, because published methods differ substantially in supervision, backbone, and inference budget.

As shown in Table 2, *Erase-to-Verify* reaches 73.0 gIoU and 68.3 cIoU on the validation split, and 67.8 gIoU and 60.5 cIoU on the test split. To control for backbone scale, we explicitly restrict the parameter-comparable comparison to the closest reported training-free configurations: EVOL-SAM3 and Tarot-SAM3 use Qwen2.5-VL-3B, while our system uses Qwen3-VL-4B. Under this 3B–4B setting, *Erase-to-Verify* improves over the stronger of these baselines, Tarot-SAM3, by 5.5 gIoU and 11.0 cIoU points on validation, and by 1.1 gIoU and 1.0 cIoU points on test. The 7B-based entries are included as broader context rather than as the basis for this scale-controlled claim. These results demonstrate that the same target-suppression framework remains effective when the initial error arises from implicit visual reasoning or semantic grounding rather than mask prediction alone.

4.3 Ablation Study

We conduct a cumulative component ablation using the same frozen Qwen3-VL and SAM3 checkpoints. Starting from a single-pass baseline, we progressively introduce overlay-based iteration, the target-suppressed observation, and residual-region zoom/point refinement. This design separates the contribution of iterative inference from that of *Erase-to-Verify* and its localized refinement mechanism.

Progressive Component Analysis. Table 3 compares four cumulative variants under the same model stack. *Base* is a single-pass Qwen3-VL–SAM3 baseline without iterative verification. + *Iteration (Overlay)* introduces the iterative controller while verifying each candidate from its mask overlay. + *Erase-to-Verify* additionally provides the verifier with the target-suppressed observation, but excludes residual-region zooming and signed point prompts; when refinement is requested, it continues with the standard proposal-and-segmentation process in the next round. The complete method further adds residual-region zooming and signed point prompting. All iterative variants use the same maximum refinement budget.

4.4 Statistical Analysis

Beyond aggregate benchmark performance, we examine two questions: whether complete refinement improves the initial masks, and which repair branch contributes the largest conditional gains. Figure 3 summarizes both analyses.

Refinement Effectiveness. Figure 3(a) follows each sample from the initialized mask M_0 to the final output. Averaged over three RefCOCO-family splits, refinement raises gIoU from 70.9 to 76.5; on ReasonSeg, it raises gIoU from 66.2 to 70.4.

Repair-Type Contribution. Figure 3(b) shows that wrong-target re-localization is selected in 6.4% of cases and yields a 39.6-point conditional gain, while incomplete-mask repair is selected in 21.1% and yields a 25.9-point gain. These are gains on routed samples, confirming that both branches provide useful corrections.

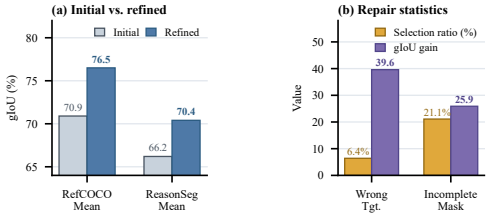

Figure 3: Statistical analysis of gIoU-based refinement behavior.

5 Conclusion

We present *Erase-to-Verify*, a training-free framework that refines referring masks by suppressing the predicted region and testing target visibility. If the target remains visible, the controller decides between re-localizing mismatched targets and recovering incomplete coverage via zooming and point-guided SAM3 prompting—all without parameter updates or learned routing. It outperforms the best training-free baseline by 6.0 gIoU across eight RefCOCO splits and also improves on ReasonSeg. Ablations confirm that improvements stem from suppression and targeted repair, not from repeated iteration alone.

AI USE STATEMENT

Generative AI tools were used for LaTeX format conversion and language editing during manuscript preparation. All AI-assisted changes were reviewed by the authors, who take responsibility for the final content of this work.

REFERENCES

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

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.

Jiru Deng, Tengjin Weng, Tianyu Yang, Wenhan Luo, Zhiheng Li, and Wenhao Jiang. Llamaseg: Image segmentation via autoregressive mask generation. *arXiv preprint arXiv:2505.19422*, 2025.

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.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738*, 2023.

Xingqi He, Yujie Zhang, Shuyong Gao, Wenjie Li, Lingyi Hong, Mingxi Chen, Kaixun Jiang, Jiyuan Fu, and Wenqiang Zhang. Rsagent: Learning to reason and act for text-guided segmentation via multi-turn tool invocations. *arXiv preprint arXiv:2512.24023*, 2025.

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

Jiaqi Huang, Zunnan Xu, Jun Zhou, Ting Liu, Yicheng Xiao, Mingwen Ou, Bowen Ji, Xiu Li, and Kehong Yuan. Sam-r1: Leveraging sam for reward feedback in multimodal segmentation via reinforcement learning. *Advances in Neural Information Processing Systems*, 38:138362–138383, 2026.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International conference on learning representations*, volume 2024, pp. 32808–32824, 2024.

Longtao Jiang, Zhendong Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Lei Shi, Dong Chen, and Houqiang Li. Smarteraser: Remove anything from images using masked-region guidance. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24452–24462, 2025.

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

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

Mengcheng Lan, Chaofeng Chen, Yue Zhou, Jiaxing Xu, Yiping Ke, Xinjiang Wang, Litong Feng, and Wei Zhang. Text4seg: Reimagining image segmentation as text generation. In *International Conference on Learning Representations*, volume 2025, pp. 1634–1661, 2025.

Mengcheng Lan, Chaofeng Chen, Jiaxing Xu, Zongrui Li, Yiping Ke, Xudong Jiang, Yingchen Yu, Yunqing Zhao, and Song Bai. Text4seg++: Advancing image segmentation via generative language modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

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, 2025.

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.

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*, 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.

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.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

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

Xiangxuan Ren, Zhongdao Wang, Liping Hou, Pin Tang, Guoqing Wang, and Chao Ma. Grounding everything in tokens for multimodal large language models. *arXiv preprint arXiv:2512.10554*, 2025.

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

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.

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. 8057–8067. IEEE, 2025a.

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, 2025b.

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.

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

Zuyao You and Zuxuan Wu. Seg-r1: Segmentation can be surprisingly simple with reinforcement learning. *arXiv preprint arXiv:2506.22624*, 2025.

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.

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.

Weiming Zhang, Dingwen Xiao, Songyue Guo, Guangyu Xiang, Shiqi Wen, Minwei Zhao, Lei Chen, and Lin Wang. Tarot-sam3: Training-free sam3 for any referring expression segmentation. *arXiv preprint arXiv:2604.07916*, 2026.

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, 2024.

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

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

A Additional Implementation Details

This appendix provides implementation and reproducibility details for *Erase-to-Verify*. The method uses a frozen Qwen3-VL-4B model for the Proposal, Verification, and Refinement Agents, together with the frozen SAM3 segmenter. No training split or ground-truth mask is accessed during inference. Each sample starts with one grounding request and one SAM3 prediction, followed by at most three verification-and-repair rounds. When a candidate is rejected, the controller either re-localizes a wrong instance or uses a zoomed residual region with signed point prompts to refine an incomplete mask.

The erasure operation is implemented as localized blur: pixels inside the candidate mask are replaced by the corresponding values from a fixed blurred image, while pixels outside the mask are copied unchanged from the original image. Thus, "erasure" refers to suppression of recognizable target evidence, not to photorealistic image reconstruction or inpainting. The verifier receives the current mask overlay together with this target-suppressed observation. If the target remains visible, the candidate is rejected; if the verifier accepts the candidate or the refinement budget is exhausted, the current mask is returned.

For the RefCOCO-family benchmarks, we use the standard validation and test splits and report gIoU. For ReasonSeg, we report both gIoU and cIoU following the protocol used in the main paper. Ground-truth masks are used only for evaluation. All models remain frozen, and experiments are run on a server with eight NVIDIA L20 GPUs.

B Inference Time

Table 4 reports the available per-sample inference times on the ReasonSeg test set. The values should be interpreted with their original measurement protocol and hardware configuration; they are provided as an efficiency reference rather than as a fully controlled runtime ablation across independently implemented systems.

Model	Time (s)
SAM3 Agent (Carion et al., 2025)	37.8
EVOL-SAM3 (Ye et al., 2025)	23.2
Tarot-SAM3 (Zhang et al., 2026)	15.2
Erase-to-Verify	15.3

Table 4: Per-sample inference time on the ReasonSeg test set, in seconds.

C Prompt Templates

The following templates are the fixed prompts used by the frozen VLM. The placeholders are filled with the current referring expression, visual context, coordinates, and module-specific limits. The verifier uses the terms `accept` and `refine`; a rejected candidate is classified as `wrong_target` or `incomplete_mask`, matching the labels used in the main text.

Subject extraction.

> Look at the image and answer what visible target this request refers to: "phrase." Return one short canonical visual subject for SAM3 and concise spatial cues for grounding. Return strict JSON with the fields `subject`, `scope`, `location_description`, `visual_cues`, and `reason`. Use a concrete visible object category or part rather than a full sentence.

Initial grounding.

> Provide the bounding box coordinates for the region described by "phrase." The inferred subject is "subject." Use the location hint "location_description" and visual cues "visual_cues" to avoid wrong instances. Return strict JSON with a list of tight `bbox_2d` boxes and a concise grounding reason. If the target cannot be localized confidently, return an empty box list.

Mask verification.

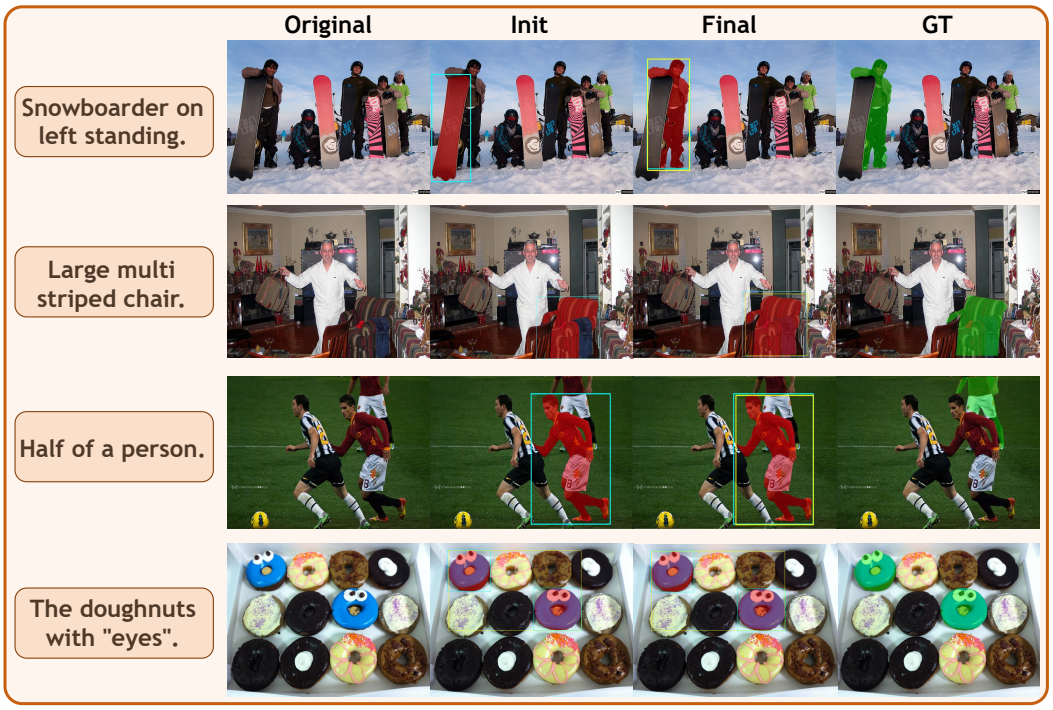

Figure 4: Qualitative cases for *Erase-to-Verify*. The first row illustrates wrong-instance correction, the second illustrates recovery of missing target regions, the third shows a failure case, and the fourth shows a ReasonSeg example.

> Image 1 is the current SAM mask overlay. Image 2 is the target-suppressed observation after localized blur inside the current mask. For target "phrase," determine whether any part of the target remains visible in Image 2. Return `accept` only when no target evidence remains; otherwise return `refine` with a short reason.

Failure classification.

> Judge whether the current mask matches the requested target using the overlay and target-suppressed observation. A mask mainly on a different object is `wrong_target`. A mask on the correct object that leaves visible target evidence is `incomplete_mask`. Return one JSON object with the fields `case`, the required coordinates, and a short reason.

Zoom and point repair.

> Use the full image for global context and the zoomed crop for precise localization. Locate visible parts of the same target that lie outside the current red mask. Return JSON containing `residual_bbox_2d`, `missing_points`, and `avoid_points`. Missing points must lie inside missed target parts; avoid points must lie on clear non-target regions that were incorrectly included.

D Qualitative Case Studies

Figure 4 shows representative behaviors of *Erase-to-Verify*. The columns show the original image, the initialized prediction, the refined prediction, and the ground-truth mask. The examples include wrong-instance correction, incomplete-mask refinement, a failure case, and a ReasonSeg example.

The cases illustrate why the target-suppressed observation can provide useful feedback: residual target evidence supports either re-localization of a wrong instance or localized repair of an incomplete mask. They also expose the method's limits when similar instances are difficult to disambiguate or when the edited observation does not contain enough information for correction.

E LIMITATIONS

The suppression test is one-sided. Target disappearance is an operational acceptance signal, but it does not certify pixel-level boundary precision, so over-segmentation can remain difficult to detect. Performance also depends on the frozen VLM's visual judgment, the initial SAM3 proposal, and the quality of the localized blur. Finally, the verification-and-repair loop uses additional inference calls and is currently evaluated only for still-image referring and reasoning segmentation. These limitations motivate future work on calibrated verification, more efficient early-exit policies, and extensions to video or other intervenable visual hypotheses.