EVIDENCE IS NOT EXTENT: SEPARATING REFERENTIAL EVIDENCE, OBJECT EXTENT, AND BOUNDARY TRUST FOR REFERRING IMAGE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Parameter-efficient referring image segmentation asks one dense score to name the referent, cover its extent, and vouch for its boundary. However, evidence that identifies the referent is often sparse, such as a color patch or an anchor object, while the final mask must cover the full object extent. Therefore, sharpening the score may truncate the object, whereas a shape prior that completes it can spill onto same-category distractors. We introduce E3-RIS, separating these decisions by authority: each source may change only the pixels it is entitled to change. Prompt-Anchored Referential Evidence (PARE) pools prompt-initialized queries into one posterior that names the referent. Evidence-Constrained Bidirectional Projection (ECBP) lets frozen SAM add or remove pixels only inside a boundary band where the posterior agrees. Boundary Trust-Region Arbitration (BTRA) admits a direction-fixed edit when a learned score clears its threshold, and otherwise keeps the projected mask. On RefCOCO val, E3-RIS lifts mIoU over TALENT from 77.8 to 78.5 and Pr@0.9 from 40.1 to 44.2. It also exceeds TALENT in mIoU across all nine official RefCOCO, RefCOCO+, and RefCOCOg splits.

1 INTRODUCTION

Referring image segmentation (RIS) predicts the mask of the object that a free-form expression describes (Hu et al., 2016; Ji et al., 2024), and parameter-efficient tuning (PET) now adapts frozen vision–language backbones (Radford et al., 2021; Oquab et al., 2024) to it with a small trainable set (Xu et al., 2023; Wang et al., 2024; Huang et al., 2025; Jin et al., 2026). The remaining failures lie inside the category. A phrase such as "the second man to the right of the woman in red" names its referent through a color patch and an ordinal relation (Yu et al., 2016; Yang et al., 2022; Wang et al., 2022), so the evidence occupies a few pixels of a mask that must cover the whole man. Therefore, a dense score sharp enough to pick the right man may produce a truncated mask, whereas a shape prior that completes the object may spill onto a same-category neighbor.

Three lines of prior work have attempted to address this bottleneck, but it remains unresolved. PET methods strengthen cross-modal fusion or suppress non-target activation (Xu et al., 2023; Wang et al., 2024; Yu et al., 2025a; Huang et al., 2025; Jin et al., 2026), while the final prediction is still obtained from a single dense score using one threshold. Sharpening this score to better distinguish the referent can therefore truncate the object. SAM-based methods either train a large vision–language encoder with SAM or use a PET model to prompt a frozen SAM (Kirillov et al., 2023; Zhang et al., 2024; Liu et al., 2025; Shang et al., 2024; Zhang et al., 2023). In both cases, the decoder outputs the final mask without explicitly restricting which pixels the proposal may claim. If the prompt lands on the wrong instance, the entire mask may follow that instance. Boundary refiners modify selected predictions in different ways: PointRend re-predicts uncertain points (Kirillov et al., 2020), SegFix transports interior labels to the boundary (Yuan et al., 2020), and mask scoring evaluates the whole mask (Huang et al., 2019). However, once a selected pixel is re-predicted, there is no default action that preserves its current label, no predefined direction for adding or removing a pixel, and no mechanism for rejecting an uncertain edit. In other words, selected pixels are modified without explicitly deciding whether a particular edit should be accepted or rejected.

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 1: E3-RIS on “big man” (RefCOCO val). Left to right: scene, PARE evidence, ECBP extent, and BTRA boundary trust. All overlays and accept/reject decisions come from the frozen single-dataset pipeline.

Better fusion can sharpen the posterior, but it does not change the fact that the same posterior is still asked to identify the referent, cover its full extent, and determine its own boundary. No single threshold can adequately serve all three decisions. We therefore introduce a rule of authority, under which each source may change only the pixels it is entitled to change.

E3-RIS assigns distinct authorities to three decisions: referential evidence, object extent, and boundary trust. Prompt-Anchored Referential Evidence (PARE) replaces the base decoder’s single weighted text vector with four prompt-initialized queries over the expression tokens and pools their responses into a single posterior that identifies the referent. In Figure 1, where several similar men are distinguished mainly by size, it selects the larger one. Evidence-Constrained Bidirectional Projection (ECBP) prompts frozen SAM with that posterior and adds or removes pixels only inside a scale-normalized boundary band where the posterior agrees. In this way, the shape prior recovers object extent without participating in identity selection, and it completes only the selected man. Boundary Trust-Region Arbitration (BTRA) keeps the projected mask M_1 by default and edits only inside an outer-addition or inner-removal ring when a learned score clears a train-selected threshold. In Figure 1, it accepts one edit on the man and rejects another. Both act near the boundary and differ in admission: ECBP applies one fixed posterior-compatible rule, BTRA a learned score that rejects by default.

Our main contributions are as follows:

- We design PARE, which pools four prompt-initialized queries over the expression tokens into a single posterior, allowing sparse referential evidence to be captured from multiple anchors rather than from the single weighted text vector used in existing PET decoders.
- We propose ECBP and BTRA to constrain frozen SAM edits. ECBP adds or removes pixels only inside a posterior-gated boundary band to recover object extent without participating in identity selection, while BTRA admits direction-fixed edits only when a learned score meets its threshold and otherwise keeps the projected mask, providing an explicit reject option.
- We show that E3-RIS improves Pr@0.9 from 40.1 to 44.2 and mIoU from 77.8 to 78.5 over TALENT on RefCOCO val, while achieving higher mIoU on all nine official RefCOCO, RefCOCO+, and G-Ref splits.

2 Related Work

One fused score for identity and extent. RIS has moved from recurrent fusion to cross-modal transformers (Hu et al., 2016; Yang et al., 2022; Wang et al., 2022; Tang et al., 2023), while modular or staged designs decompose the expression or reasoning before producing a single mask (Yu et al., 2018; Dai et al., 2025; Nguyen-Truong et al., 2025). Parameter-efficient variants freeze the encoders and train bridges or adapters (Xu et al., 2023; Wang et al., 2024; Yu et al., 2025a; Huang et al., 2025). TALENT further suppresses non-target activation (Jin et al., 2026). Despite these advances, the final prediction still relies on a single dense score, and one threshold must both separate the referent from its neighbors and cover its full extent. Referential evidence is often concentrated in only a few pixels, so a score sharp enough to identify the correct instance may weaken over the rest of the object. E3-RIS retains a single posterior but uses it to gate a separate extent operator.

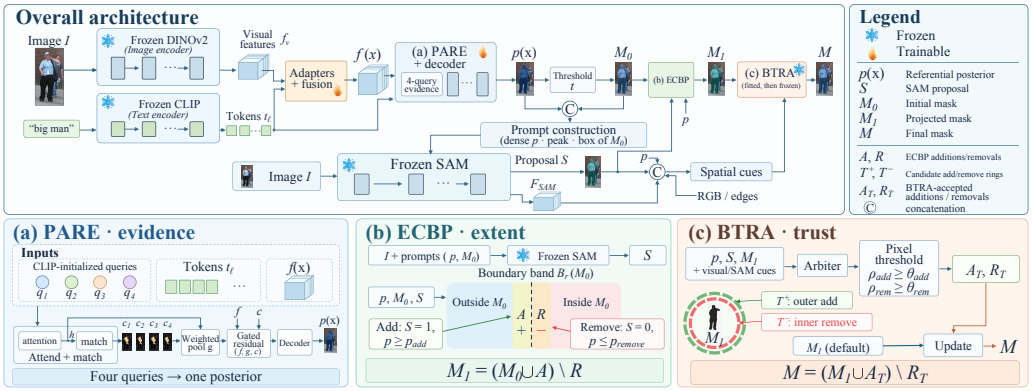

Figure 2: E3-RIS overview on the Figure 1 case: PARE predicts $p(x)$, ECBP projects M_0 to M_1, and BTRA admits direction-fixed boundary edits. Image/text encoders and SAM are frozen.

Prompting hands SAM the mask. SAM supplies a class-agnostic shape prior (Kirillov et al., 2023). One line of RIS systems couples a large vision–language encoder or a multimodal LLM to SAM and trains the pair (Zhang et al., 2024; Liu et al., 2025; Li et al., 2026; Lai et al., 2024). A parameter-efficient line instead freezes SAM and learns only the prompts or adapters that steer it (Shang et al., 2024; Zhang et al., 2023; Ito, 2025). In both cases, the decoder emits the final mask, while SAM itself does not read the referring expression. A prompt that lands on a same-category neighbor may therefore complete the wrong instance as a full mask, causing the intended referent to be lost. Such failures can offset gains elsewhere when results are pooled over pixels, and on our own logits whole-mask replacement leaves oIoU unchanged (Table 4). E3-RIS instead constrains SAM's access, allowing the proposal to add or remove pixels only inside a boundary band and only where the posterior agrees.

Refinement without a default. Boundary refinement methods typically improve segmentation by revisiting uncertain or boundary pixels. PointRend re-predicts labels at uncertain contour points, while SegFix transports interior labels outward to boundary pixels (Kirillov et al., 2020; Yuan et al., 2020). In both cases, selected pixels receive new labels, leaving no default action that preserves the current prediction and no predefined edit direction. Mask scoring and selective prediction instead attach confidence to whole predictions and allow abstention (Huang et al., 2019; Geifman & El-Yaniv, 2019), indicating whether a mask should be trusted rather than which pixels should change. E3-RIS adopts a different rule: BTRA keeps the projected mask by default and admits a direction-fixed edit inside a predeclared ring only when a learned score meets a train-selected threshold. The matched-capacity control in Table 4 tests this design, with extended context provided in Appendix I.

3 Method

3.1 Overview

One posterior, three decisions. Given an image I and an expression Q, RIS estimates a posterior $p(x) = \sigma(z(x))$ over pixels x and thresholds it once, $M_0 = \{x \mid p(x) > t\}$, so the same dense score must identify the referent, cover its extent, and determine its boundary. To separate these roles, E3-RIS retains this output interface but specifies which source may change which pixels (Figure 2). PARE supplies the *referential evidence* through a multi-query representation whose pooled responses form the posterior. ECBP adjusts *object extent* only inside a posterior-gated boundary band around M_0, using a frozen SAM proposal to guide local edits. BTRA adjusts *boundary pixels* only inside an outer-addition ring and an inner-removal ring around the projected mask M_1, and only when a learned score exceeds a train-selected threshold. Pixels outside these regions remain unchanged.

3.2 PROMPT-ANCHORED REFERENTIAL EVIDENCE

Prompt-anchored query pool. The base decoder aggregates the expression tokens into one weighted text vector before interacting with the visual features (Appendix A), so the cue that separates the referent from a same-category neighbor competes with every other token for that vector. We instead read the token sequence with $K = 4$ parallel queries q_k, each initialized with the frozen CLIP embedding of a distinct fixed prompt (strings in Appendix A) and jointly adapted during training. Let t_ℓ be the CLIP feature of content token ℓ and $f(x)$ the visual feature at pixel x. With $\tau_t = 0.07$ and $\tau_e = 0.2$,

$$\begin{aligned} a_{k\ell} &= \mathrm{softmax}_\ell(\cos(q_k, t_\ell)/\tau_t), & h_k &= \mathrm{LN}\big(q_k + \textstyle\sum_\ell a_{k\ell} t_\ell\big), \\ e_k(x) &= \cos(W_v f(x), W_k h_k)/\tau_e, & c_k(x) &= \sigma(e_k(x)), \end{aligned} \tag{1}$$

where W_v and W_k are linear projections, h_k is the token-conditioned representation of query k, and $c_k(x)$ is its query map. The pooled evidence context $g(x) = K^{-1} \sum_k c_k(x) W_k h_k$ enters the decoder through a gated residual initialized near zero, $f'(x) = f(x) + \tanh(\gamma)\, \phi([f(x), g(x), c_1(x), \ldots, c_4(x)])$, with ϕ a two-layer 1×1 convolution and γ a learned scalar.

Containment, one posterior. A weak localization loss encourages two of the four maps to localize the referent,

$$\mathcal{L}_{\mathrm{loc}} = -\frac{1}{2} \sum_{k=1}^{2} \log \sum_x Y(x)\ \mathrm{softmax}_x(e_k(x)), \tag{2}$$

where $Y \in \{0,1\}^{H \times W}$ is the ground-truth mask; the other two maps receive no containment constraint, so part of the pool is tied to the target and the rest remains free to respond to context (Figure 5). Two small-weight regularizers, a diversity term on the attention vectors and an anchoring term that keeps each query near its prompt initializer, complete the objective (Appendix A). There is one mask head, and the prompt strings specify initializers rather than learned roles: the four maps together form an evidence representation for a single posterior. Table 7 reports the stage gain over the B3 base before any SAM edit; the frozen PARE posterior is the input to every downstream gain in Table 3.

3.3 EVIDENCE-CONSTRAINED BIDIRECTIONAL PROJECTION

Evidence is not extent. At full budget, PARE improves over B3 on every reported metric (Table 7), yet the frozen validation cases in Figure 3 show that the sharper posterior can still under-cover the referent. This indicates that identifying the referent and covering its full extent are different decisions, while a single threshold on one score directly addresses only the former. Handing the mask to the shape prior (Shang et al., 2024; Ito, 2025) repairs extent but lets a wrong-instance proposal take the whole mask.

Posterior-gated band. We instead let SAM provide a proposal while the posterior decides where it may be used. Specifically, we threshold $p(x)$ to obtain M_0, then prompt frozen SAM ViT-B with the posterior as a dense mask prompt, its peak point, and the box of M_0 to obtain a single proposal S (Appendix A). We then define $B_r(M_0)$ as a scale-normalized boundary band of width $b(M_0) = \max(2, r\sqrt{|M_0|})$ around the current boundary. Eligible additions and removals are

$$\begin{aligned} A &= \bar{M}_0 \cap S \cap B_r(M_0) \cap \{x : p(x) \geq p_{\mathrm{add}}\}, \\ R &= M_0 \cap \bar{S} \cap B_r(M_0) \cap \{x : p(x) \leq p_{\mathrm{remove}}\}, \end{aligned} \tag{3}$$

where $\bar{M}_0$ and $\bar{S}$ are complements, while p_{add} and p_{remove} are posterior gates. The projected mask is then $M_1 = (M_0 \cup A) \setminus R$.

Three consequences. The intersection rule restricts what SAM may change. A SAM pixel rejected by the posterior ($p(x) < p_{\mathrm{add}}$) cannot be added, so SAM does not determine identity. Likewise, a pixel with $p(x) > p_{\mathrm{remove}}$ cannot be removed, protecting the high-confidence core. Both directions are retained because extent errors are two-sided: undercoverage requires addition, while spill onto a neighbor requires removal. Figure 3 illustrates both directions on frozen validation examples,

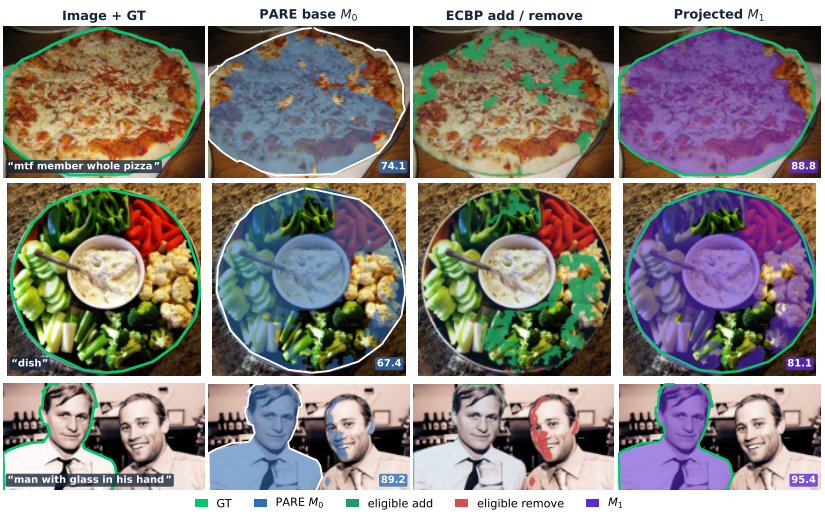

Figure 3: ECBP edits on three RefCOCO val cases from the frozen single-dataset chain. Columns: image/ground truth (green), M_0 (blue), eligible additions A (green)/removals R (red), and M_1 (purple). Some images are cropped to highlight segmentation details.

where eligible additions recover missing extent and eligible removals suppress spill onto neighboring instances. Across all four validation protocols, this rule improves mIoU by $+0.43$ to $+0.49$ pp from each final PARE checkpoint (Table 3). Under the same logits and proposal, Table 4 removes locality and posterior gating in turn, showing that only the complete chain performs best on every metric. The posterior veto is critical for protecting identity: unrestricted whole-mask replacement causes 25 catastrophic failures, the veto-free local rule two, and the vetoed rule none (Appendix F.4).

3.4 Boundary Trust-Region Arbitration

Fixed rules cannot weigh evidence. Both ECBP and BTRA operate near the boundary, but they differ in how edits are admitted. ECBP uses a fixed posterior-gated rule and allows edits only where SAM and the posterior agree. In contrast, boundary refiners that re-predict selected pixels (Kirillov et al., 2020; Yuan et al., 2020) have no default state or predefined edit direction, leaving no explicit way to reject an uncertain edit. BTRA instead learns edit admission, keeping the projected mask by default and applying a local edit only when its score supports the change.

Trust region and edit scores. The editable region consists of two morphological rings around the M_1 boundary, with width $b(M_1) = \mathrm{clip}(\mathrm{round}(0.02\sqrt{|M_1|}), 2, 8)$: an outer ring T^+ for candidate additions and an inner ring T^- for candidate removals. Pixels outside $T^+ \cup T^-$, including the protected mask interior, cannot be modified, and each ring allows only one edit direction. The arbiter therefore decides whether a candidate edit should be applied, rather than which label a pixel should take. At 256×256 resolution, a shallow encoder–decoder takes ten spatial channels (posterior, masks, rings, boundary geometry, and RGB cues) together with frozen SAM features and predicts per-pixel addition and removal logits (Appendix A). For member $m \in \{1, 2, 3\}$ and edit type $c \in \{+, -\}$ with logit $\ell_c^{(m)}(x)$, the ensemble reliability is

$$\rho_c(x) = \frac{1}{3} \sum_{m=1}^{3} \sigma\left(\ell_c^{(m)}(x)\right). \tag{4}$$

The members train separately with AdamW for 24 epochs from different fixed seeds on a binary cross-entropy objective whose labels are the ground truth on T^+ and its complement on T^- (Appendix A).

Default-reject arbitration. An edit is admitted only when its reliability meets or exceeds the corresponding threshold; otherwise the pixel retains its M_1 label:

$$\begin{aligned} A_T &= \{x \in T^+ : \rho_{\mathrm{add}}(x) \geq \theta_{\mathrm{add}}\}, \qquad R_T = \{x \in T^- : \rho_{\mathrm{rem}}(x) \geq \theta_{\mathrm{rem}}\}, \\ M &= (M_1 \cup A_T) \setminus R_T, \end{aligned} \tag{5}$$

with $\rho_{\text{add}} = \rho_+$ and $\rho_{\text{rem}} = \rho_-$. Each threshold controls admission within its designated ring, while rejected candidates retain their M_1 labels. Projection parameters, arbiter weights, and edit thresholds are selected using training expressions or train-only, image-disjoint calibration partitions, and are frozen before official evaluation (Tables 11 and 12; Appendix E). Section 4.3 tests this design against a matched-capacity refiner, unrestricted replacement, and admission controls using the same posterior.

4 EXPERIMENTS

4.1 SETUP

Benchmarks and metrics. We evaluate on RefCOCO, RefCOCO+, and RefCOCOg (G-Ref) under the UMD and Google partitions (Lin et al., 2014; Kazemzadeh et al., 2014; Yu et al., 2016; Mao et al., 2016; Nagaraja et al., 2016); RefCOCO+ removes absolute-location words, G-Ref contains longer expressions, and the Google partition provides only a validation split. mIoU is our primary metric and gives equal weight to each expression. oIoU instead pools pixels across expressions, while Pr@X denotes the fraction of expressions whose IoU exceeds X, for $X \in \{0.5, 0.7, 0.9\}$. Deltas are paired differences in percentage points.

Models and protocol. B3 denotes our TALENT reconstruction with frozen DINOv2-B/14 and CLIP ViT-B/16. E3-RIS-B adds PARE to this baseline, while E3-RIS-L pairs frozen DINOv2-L/14 with CLIP ViT-B/16. Frozen SAM ViT-B provides the proposal and features used by ECBP and BTRA at inference for both model scales. Downstream parameters and thresholds are selected using training expressions or train-only, image-disjoint partitions and are frozen before evaluation on any official split. Checkpoints and posterior thresholds are selected by validation mIoU, while test splits remain held out. Table 8 reports mean$\pm$s.d. over three training seeds on the reduced-budget screen. The three BTRA seeds are ensemble members, and our paired image-cluster bootstrap 95% intervals quantify evaluation uncertainty of frozen models. Mixed training follows the DETRIS mixed protocol with one checkpoint, one ECBP rule, and one BTRA ensemble per scale across all eight splits. Among PET methods that freeze SAM and report official RefCOCO results, we compare with RISAM (frozen SAM ViT-H backbone; oIoU and Pr@X) and Prompt-RIS (one CLIP+SAM model trained on all three training sets, †). Appendix H records all SAM-driven settings cited in our comparisons. Component analyses use the B-scale single-dataset models (Appendix A).

Cost. E3-RIS remains within the PET training budget, with 23.66M trainable parameters compared with 22.78M for TALENT. At inference, E3-RIS and Prompt-RIS use SAM ViT-B; RISAM uses ViT-H. Table 13 lists deployed parameters, latency, and memory for this family.

4.2 MAIN RESULTS

Overlap benchmarks. E3-RIS-B exceeds TALENT and DETRIS-B in mIoU on all nine official splits (Table 1). With the larger backbone, E3-RIS-L achieves the best single-dataset results on all nine splits, including 80.21/81.74/77.49 on RefCOCO val/testA/testB and 73.81 on RefCOCO+ val. Notably, E3-RIS-L also exceeds Prompt-RIS's three-set model on all nine mIoU cells with less training data and no mixed-protocol image reuse. Under mixed training, E3-RIS-L exceeds DETRIS-L on all eight mIoU and oIoU cells (Table 6), while both mixed E3-RIS checkpoints exceed Prompt-RIS on the eight shared mIoU cells. For oIoU, single-dataset E3-RIS-L also exceeds the SAM-based RISAM on all nine splits, reaching 77.71/80.00/74.65 on RefCOCO val/testA/testB.

Tight overlap. Table 2 shows that the system-level gain is concentrated at tight overlap: Pr@0.9 rises from 40.1 to 44.2 over TALENT, and E3-RIS also exceeds both PET systems that use SAM, 39.8 for Prompt-RIS with SAM (mixed) and 42.4 for RISAM (SAM ViT-H). The two Prompt-RIS rows correspond to its CLIP+CMP+ICL model and the same mixed model with SAM; RISAM's Pr@0.5/0.7 come from its default ablation, while its Pr@0.9 matches the official table. On identical PARE logits, unrestricted whole-mask replacement reaches 43.75, whereas the authority rule reaches 44.25, performs best on every column, and avoids the catastrophic losses caused by replacement (Table 4). Section 4.3 provides the full comparison. Figure 4 presents case-by-case results across difficulty bands using the official frozen settings for DETRIS and TALENT and the single-dataset frozen protocol for E3-RIS; Appendix G adds 18 cases across eight scene categories.

Table 1: Benchmark results (%): single-dataset mIoU/oIoU and mixed-training mIoU. u/g: UMD/Google; *: DETRIS mixed protocol; †: Prompt-RIS mixed protocol; unmarked: single-dataset. All E3-RIS rows use frozen SAM ViT-B.

Method	RefCOCO			RefCOCO+			G-Ref		
	val	testA	testB	val	testA	testB	val(u)	test(u)	val(g)
ETOG (Yu et al., 2025a)	73.4	76.9	69.3	66.0	71.5	56.9	63.8	64.6	–
RISCLIP (Kim et al., 2024)	75.7	78.0	72.5	69.2	73.5	60.7	67.6	68.0	–
DETRIS-B (Huang et al., 2025)	76.0	78.2	73.5	68.9	74.0	61.5	67.9	68.1	65.9
DETRIS-L (Huang et al., 2025)	77.3	79.0	75.2	70.8	75.3	64.7	69.3	70.2	67.9
TALENT (Jin et al., 2026)	77.8	79.4	74.8	70.1	74.9	63.4	69.7	69.1	68.4
E3-RIS-B	78.5	80.7	75.6	71.4	75.7	63.8	70.3	71.0	68.6
E3-RIS-L	**80.21**	**81.74**	**77.49**	**73.81**	**77.53**	**66.95**	**71.99**	**72.56**	**70.74**
oIoU, single-dataset training									
BarLeRIa (Wang et al., 2024)	72.4	75.9	68.3	65.0	70.8	56.9	63.4	63.8	61.6
RISAM (Zhang et al., 2023)	76.2	78.9	71.8	66.4	72.1	57.3	65.5	66.6	64.8
TALENT (Jin et al., 2026)	75.9	78.3	72.8	66.9	72.3	58.8	65.9	66.8	64.7
E3-RIS-B	75.7	79.0	72.2	67.0	71.8	57.3	65.7	67.7	63.6
E3-RIS-L	**77.71**	**80.00**	**74.65**	**69.14**	**74.56**	**60.93**	**67.83**	**68.65**	**66.28**
mIoU, mixed training									
Prompt-RIS† (Shang et al., 2024)	78.1	81.2	74.6	71.1	76.6	64.3	70.5	71.3	**69.2**
DETRIS-L* (Huang et al., 2025)	81.0	81.9	79.0	75.2	78.6	70.2	74.6	75.3	–
E3-RIS-B*	81.1	82.6	78.8	74.7	78.7	68.4	75.0	75.6	–
E3-RIS-L*	**82.8**	**84.2**	**80.4**	**77.5**	**80.7**	**72.0**	**77.0**	**77.1**	–

Table 2: Precision@X (%) on RefCOCO val; our row is E3-RIS-B.

Method	Pr@0.5	Pr@0.7	Pr@0.9
ETRIS (Xu et al., 2023)	83.4	72.7	17.4
ETOG (Yu et al., 2025a)	83.2	73.0	26.2
CGFormer (Tang et al., 2023)	87.2	78.7	38.8
RISAM (Zhang et al., 2023)	85.8	80.4	42.4
Prompt-RIS w/o SAM (Shang et al., 2024)	85.6	76.9	26.2
Prompt-RIS (Shang et al., 2024)	88.5	82.1	39.8
DETRIS (Huang et al., 2025)	87.9	80.2	27.5
TALENT (Jin et al., 2026)	**88.6**	82.3	40.1
E3-RIS	88.4	**82.7**	**44.2**

4.3 WHERE THE GAIN COMES FROM

Stage-wise gains. Table 3 compares ECBP and BTRA with their inputs across four validation protocols, while Table 7 compares PARE with B3 (Appendix D). On RefCOCO val, PARE improves mIoU from 77.46 to 77.80 ($+0.34$) and Pr@0.9 from 37.30 to 39.42 before any SAM edit. ECBP then adds $+0.43$ to $+0.49$ mIoU, and BTRA another $+0.22$ to $+0.30$ while rejecting 92.5% of candidate pixels. The stage-gain intervals exclude zero, and BTRA improves both mIoU and oIoU on all nine official splits (Table 10; Appendix F).

What the evidence stage changes. The fixed-threshold experiment in Table 8 separates changes in target discrimination from coverage. With the same 25% training subset, five epochs, and $t = 0.35$, PARE reduces same-category leakage by 1.30 ± 0.38 pp and background false positives by 1.09 ± 0.63 pp over three seeds, while completeness falls by 1.65 ± 0.70 pp. mIoU nevertheless rises by 1.33 ± 0.27 pp. This reduced-budget result reveals the trade-off motivating ECBP: suppressing non-target pixels improves overlap but leaves missing target extent to a separate operator.

Authority versus access and capacity. Table 4 compares paired predictions while varying access to the same posterior and proposal. Whole-mask replacement gains $+0.42$ mIoU and $+4.33$ Pr@0.9 over the posterior but changes oIoU by only $+0.01$ pp and causes 25 expressions to drop by at

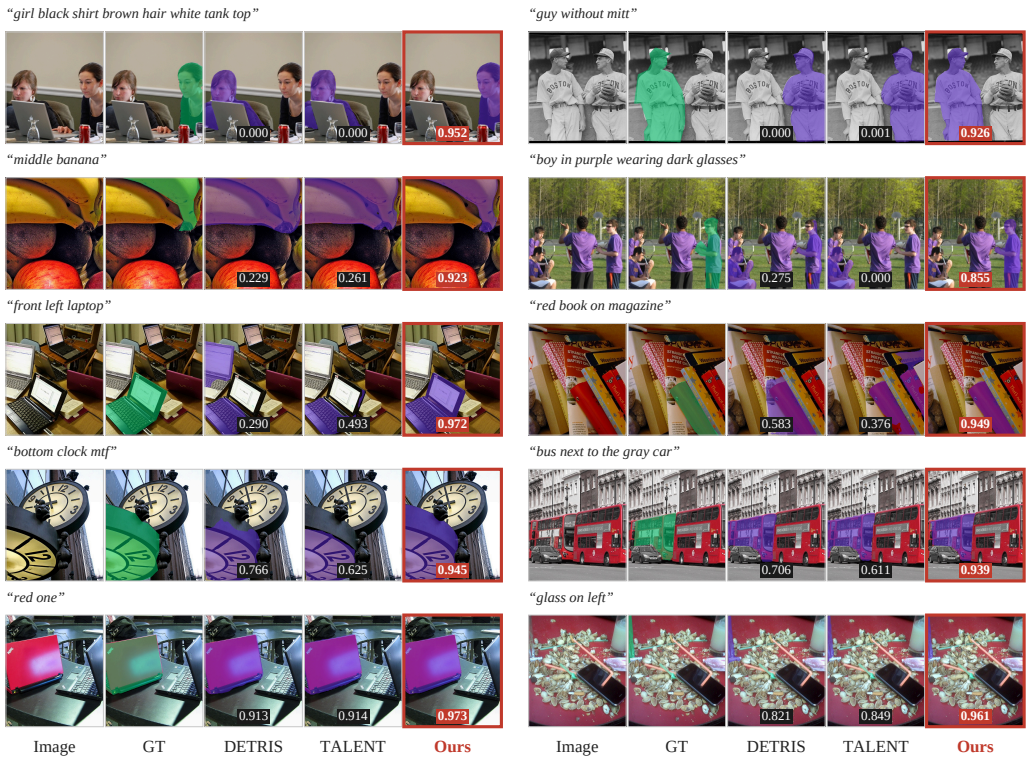

Figure 4: Difficulty-stratified qualitative comparison on RefCOCO val, two cases per stronger-baseline IoU band. Green is ground truth and purple is the prediction; each prediction prints its per-expression IoU. Some images are cropped to highlight segmentation details.

Table 3: B-scale single-dataset validation chain (mIoU, %). Δ_E/Δ_B: successive ECBP/BTRA gains (pp); CIs: paired image-cluster bootstrap 95% intervals for frozen models.

Dataset	PARE	+ECBP	Δ_E	95% CI	+BTRA	Δ_B	95% CI
RefCOCO	77.80	78.23	+0.43	$[+0.41, +0.46]$	**78.45**	+0.22	$[+0.20, +0.24]$
RefCOCO+	70.73	71.16	+0.43	$[+0.40, +0.46]$	**71.38**	+0.23	$[+0.21, +0.25]$
G-Ref (UMD)	69.56	70.05	+0.49	$[+0.45, +0.53]$	**70.29**	+0.24	$[+0.21, +0.28]$
G-Ref (Google)	67.83	68.29	+0.46	$[+0.43, +0.49]$	**68.59**	+0.30	$[+0.28, +0.33]$

least 0.20 IoU, which we call catastrophic. Because oIoU pools pixels across expressions, gains on many cases can be offset by failures where a wrong proposal replaces the referent. Removing the posterior veto produces two catastrophic drops, whereas ECBP produces none, showing that the veto keeps identity control with the posterior. Adding BTRA on top of ECBP then exceeds the veto-free variant on all five metrics and is the only setting that performs best on every metric. The generic refiner, matched to BTRA in everything except action semantics (168,978 parameters per member; Appendix F.2), leaves ECBP mIoU unchanged at 78.23, showing that matched capacity alone does not explain the gain. Keeping the same arbiter but setting both thresholds to zero admits every ring candidate and lowers mIoU to 72.47 and Pr@0.9 to 3.64 (Appendix F.3), confirming the importance of default rejection. Finally, on the 10,522 RefCOCO val expressions with a same-category competitor, the full chain gains +0.66 pp mIoU over PARE with no catastrophic drops (Appendix F.4).

Selecting edits at matched coverage. Table 5 matches the accepted addition and removal counts of BTRA for each expression, so every control edits the same number of pixels. Posterior direction ranks additions by p and removals by $1 - p$; SAM direction ranks proposal membership and non-membership, with posterior tie-breaking; image edge uses Sobel magnitude. Random averages five fixed seeds. BTRA exceeds the strongest simple ranking, SAM direction, by 0.21 mIoU and 0.16

Table 4: Authority ablation (%) on all 10,834 RefCOCO val expressions, using identical PARE logits and SAM proposals. All settings are frozen.

Method	mIoU	oIoU	Pr@0.5	Pr@0.7	Pr@0.9
PARE posterior only (M_0)	77.80	75.31	88.19	82.23	39.42
Whole-mask SAM replacement	78.22	75.32	88.22	82.57	43.75
Local bidirectional, no posterior veto	78.33	75.48	88.31	82.54	44.03
ECBP (local, vetoed)	78.23	75.51	88.31	82.55	42.63
ECBP + generic refiner (matched capacity)	78.23	75.49	88.31	82.57	42.94
ECBP + BTRA (complete)	**78.45**	**75.67**	**88.38**	**82.69**	**44.25**

Table 5: Matched-coverage controls on RefCOCO val (%). Addition/removal counts match frozen BTRA per expression and direction. Correct: pixel-weighted fraction of edits agreeing with ground truth; random: five-seed mean.

Admission score	mIoU	oIoU	Accepted correct (%)
Random	77.74	75.12	35.81
Image edge	77.83	75.20	38.84
Posterior direction	77.92	75.35	45.31
SAM direction	78.24	75.51	51.95
BTRA	**78.45**	**75.67**	**58.30**

oIoU, while raises accepted-edit correctness from 51.95 to 58.30%. The gain therefore depends on selecting useful pixels within the same edit budget.

Transfer without downstream retuning. Table 9 applies the RefCOCO input threshold, ECBP rule, BTRA weights, and edit thresholds unchanged to PARE logits on RefCOCO+. The transferred chain improves mIoU/oIoU by $+0.65/+0.33$ pp on val and by at least $+0.53/+0.22$ pp on testA and testB over the source-threshold posterior. Its mIoU is within 0.03 pp of the independently fitted target policy on every split. The 1,546 fitting and 168 calibration images are disjoint from all RefCOCO+ evaluation images. Thus, the learned downstream policy transfers across the annotation shift while the upstream evidence model remains target-specific.

Complementary directions and prompt control. With scores and thresholds fixed, addition alone contributes $+0.07$ pp mIoU over ECBP, removal alone $+0.15$, and combining both yields $+0.22$ (Table 19). Addition increases completeness by 0.280 pp, while removal reduces commission by 0.493 pp, showing that the two directions address different mask errors. Dropping the dense SAM prompt while keeping the peak point and M_0 box still gives ECBP a $+0.37$ pp mIoU gain, followed by another $+0.24$ from a refit BTRA (Table 17). The chain reaches 78.41 mIoU, within 0.04 of the dense-prompt chain, showing that the gain also holds with point-and-box prompting.

Local sensitivity. Post-hoc perturbations around the frozen policy reveal a stable local plateau (Table 15). The maximum mIoU change is only 0.05 pp for 20% changes in ECBP thresholds or band width, 0.03 pp for 10% changes in BTRA thresholds, and 0.01 pp for 20% changes in trust-region scale. No validation setting is reselected.

5 Conclusion

E3-RIS separates referential evidence, object extent, and boundary trust while retaining a single posterior. PARE pools prompt-anchored queries to identify the referent, ECBP constrains frozen SAM to recover object extent through posterior-compatible local edits, and BTRA admits only trusted direction-fixed boundary edits while otherwise preserving the projected mask. Extensive experiments on RefCOCO, RefCOCO+, and G-Ref demonstrate that E3-RIS achieves highly competitive performance among current parameter-efficient RIS methods.

AI USE STATEMENT

We used generative AI tools to polish the language of the manuscript. They were not used to generate experimental numbers, invent citations, or create qualitative results. All AI-assisted text was reviewed and verified by the authors, who take responsibility for the final content.

ETHICS STATEMENT

RIS drives language-guided image editing and robot instruction following, and E3-RIS makes the authority over each decision explicit at inference. All experiments use the public COCO-derived RefCOCO, RefCOCO+, and RefCOCOg benchmarks, and no new data were collected.

REPRODUCIBILITY STATEMENT

The three mechanisms are specified in Section 3. Training recipes, prompt strings, adapter placement, architecture, and loss details are in Appendix A; the frozen inference hyperparameters are in Tables 11 and 12; the selection and image-disjoint confirmation protocol is in Appendix E; metric and diagnostic definitions are in Appendix B. Stage gains of frozen models (Tables 3, 10, and the generic-refiner contrast in Appendix F.2) carry paired image-cluster bootstrap intervals; upstream training uses seed 0, Table 8 reports mean$\pm$s.d. over three training seeds for every metric, and the three BTRA seeds are ensemble members. For single-dataset B-scale models, the posterior operating points are selected by validation mIoU over a fixed grid, listed in Table 11, and frozen before test reporting, so test labels never enter selection. Code and the frozen selection manifests will be released with the camera-ready version.

REFERENCES

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1–9, 2022. doi: 10.18653/v1/2022.acl-short.1.

Bo Chen, Zhiwei Hu, Zhilong Ji, Jinfeng Bai, and Wangmeng Zuo. Position-aware contrastive alignment for referring image segmentation. *arXiv preprint arXiv:2212.13419*, 2022a.

Shoufa Chen, Chongjian GE, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. In *Advances in Neural Information Processing Systems*, volume 35, pp. 16664–16678, 2022b. doi: 10.52202/068431-1212.

Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. In *International Conference on Learning Representations*, 2023.

Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C. Berg, and Alexander Kirillov. Boundary iou: Improving object-centric image segmentation evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15334–15342, 2021.

Yong Xien Chng, Henry Zheng, Yizeng Han, Xuchong Qiu, and Gao Huang. Mask grounding for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26573–26583, 2024.

Ming Dai, Wenxuan Cheng, Jiang-jiang Liu, Sen Yang, Wenxiao Cai, Yanpeng Sun, and Wankou Yang. DeRIS: Decoupling perception and cognition for enhanced referring image segmentation through loopback synergy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19936–19946, 2025.

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

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. CLIP-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2023. doi: 10.1007/s11263-023-01891-x.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 2151–2159, 2019.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 1321–1330, 2017.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 2790–2799. PMLR, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *European Conference on Computer Vision*, pp. 108–124, 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*, pp. 4067–4077, 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. doi: 10.1609/aaai.v39i4.32380.

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.

Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6409–6418, 2019.

Koichiro Ito. Feature design for bridging SAM and CLIP toward referring image segmentation. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 8357–8367, 2025.

Lixia Ji, Yunlong Du, Yiping Dang, Wenzhao Gao, and Han Zhang. A survey of methods for addressing the challenges of referring image segmentation. *Neurocomputing*, 583:127599, 2024. doi: 10.1016/j.neucom.2024.127599.

Shibo Jie, Zhi-Hong Deng, Shixuan Chen, and Zhijuan Jin. Convolutional bypasses are better vision transformer adapters. In *European Conference on Artificial Intelligence*, pp. 202–209, 2024. doi: 10.3233/FAIA240489.

Shuo Jin, Siyue Yu, Bingfeng Zhang, Chao Yao, Meiqin Liu, and Jimin Xiao. TALENT: Target-aware efficient tuning for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 7472–7482, 2026.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. MDETR – modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1780–1790, 2021.

Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. Compacter: Efficient low-rank hypercomplex adapter layers. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1022–1035, 2021.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 787–798, 2014. doi: 10.3115/v1/D14-1086.

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.

Seoyeon Kim, Minguk Kang, Dongwon Kim, Jaesik Park, and Suha Kwak. Extending CLIP's image-text alignment to referring image segmentation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4611–4628, 2024. doi: 10.18653/v1/2024.naacl-long.258.

Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9799–9808, 2020.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. 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.

Hongjun Li, Nan Wang, Feng Gu, Jiang Liu, Hantao Liu, Paul L. Rosin, Wanli Ma, Wei Zhou, and Qijun Yang. E-ReferSAM: Enhanced ReferSAM with fine-grained alignment and boundary-aware fusion for referring image segmentation. *SSRN preprint*, 2026. doi: 10.2139/ssrn.6760539.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 12888–12900. PMLR, 2022a.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10965–10975, 2022b.

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.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597, 2021. doi: 10.18653/v1/2021.acl-long.353.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014.

Chang Liu, Henghui Ding, and Xudong Jiang. GRES: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23592–23601, 2023a.

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.

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

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. doi: 10.1109/TPAMI.2021.3079993.

Sun-Ao Liu, Yiheng Zhang, Zhaofan Qiu, Hongtao Xie, Yongdong Zhang, and Ting Yao. CARIS: Context-aware referring image segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 779–788, 2023c.

Sun-Ao Liu, Hongtao Xie, Jiannan Ge, and Yongdong Zhang. ReferSAM: Unleashing segment anything model for referring image segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(5):4910–4922, 2025. doi: 10.1109/TCSVT.2024.3524543.

Ting Liu, Xuyang Liu, Siteng Huang, Honggang Chen, Quanjun Yin, Long Qin, Donglin Wang, and Yue Hu. DARA: Domain- and relation-aware adapters make parameter-efficient tuning for visual grounding. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2024a. doi: 10.1109/ICME57554.2024.10688132.

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, 2024b.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, 2015.

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

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

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.

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

Hai Nguyen-Truong, E-Ro Nguyen, Tuan-Anh Vu, Minh-Triet Tran, Binh-Son Hua, and Sai-Kit Yeung. Vision-aware text features in referring image segmentation: From object understanding to context understanding. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 4988–4998, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763. PMLR, 2021.

Nisarg A. Shah, Vibashan VS, and Vishal M. Patel. LQMFormer: Language-aware query mask transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12903–12913, 2024.

Chao Shang, Zichen Song, Heqian Qiu, Lanxiao Wang, Fanman Meng, and Hongliang Li. Prompt-driven referring image segmentation with instance contrasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4124–4134, 2024.

Jiajin Tang, Ge Zheng, Cheng Shi, and Sibei Yang. Contrastive grouping with transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23570–23580, 2023.

Yaoming Wang, Jin Li, Xiaopeng Zhang, Bowen Shi, Chenglin Li, Wenrui Dai, Hongkai Xiong, and Qi Tian. BarLeRIa: An efficient tuning framework for referring image segmentation. In *International Conference on Learning Representations*, 2024.

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.

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.

Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15325–15336, 2023.

Yuhuan Yang, Chaofan Ma, Jiangchao Yao, Zhun Zhong, Ya Zhang, and Yanfeng Wang. ReMamber: Referring image segmentation with mamba twister. In *European Conference on Computer Vision*, pp. 108–126, 2024.

Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip H. S. 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.

Houjian Yu, Mingen Li, Alireza Rezazadeh, Yang Yang, and Changhyun Choi. A parameter-efficient tuning framework for language-guided object grounding and robot grasping. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 14353–14360, 2025a. doi: 10.1109/ICRA55743.2025.11128679.

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, Joonbeom Hong, Joonseok Lee, and Jeany Son. Latent expression generation for referring image segmentation and grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21374–21383, 2025b.

Yuhui Yuan, Jingyi Xie, Xilin Chen, and Jingdong Wang. Segfix: Model-agnostic boundary refinement for segmentation. In *European Conference on Computer Vision*, pp. 489–506, 2020. doi: 10.1007/978-3-030-58610-2_29.

Mengxi Zhang, Yiming Liu, Xiangjun Yin, Huanjing Yue, and Jingyu Yang. RISAM: Referring image segmentation via mutual-aware attention features. *arXiv preprint arXiv:2311.15727*, 2023.

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*, 2024.

Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5729–5739, 2023.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

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. In *Advances in Neural Information Processing Systems*, volume 36, pp. 19769–19782, 2023. doi: 10.52202/075280-0868.

A Implementation details for reproduction

B-scale B3 and PARE. *B3* denotes our TALENT reconstruction: frozen DINOv2-B/14 and CLIP ViT-B/16 encoders, 448^2 input, 128-dimensional visual adapters and 64-dimensional text adapters at layers $1, 3, 5, 7, 9, 11$, RCA with weighted-text aggregation, and the reported CPCL and mask-aware TCCL terms (both weighted 0.1). Its segmentation objective is pixel BCE plus soft IoU loss; B3 contains no PARE, SAM, ECBP, or BTRA component. PARE appends four 512-dimensional trainable queries, each initialized once from the frozen CLIP text embedding of a distinct fixed prompt: "the target object", "the distinguishing visual attribute", "the spatial or semantic relation", and "the reference object or anchor" for q_1–q_4. The strings are initializers only; no learned role is assigned to any query, and q_1, q_2 are the two maps that receive $\mathcal{L}_{\mathrm{loc}}$. $\mathcal{L}_{\mathrm{anc}}$ keeps each trained query near its initializer. Token attention uses the 15 content positions of the length-17 CLIP sequence after masking padding and start/end tokens. $W_k : \mathbb{R}^{512} \to \mathbb{R}^{128}$ and $W_v : \mathbb{R}^{512} \to \mathbb{R}^{128}$ are bias-free linear/1×1 projections. The residual map ϕ is $644 \to 128$ bias-free 1×1 convolution, GroupNorm(32), GELU, then a bias-free $128 \to 512$ 1×1 convolution; its final kernel is initialized from $\mathcal{N}(0, 10^{-6})$ variance and scalar γ from 0.05. Adam uses learning rate 10^{-4}, zero weight decay, global batch 32, and a 0.1 learning-rate multiplier at epochs 35 and 55 (epoch 35 only for the 50-epoch Google schedule). Reported upstream runs use seed 0, deterministic aspect-preserving affine resize with mean padding, and one uniformly sampled training expression per referent; there is no random color or geometric image augmentation. We validate every epoch after 128-batch BN recalibration and retain the highest validation-mIoU checkpoint.

PARE regularizers. Two regularizers accompany the containment loss of Section 3. Let $\hat{a}_k = a_k / \|a_k\|_2$ be the normalized content-token attention of query k, and let q_k^0 be its frozen initializer. The regularizers are

$$\mathcal{L}_{\mathrm{div}} = \frac{1}{K(K-1)} \sum_{k \neq j} \hat{a}_k^\top \hat{a}_j, \qquad \mathcal{L}_{\mathrm{anc}} = \frac{1}{K} \sum_k \left(1 - \cos(q_k, q_k^0)\right). \tag{6}$$

We optimize $0.1\mathcal{L}_{\mathrm{loc}} + 0.01\mathcal{L}_{\mathrm{div}} + 0.01\mathcal{L}_{\mathrm{anc}}$ in addition to the base segmentation loss. Diversity is an anti-collapse regularizer: it keeps the four queries from becoming copies of one another. Anchoring keeps each query near its prompt initializer, which is the sense in which the evidence is prompt-anchored.

B-scale mixed PARE. Mixed E3-RIS-B uses the same encoders, adapters, prompts, and evidence losses as the single-dataset B-scale model. It concatenates the RefCOCO, RefCOCO+, and G-Ref UMD training sets and trains with Adam at learning rate 10^{-4} and global batch 32 for 65 epochs, multiplying the learning rate by 0.1 at epochs 35 and 55. We evaluate after 128-batch BN recalibration each epoch and retain the checkpoint with the highest mixed-validation mIoU.

L-scale PARE. E3-RIS-L uses frozen DINOv2-L/14 and CLIP ViT-B/16, with 384-dimensional visual adapters at layers 1–23 and 384-dimensional text adapters at layers 1–11. The FPN reads DINOv2 layers 8 and 16 at width 1024; the ladder width is 64. The four prompt initializers and the evidence-loss weights match the B-scale PARE recipe; the sequence length is 17, the input

size is 448^2, and the seed is 0. Single-dataset PARE-L trains with Adam at learning rate 10^{-4} and global batch 32 for 50 epochs, multiplying the learning rate by 0.1 at epoch 35, and retains the highest validation-mIoU checkpoint on that dataset. Mixed PARE-L trains with Adam at learning rate 2×10^{-4} and global batch 64 for 50 epochs, multiplying the learning rate by 0.1 at epoch 35. We evaluate after 128-batch BN recalibration each epoch and retain the checkpoint with the highest mixed-validation mIoU.

The L-scale downstream pipeline uses frozen SAM ViT-B. Its ECBP rule is selected on the pooled training union, and its three BTRA members use seeds 6407/6408/6409 and the 24-epoch AdamW schedule described below. Edit thresholds are selected on an image-disjoint calibration partition of the pooled training union; the posterior threshold uses mixed validation. Table 12 gives the resulting settings, applied unchanged across all eight splits.

SAM and morphology. For nonempty M_0, the positive point is $\arg\max_x p(x)$ and the box is the inclusive min/max rectangle of M_0. The dense mask prompt is $\operatorname{logit} p - \operatorname{logit} t$, resized with SAM's longest side to 1024, bottom/right padded with -10, area-resized to 256^2, and clipped to $[-10, 10]$. We call official SAM ViT-B with the point, box, and dense mask and set `multimask_output=False`; hence there is one proposal and no score-based proposal selection. If M_0 is empty, $S = M_0$ and ECBP makes no edit. ECBP uses OpenCV L2 distance transforms with a 5×5 mask: outside distance for additions, inside distance for removals, and band radius $\max(2, r\sqrt{|M_0|})$ in original-image pixels. BTRA instead applies an elliptical morphological kernel of integer radius $\operatorname{clip}(\operatorname{round}(0.02\sqrt{|M_1|}), 2, 8)$; dilation minus M_1 defines T^+ and M_1 minus erosion defines T^-. Thus BTRA may learn new local corrections relative to ECBP, but cannot change any pixel outside these direction-fixed rings.

BTRA input channels. At 256×256 resolution, the arbiter receives posterior, threshold margin, posterior gradient, M_1, SAM proposal, T^+, T^-, signed distance from M_0, luminance, and image edges. A frozen $256 \times 64 \times 64$ SAM embedding is projected to 16 channels by 1×1 convolution–GroupNorm–GELU–3×3 convolution.

BTRA optimization. All spatial channels are constructed at 256^2. Posterior, luminance, and binary mask/ring channels lie in $[0, 1]$; luminance is $0.299R + 0.587G + 0.114B$ after RGB division by 255. Threshold margin is $\operatorname{clip}((p - t)/0.5, -1, 1)$. Posterior gradient and image edge are OpenCV 3×3 Sobel magnitudes divided by four and clipped to $[0, 1]$. The distance channel retains the geometry of the initial mask: it is $\operatorname{clip}((d_{\text{in}} - d_{\text{out}})/\sqrt{\max(|M_0|, 1)}, -1, 1)$, with area and OpenCV L2 distances computed from M_0 at 256^2 using mask size five. The ten spatial channels and the 16-channel SAM projection are followed by two 3×3 24-channel blocks, stride-2 $24 \to 48$, another 48-channel block, dilated $48 \to 64 \to 64 \to 48$ context with dilation $2/4/2$, bilinear upsampling, a $72 \to 32 \to 24$ fusion path, and a two-logit 1×1 head. Convolutions use GroupNorm and GELU except the output. The three seeds are 6407/6408/6409. Each member uses AdamW, learning rate 10^{-3}, weight decay 10^{-3}, batch 32, AMP, 24 epochs, and no horizontal flip because cached SAM features retain absolute position. For single-dataset B-scale models, final refit/calibration expression counts are respectively 11,045/1,243 for RefCOCO, 11,108/1,180 for RefCOCO+, 10,981/1,225 for G-Ref UMD, and 10,975/1,234 for G-Ref Google. Splits are grouped by COCO image with seed 7407; 82/79 extra G-Ref UMD/Google rows sharing calibration images are excluded from fitting. The 0.02 threshold grid maximizes calibration mIoU subject to at least $+0.10$ pp calibration oIoU; official splits are read only after freezing.

BTRA training loss. Addition labels are $y_+(x) = Y(x)$ on T^+; removal labels are $y_-(x) = 1 - Y(x)$ on T^-. Let $\mathcal{T}_{ic}$ contain the candidate pixels for expression i and $\mathcal{I}_c$ the expressions with

nonempty $\mathcal{T}_{ic}$. Each member is trained with

$$L_c^{\text{expr}} = \frac{1}{|\mathcal{I}_c|} \sum_{i \in \mathcal{I}_c} \frac{1}{|\mathcal{T}_{ic}|} \sum_{x \in \mathcal{T}_{ic}} \text{BCE}(\ell_{ic}(x), y_{ic}(x)), \tag{7}$$

$$L_c^{\text{pool}} = \frac{\sum_i \sum_{x \in \mathcal{T}_{ic}} \text{BCE}(\ell_{ic}(x), y_{ic}(x))}{\sum_i |\mathcal{T}_{ic}|}, \tag{8}$$

$$\mathcal{L}_{\text{BTRA}} = \frac{1}{2} \sum_{c \in \{+,-\}} \left[\lambda L_c^{\text{expr}} + (1-\lambda) L_c^{\text{pool}}\right], \quad \lambda = 0.5. \tag{9}$$

The expression term prevents large masks from dominating, while the pooled term retains pixel-level statistical efficiency. Ground truth supervises the edit labels only during training; at inference the frozen ensemble reliability of Section 3 and the frozen thresholds decide every edit.

Mixed-protocol image overlap. Mixed training follows the DETRIS mixed protocol over the 28,158 unique COCO images of the three training sources. Same-dataset train/evaluation image overlap is zero for every official split.

Dataset sizes. RefCOCO has 10,834/5,657/5,095 val/testA/testB expressions and RefCOCO+ has 10,758/5,726/4,889. G-Ref UMD has 4,896 validation and 9,602 test expressions; Google has 9,536 validation expressions.

B METRIC DEFINITIONS AND DIAGNOSTIC SCOPE

For prediction P_i and nonempty target Y_i, let $u_i = |P_i \cup Y_i|$ and $z_i = |P_i \cap Y_i|$. The public benchmark metrics are $\text{mIoU} = N^{-1} \sum_i z_i/u_i$, $\text{oIoU} = \sum_i z_i / \sum_i u_i$, and $\Pr@X = N^{-1} \sum_i \mathbf{1}[z_i/u_i > X]$. Thus mIoU weights expressions equally, oIoU pools their pixel mass, and $\Pr@X$ is the fraction of expressions whose IoU strictly exceeds the threshold. The following diagnostics separate target coverage, contamination, and distractor errors. Let D_i be the union of all annotated non-target instances of the target's category in that image. We compute

$$\text{Comp}_i = \frac{|P_i \cap Y_i|}{|Y_i|}, \qquad \text{Comm}_i = \frac{|P_i \setminus Y_i|}{|P_i|}, \tag{10}$$

$$\text{Leak}_i = \frac{|P_i \cap D_i|}{|P_i|}, \qquad \text{BFP}_i = \frac{|P_i \cap B_i|}{|P_i|}. \tag{11}$$

where B_i is the complement of $Y_i \cup D_i$. Reported values are expression-level means (zero predicted area contributes zero to commission, leakage, and background false positives). *Foreground completeness* measures target-pixel recall. *Commission* measures the fraction of predicted foreground outside the target. *Distractor leakage* attributes part of that contamination specifically to same-category competing instances. *Background false positives* are the remaining share of the prediction, the pixels on neither the target nor a same-category distractor; the denominator is predicted area, not the number of background pixels. Together, completeness and commission distinguish omission from contamination.

The *same-category-competitor set* consists of the expressions whose D_i contains at least one instance; it covers 97% of RefCOCO val. On this subset, a *catastrophic drop* is a stage-wise decrease of at least 0.20 absolute IoU from the frozen PARE mask; it is evaluated post hoc.

Boundary IoU compares the internal boundary bands of P_i and Y_i. A band is $B_w(M) = M \setminus \text{erode}_w(M)$ with $w = \max(1, \text{round}(0.02\sqrt{H^2 + W^2}))$ pixels; thus $\text{BIoU}_i = |B_w(P_i) \cap B_w(Y_i)|/|B_w(P_i) \cup B_w(Y_i)|$. It is included because mIoU can hide contour changes on large objects, whereas ECBP and BTRA act locally near the boundary.

For BTRA, the *rejected-pixel fraction* is one minus the number of candidate addition/removal pixels admitted by the arbiter divided by the total number of candidate pixels; it is pixel-weighted, not the fraction of expressions or connected components rejected.

C oIoU COMPARISON

Table 6 reports single-dataset oIoU at both scales and mixed-training oIoU at both scales. Because $\text{oIoU} = \sum_i u_i \text{IoU}_i / \sum_i u_i$, it is exactly a union-area-weighted mean of expression IoUs: one error on a large referent can outweigh improvements on many small referents. This differs from mIoU, which assigns one vote to every expression, so the two metrics answer different questions. In single-dataset training, E3-RIS-L records 77.71/80.00/74.65 on RefCOCO val/testA/testB ($+1.81/+1.70/+1.85$ over TALENT) and 69.14/74.56/60.93 on RefCOCO+, and exceeds TALENT and the SAM-driven RISAM (Zhang et al., 2023) on all nine official cells. E3-RIS-B exceeds TALENT on RefCOCO testA, RefCOCO+ val, and G-Ref test(u) by 0.7, 0.1, and 0.9 percentage points, respectively. Under mixed training, E3-RIS-L reaches 81.5/83.8/78.3 on RefCOCO, 74.3/78.5/67.3 on RefCOCO+, and 74.4/74.8 on G-Ref UMD, exceeding mixed DETRIS-L on all eight cells by 0.1–1.6 points (mean 0.95). RISAM accumulates intersection and union over test samples, corresponding to oIoU. Prompt-RIS uses one model trained on the three-set union with evaluation images removed. Both E3-RIS scales use a PET encoder and frozen SAM ViT-B; single-dataset L-scale values retain two decimals as logged. Table 1 gives the mIoU comparison.

Table 6: oIoU (%) on RefCOCO, RefCOCO+, and G-Ref. u/g: UMD/Google; *: DETRIS mixed protocol; †: Prompt-RIS mixed protocol; unmarked: single-dataset. –: unavailable.

Method	RefCOCO			RefCOCO+			G-Ref		
	val	testA	testB	val	testA	testB	val(u)	test(u)	val(g)
ReSTR (Kim et al., 2022)	67.2	69.3	64.5	55.8	60.4	48.3	54.5	–	54.5
CRIS (Wang et al., 2022)	70.5	73.2	66.1	62.3	68.1	53.7	59.9	60.4	–
LAVT (Yang et al., 2022)	72.7	75.8	68.8	62.1	68.4	55.1	61.2	62.1	60.5
VPD (Zhao et al., 2023)	73.5	–	–	63.9	–	–	63.1	–	–
DMMI (Hu et al., 2023)	74.1	77.1	70.2	64.0	69.7	57.0	63.5	64.2	62.0
ReLA (Liu et al., 2023a)	73.8	76.5	70.2	66.0	71.0	57.7	65.0	66.0	62.7
CGFormer (Tang et al., 2023)	74.8	77.3	70.6	64.5	71.0	57.1	64.7	65.1	62.5
MagNet (Chng et al., 2024)	75.2	78.2	71.1	66.2	71.3	58.1	65.4	66.0	63.1
LQMFormer (Shah et al., 2024)	74.2	76.8	71.0	65.9	71.8	57.6	64.7	66.0	63.0
ReMamber (Yang et al., 2024)	74.5	76.7	70.9	65.0	70.8	57.5	63.9	64.0	–
CoHD (Luo et al., 2025)	75.1	78.3	71.0	66.8	71.6	58.4	65.8	66.7	–
ETRIS (Xu et al., 2023)	70.5	73.5	66.6	60.1	66.9	50.2	59.8	59.9	57.9
RISCLIP (Kim et al., 2024)	73.6	76.5	69.8	65.5	70.6	55.5	64.1	65.1	–
ETOG (Yu et al., 2025a)	71.4	76.1	66.7	62.3	68.5	51.9	61.1	62.9	–
BarLeRIa (Wang et al., 2024)	72.4	75.9	68.3	65.0	70.8	56.9	63.4	63.8	61.6
RISAM (Zhang et al., 2023)	76.2	78.9	71.8	66.4	72.1	57.3	65.5	66.6	64.8
TALENT (Jin et al., 2026)	75.9	78.3	72.8	66.9	72.3	58.8	65.9	66.8	64.7
E3-RIS-B	75.7	79.0	72.2	67.0	71.8	57.3	65.7	67.7	63.6
E3-RIS-L	**77.71**	**80.00**	**74.65**	**69.14**	**74.56**	**60.93**	**67.83**	**68.65**	**66.28**
Mixed training									
PolyFormer-L* (Liu et al., 2023b)	76.0	78.3	73.3	69.3	74.6	61.9	69.2	70.2	–
UNINEXT-L* (Yan et al., 2023)	80.3	82.6	77.8	70.0	74.9	62.6	73.4	73.7	–
Prompt-RIS† (Shang et al., 2024)	76.4	80.4	72.3	67.1	73.6	59.0	67.2	69.0	**64.8**
DETRIS-L* (Huang et al., 2025)	80.7	82.2	77.9	73.1	77.5	66.0	73.2	74.7	–
E3-RIS-B*	79.6	81.4	76.7	70.9	76.0	62.3	71.6	73.3	–
E3-RIS-L*	**81.5**	**83.8**	**78.3**	**74.3**	**78.5**	**67.3**	**74.4**	**74.8**	–

D COMPONENT TABLES

This appendix holds the per-stage tables behind Section 4.3: the full-budget PARE-versus-B3 comparison on RefCOCO val (Table 7), the reduced-budget mechanism screen (Table 8), the PARE query maps (Figure 5), the RefCOCO-to-RefCOCO+ policy transfer (Table 9), and the frozen

BTRA gains on all nine official splits (Table 10). Every row of Table 8 averages three independent 25%/five-epoch trainings.

Table 7: Full-data, 65-epoch RefCOCO evidence-stage comparison on all 10,834 validation expressions.

Metric	B3	PARE	Δ
mIoU	77.46	77.80	**+0.34**
oIoU	75.15	75.31	**+0.16**
Pr@0.5	88.17	88.19	**+0.02**
Pr@0.7	81.94	82.23	**+0.29**
Pr@0.9	37.30	39.42	**+2.12**

Table 8: Reduced-budget mechanism screen on RefCOCO val (25% training subset, five epochs, fixed $t = 0.35$, all 10,834 expressions). Mean±s.d. over three training seeds $(0, 892, 940)$.

Metric	B3	PARE	Δ
mIoU	62.58 ± 1.21	63.90 ± 1.11	$+1.33 \pm 0.27$
oIoU	58.20 ± 1.26	60.17 ± 1.21	$+1.98 \pm 0.21$
Pr@0.9	12.82 ± 0.47	15.57 ± 2.20	$+2.75 \pm 1.95$
Completeness	79.25 ± 0.46	77.59 ± 0.39	-1.65 ± 0.70
Distractor leakage	16.18 ± 1.33	14.88 ± 1.55	-1.30 ± 0.38
Background FP	16.51 ± 2.26	15.42 ± 2.75	-1.09 ± 0.63
Boundary IoU	41.62 ± 0.69	43.58 ± 1.82	$+1.96 \pm 1.18$

Figure 5 shows the four PARE query maps and their content-token attention on the third Figure 3 case. Query q_2 attends to *glass* (0.95) and q_3 to *man* (0.93), illustrating distinct token responses within the shared evidence representation. Queries q_1, q_2 receive $\mathcal{L}_{\text{loc}}$; q_3, q_4 do not. All four maps feed one mask head.

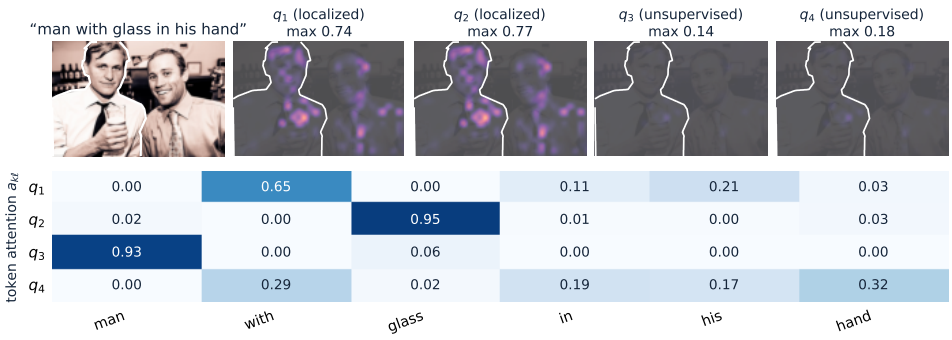

Figure 5: PARE maps for Figure 3, row 3, using the frozen single-dataset RefCOCO checkpoint. Top: image/ground-truth outline and $c_k(x)$ (scale $[0, 1]$, maxima printed), upsampled from 32×32. Bottom: token attention $a_{k\ell}$; q_1, q_2 receive containment supervision.

The RefCOCO+ PARE checkpoint supplies target-dataset logits, but no downstream parameter is fit or recalibrated on RefCOCO+. The input-mask threshold, projection tuple, arbiter weights, and edit thresholds remain fixed at their RefCOCO values. The target-specific policy is an independently fitted reference, not a selection target. Target labels evaluate each frozen policy once and never select a row or setting. On val/testA/testB, the transferred final policy differs from the independently target-fitted reference by only $-0.01/-0.03/+0.01$ pp mIoU. The 1,546 RefCOCO BTRA-fitting images and 168 disjoint calibration images have zero overlap with the 1,500 RefCOCO+ validation images; the same exact audit finds zero overlap with all 750 testA and 750 testB images. Thus policy fitting is image-disjoint from target evaluation, and the result establishes policy portability under annotation shift.

Table 9: Frozen RefCOCO-to-RefCOCO+ downstream-policy transfer. Paired entries: mIoU/oIoU (%); Δ: full chain minus PARE@src (pp); CIs: paired image-cluster bootstrap 95% intervals. PARE@src uses the source input threshold.

Split	PARE@src	Transferred	Δm/o	mIoU 95% CI	vs. target-specific m/o
val	70.72/66.58	71.37/66.91	+0.65/+0.33	$[+0.61, +0.69]$	$-0.01/-0.04$
testA	75.06/71.41	75.69/71.71	+0.63/+0.30	$[+0.57, +0.68]$	$-0.03/-0.04$
testB	63.26/57.04	63.79/57.26	+0.53/+0.22	$[+0.48, +0.59]$	$+0.01/-0.01$

Table 10: Frozen BTRA vs. its ECBP input (mIoU). CIs are paired image-cluster bootstrap 95% intervals for a frozen model.

Dataset	Split	ECBP	+BTRA	Δ	95% CI
RefCOCO	val	78.23	**78.45**	+0.22	$[+0.20, +0.24]$
	testA	80.46	**80.70**	+0.24	$[+0.22, +0.27]$
	testB	75.47	**75.62**	+0.15	$[+0.13, +0.18]$
RefCOCO+	val	71.16	**71.38**	+0.23	$[+0.21, +0.25]$
	testA	75.46	**75.72**	+0.26	$[+0.23, +0.28]$
	testB	63.65	**63.78**	+0.13	$[+0.10, +0.16]$
G-Ref (UMD)	val(u)	70.05	**70.29**	+0.24	$[+0.21, +0.28]$
	test(u)	70.69	**70.96**	+0.28	$[+0.25, +0.30]$
G-Ref (Google)	val(g)	68.29	**68.59**	+0.30	$[+0.28, +0.33]$

E PROTOCOL DETAILS

Table 11 gives the single-dataset B-scale projection and arbitration settings; Table 12 gives mixed L-scale settings. For single-dataset ECBP, parameter selection uses training expressions $[8192, 9216)$, and confirmation sets are constructed from training expressions $[9216, 10240)$. For RefCOCO, this construction excludes three expressions associated with the one COCO image shared with the selection block. The resulting 1,021-expression confirmation set is image-disjoint from the selection set. The reported $+0.2975$ pp mIoU gain and image-cluster 95% CI $[+0.2377, +0.3593]$ are computed on this set.

Confidence intervals use paired bootstrap resampling; frozen BTRA evaluation uses 10,000 resamples. Because several expressions describe the same image, ECBP and BTRA additionally require a positive image-cluster interval. These paired intervals quantify evaluation uncertainty for a frozen model.

On RefCOCO, BTRA development uses the first 8,192 training expressions, split by COCO image (seed 7407) into 5,739 fitting, 1,243 calibration, and 1,210 development-audit expressions. The image-disjoint audit set evaluates the development arbiter with calibration-selected thresholds; checks include gains of at least $+0.20$ pp mIoU and $+0.10$ pp oIoU over ECBP, and a positive lower bound of the image-cluster mIoU confidence interval. The final refit combines the original fitting and development-audit expressions with additional training expressions, excluding all expressions associated with calibration images. This yields 11,045 fitting expressions while preserving the original 1,243-expression calibration set. The arbiter is fit for 24 epochs. Official evaluation uses the frozen arbiter and decision thresholds. The architecture, loss, trust-region definition, fitting schedule, and selection objective were frozen after RefCOCO; other datasets fit only their own arbiter weights and thresholds.

Table 13 measures 200 expressions over three repeats; latency is the mean of repeat means. The complete B-scale chain trains 23.66M parameters and deploys 267.40M, including the frozen backbones. RISAM uses frozen SAM ViT-H and Prompt-RIS uses CLIP+SAM at inference; their sources omit latency and memory. The deployed lower bounds follow the backbone sizes and reported trainable fractions in Table 21.

Table 11: Frozen inference hyperparameters for single-dataset B-scale evaluation. The posterior threshold t is selected by validation mIoU over a fixed grid; projection parameters and BTRA thresholds are train-selected.

Parameter	RefCOCO	RefCOCO+	G-Ref UMD	G-Ref Google
posterior t	0.475	0.475	0.45	0.475
p_{add}	0.10	0.10	0.10	0.15
p_{remove}	0.625	0.60	0.55	0.575
r	0.06	0.06	0.06	0.06
$\theta_{\text{add}}/\theta_{\text{rem}}$	0.68/0.66	0.62/0.68	0.62/0.64	0.62/0.64

Table 12: Frozen inference hyperparameters for mixed L-scale evaluation. One posterior threshold, one projection rule, and one arbiter are shared across all eight official splits.

Parameter	Mixed L
posterior t	0.525
p_{add}	0.20
p_{remove}	0.60
r	0.06
$\theta_{\text{add}}/\theta_{\text{rem}}$	0.76/0.74

Table 13: Single-dataset B-scale inference cost: RTX 3080, batch 1, FP32, no SAM embedding cache. Train./Deploy.: trainable/deployed parameters; Thr.: throughput; Mem.: peak memory; n/r: unreported; $>$: lower bound. †: Prompt-RIS mixed protocol.

Method	Train. (M)	Deploy. (M)	Latency (ms)	Thr. (expr/s)	Mem. (GiB)
TALENT	22.78	172.79	49.56	20.18	0.80
PARE	23.15	173.16	51.87	19.28	0.83
PARE+ECBP	23.15	266.90	101.70	9.83	3.36
PARE+ECBP+BTRA	23.66	267.40	154.60	6.47	3.36
RISAM (Zhang et al., 2023)	6	$>$600	n/r	n/r	n/r
Prompt-RIS† (Shang et al., 2024)	n/r	$>$240	n/r	n/r	n/r

F FULL BTRA TABLE

Table 14 pairs the frozen arbiter against its frozen ECBP input on every official split in both overlap metrics; it supports the nine-split claim of Section 4.3.

Table 14: Frozen BTRA paired against its frozen ECBP input on every completed official split, including oIoU.

Dataset	Split	ECBP mIoU	+BTRA	ΔmIoU	ECBP oIoU	+BTRA	ΔoIoU
RefCOCO	val	78.23	78.45	+0.22	75.51	75.67	+0.16
	testA	80.46	80.70	+0.24	78.80	78.99	+0.19
	testB	75.47	75.62	+0.15	72.11	72.22	+0.11
RefCOCO+	val	71.16	71.38	+0.23	66.77	66.95	+0.18
	testA	75.46	75.72	+0.26	71.55	71.75	+0.20
	testB	63.65	63.78	+0.13	57.14	57.27	+0.13
G-Ref (UMD)	val(u)	70.05	70.29	+0.24	65.51	65.65	+0.14
	test(u)	70.69	70.96	+0.28	67.55	67.72	+0.17
G-Ref (Google)	val(g)	68.29	68.59	+0.30	63.40	63.57	+0.17

F.1 ROBUSTNESS AND EDIT RELIABILITY

The rejected-pixel fraction is 92.7% on RefCOCO train confirmation and 92.5% on full validation.

We measure local sensitivity around the frozen parameters on all 10,834 RefCOCO validation expressions. Table 15 reports the largest absolute change over the indicated neighborhood. The 256×256 ECBP cache reconstruction differs from the original masks on 0.112% of pixels, within the predeclared 0.5% tolerance. Full-resolution BTRA evaluation reproduces 78.45/75.67 mIoU/oIoU. BTRA edit thresholds exhibit a local plateau within the $\pm 10\%$ neighborhood, with maximum absolute changes of 0.03 pp in both mIoU and oIoU. Expanding the neighborhood to $\pm 20\%$ gives maximum absolute changes of 0.13 and 0.09 pp, respectively. ECBP thresholds/band width and BTRA trust-region scale remain within 0.05 pp mIoU under $\pm 20\%$ perturbations.

Table 15: Post-hoc local sensitivity around frozen RefCOCO-val parameters. Entries are the maximum absolute metric change (percentage points) over one-factor and coupled perturbations; no setting is reselected.

Component	Neighborhood	max $\lvert\Delta\rvert$ mIoU	max $\lvert\Delta\rvert$ oIoU
ECBP thresholds/band	$\pm 10\%$	0.02	0.01
	$\pm 20\%$	0.05	0.03
BTRA edit thresholds	$\pm 10\%$	0.03	0.03
	$\pm 20\%$	0.13	0.09
BTRA trust-region scale	$\pm 10\%$	0.00	0.00
	$\pm 20\%$	0.01	0.00

Table 5 in Section 4.3 reports the matched-coverage comparison. SAM-direction ranking uses posterior scores to break proposal-membership ties; every deterministic ranking then breaks remaining ties by flattened pixel index. The random control averages seeds 2027–2031. Accepted-edit correctness is pixel-weighted across expressions and directions. The generic-refiner control below separately compares action semantics at matched learned capacity.

F.2 PARAMETER-MATCHED LEARNED-REFINER CONTROL

To isolate the effect of edit selection at matched capacity, we train a generic local-mask refiner with the exact ten-channel spatial input, frozen SAM embedding, boundary-band support, architecture tensor shapes, and optimization budget used by BTRA. Both methods contain 168,978 parameters per member, use seeds 6407/6408/6409, and fit on 11,045 expressions with 1,243 image-disjoint train-only calibration expressions. The sole intended difference is action semantics: the generic

two-logit head directly reassigns every band pixel as background or foreground, without direction-specific addition/removal heads or a default-reject path. The generic refiner can reassign either label throughout the matched local band.

Before reading validation we freeze the generic threshold at $t = 0.50$, which maximizes its train-calibration mIoU; validation labels select nothing. The generic-refiner row of Table 4 reports the resulting one-shot evaluation on all 10,834 RefCOCO-val expressions.

The generic refiner leaves ECBP mIoU unchanged at 78.23 and changes oIoU by -0.02 pp; its image-cluster mIoU 95% CI is $[-0.0304, +0.0359]$ pp. Under matched capacity and training budget, BTRA exceeds the generic refiner by $+0.22/+0.18$ pp mIoU/oIoU and by $+1.31$ pp at Pr@0.9. This RefCOCO-val control supports the combined direction-specific, selective arbitration design under a matched network and ensemble budget.

F.3 Default-reject and prompt controls

Two further controls hold the PARE logits fixed and change one element of the downstream chain on all 10,834 RefCOCO-val expressions.

Every candidate admitted. Table 16 keeps the frozen ensemble weights and sets both edit thresholds to zero, so every ring candidate is admitted and the default is removed. mIoU falls 5.76 pp below the ECBP input and Pr@0.9 from 42.63 to 3.64, with 9,937 of the 10,834 expressions losing IoU, 501 improved, and 396 unchanged. The ring, the scores, and the members are identical in the two rows, so the difference is the admission rule alone: the arbiter's contribution is what it declines.

Table 16: Default-reject control on RefCOCO val (%). Same frozen BTRA and ECBP input. Accepted: pixel-weighted ring coverage; worsened: expressions losing IoU relative to ECBP. $\theta = 0$ admits every candidate.

Row	Accepted (%)	mIoU	oIoU	Pr@0.5	Pr@0.7	Pr@0.9	Worsened
ECBP input	–	78.23	75.51	88.31	82.55	42.63	–
ECBP + BTRA, frozen θ (0.68/0.66)	7.5	**78.45**	**75.67**	**88.38**	**82.69**	**44.25**	3,569
ECBP + BTRA, $\theta = 0$	100.0	72.47	69.99	87.80	79.87	3.64	9,937

Point and box prompt only. Table 17 drops the dense mask prompt and calls frozen SAM ViT-B with the posterior peak point and the M_0 box alone. PARE and the frozen ECBP rule are unchanged; BTRA is refit on this proposal with the official recipe (seeds 6407/6408/6409, 24 AdamW epochs, 11,045 fitting and 1,243 image-disjoint calibration expressions, thresholds 0.68/0.66 selected on calibration only). ECBP adds $+0.37$ mIoU over PARE and BTRA a further $+0.24/+0.18$ mIoU/oIoU (image-cluster 95% CI $[+0.217, +0.260]$ pp) while admitting 9.3% of ring pixels; the complete chain reaches 78.41, within 0.04 of the dense-prompt chain, and leads it at Pr@0.7. The authority rule does not depend on the dense prompt.

Table 17: Prompt control on RefCOCO val with frozen PARE logits and frozen SAM ViT-B. Dense rows are the official chain (point, box, and dense mask prompt); point+box rows drop the dense mask. Δ is mIoU over PARE.

Prompt	Stage	mIoU	oIoU	Pr@0.5	Pr@0.7	Pr@0.9	Δ
–	PARE posterior (M_0)	77.80	75.31	88.19	82.23	39.42	–
point+box+dense	ECBP	78.23	75.51	88.31	82.55	42.63	+0.43
	ECBP + BTRA	**78.45**	**75.67**	**88.38**	82.69	**44.25**	**+0.65**
point+box	ECBP	78.17	75.46	88.25	82.59	42.27	+0.37
	ECBP + BTRA	78.41	75.64	88.27	**82.72**	44.04	+0.61

F.4 Expressions with a same-category competitor: catastrophic-drop control

Table 18 evaluates the frozen same-SAM variants on every RefCOCO-val expression for which another annotated instance of the target's COCO category occurs in the same image. A same-

category competitor is nearly universal in RefCOCO: the set contains 10,522 of the 10,834 validation expressions (97%) from 1,405 images, so its mIoU tracks the full-set value (the 312 expressions without a competitor average a lower IoU, which is why the subset mean of PARE, 77.89, sits above its full-set 77.80). The control is therefore a per-expression stability audit rather than a hard subset. The table reports subset mIoU and the catastrophic-drop rate defined in Appendix B, in percent of subset expressions. The constrained ECBP and ECBP+BTRA rows incur no ≥ 0.20-IoU drop from PARE; the veto-free local variant incurs two (0.02%) and whole-SAM replacement 25 (0.24%). The complete chain also has the highest subset mIoU. Table 4 reports the full-validation counterpart.

Table 18: Post-hoc control on the 10,522 RefCOCO-val expressions with a same-image, same-category competitor (97% of the split). Cat. drop is the percentage of these expressions losing at least 0.20 IoU from PARE (0.24% is 25 expressions). No label selects a method or threshold.

Method	mIoU	Cat. drop (%)
PARE	77.89	0.00
Naive whole-SAM	78.35	0.24
Local, no veto	78.44	0.02
ECBP	78.33	**0.00**
ECBP+BTRA	**78.55**	**0.00**

The two edit directions contribute complementary gains. Holding the released scores and thresholds fixed, addition-only changes ECBP by $+0.07/+0.04$ pp mIoU/oIoU, removal-only by $+0.15/+0.12$ pp, and their composition by $+0.22/+0.16$ pp (Table 19). Addition recovers 0.280 pp completeness at the cost of 0.215 pp commission, whereas removal reduces commission by 0.493 pp at a 0.354 pp completeness cost; the two directions therefore perform complementary rather than interchangeable operations.

Table 19: Frozen directional BTRA ablation on all 10,834 RefCOCO-val expressions. No score, threshold, or model is refit. Coverage is the accepted fraction of all outer- and inner-ring candidates.

Accepted directions	mIoU	oIoU	Δ mIoU	Coverage (%)
None (ECBP)	78.23	75.51	0.00	0.00
Addition only	78.30	75.55	+0.07	2.78
Removal only	78.38	75.63	+0.15	4.74
Addition + removal	**78.45**	**75.67**	**+0.22**	7.52

Table 20 compares every stratum to its own frozen ECBP input. Area cut points were fixed from the ground-truth area distribution before inspecting prediction effects. Shape bins use the isoperimetric quotient $P^2/(4\pi A)$ of the ground-truth mask, where P is the total length of its external contours and A is its area: compact is at most 2, irregular is above 2 and at most 5, and thin/complex is above 5. The topology proxy counts 8-connected foreground components of that same mask; more than one component is the multi-component bin. BTRA improves both overlap metrics in all area, input-error, shape, and topology bins. Its update consistently trades a small completeness decrease for a larger commission reduction. Gains are smallest when ECBP completeness is already below 0.90 or commission is already below 0.04; the policy is most useful when a reasonably complete mask still contains removable spill. Thin/complex masks remain positive, while the multi-component visibility-fragment proxy has the smallest oIoU gain. BTRA improves pooled overlap relative to its upstream input across every one of these strata.

The sigmoid scores rank edit utility. Using ten equal-width score bins over all ring candidates, addition/removal ECE is 0.081/0.096. Under the counterfactual "apply this edit" label, accepted additions are correct 47.74% of the time against 24.67% among rejected candidates, and accepted removals 64.50% against 35.89%: in both directions the arbiter admits the edits that are more often right and declines the rest, with removal the stronger absolute signal. Here *trust* denotes learned edit admission inside a bounded action space.

Table 20: Frozen BTRA minus frozen ECBP by RefCOCO-val diagnostic stratum. Area bins use target/image area; completeness and commission bins use the ECBP input. Multi-comp is an annotation-free visibility-fragment proxy, not a ground-truth occlusion label.

Axis	Stratum	N	ΔmIoU	ΔoIoU	ΔComp.	ΔComm.
Target area	$\leq 6\%$	3,544	+0.28	+0.20	-0.06	-0.31
	$(6\%, 11\%]$	3,610	+0.23	+0.19	-0.10	-0.33
	$> 11\%$	3,680	+0.15	+0.12	-0.07	-0.23
Input completeness	< 0.90	2,826	+0.10	+0.02	-0.11	-0.19
	$[0.90, 0.97)$	4,319	+0.28	+0.23	-0.05	-0.30
	≥ 0.97	3,689	+0.24	+0.20	-0.08	-0.28
Input commission	≤ 0.04	3,495	+0.07	+0.05	-0.12	-0.20
	$(0.04, 0.10]$	4,200	+0.35	+0.31	-0.04	-0.38
	> 0.10	3,139	+0.21	+0.11	-0.05	-0.21
Shape complexity	compact	3,532	+0.16	+0.11	-0.10	-0.24
	irregular	5,897	+0.24	+0.19	-0.07	-0.30
	thin/complex	1,405	+0.27	+0.19	-0.02	-0.27
Topology	single component	9,645	+0.22	+0.16	-0.06	-0.27
	multi-comp proxy	1,189	+0.21	+0.10	-0.17	-0.30

G SCENE-STRATIFIED QUALITATIVE COMPARISONS

Figures 6 and 7 expand the main-paper comparison to 18 frozen RefCOCO-val cases across color, clothing/attribute composition, ordinal language, orientation, negation, relative position, complex anchor relations, and implicit or uncommon descriptions. The set spans all five stronger-baseline IoU bands. In the low-band cases, the baselines' masks sit on a same-class distractor and ours lies on the annotated object; in the moderate-band cases, the baselines select a plausible target but miss substantial extent; and in the high-band cases, all three masks share the target and differ in residual extent and boundary detail. All three predictions use the same purple overlay. The stronger-baseline IoU bands $[0, .2)$, $[.2, .4)$, $[.4, .6)$, $[.6, .8)$, and $[.8, 1]$ contain 3/5/3/3/4 cases, respectively.

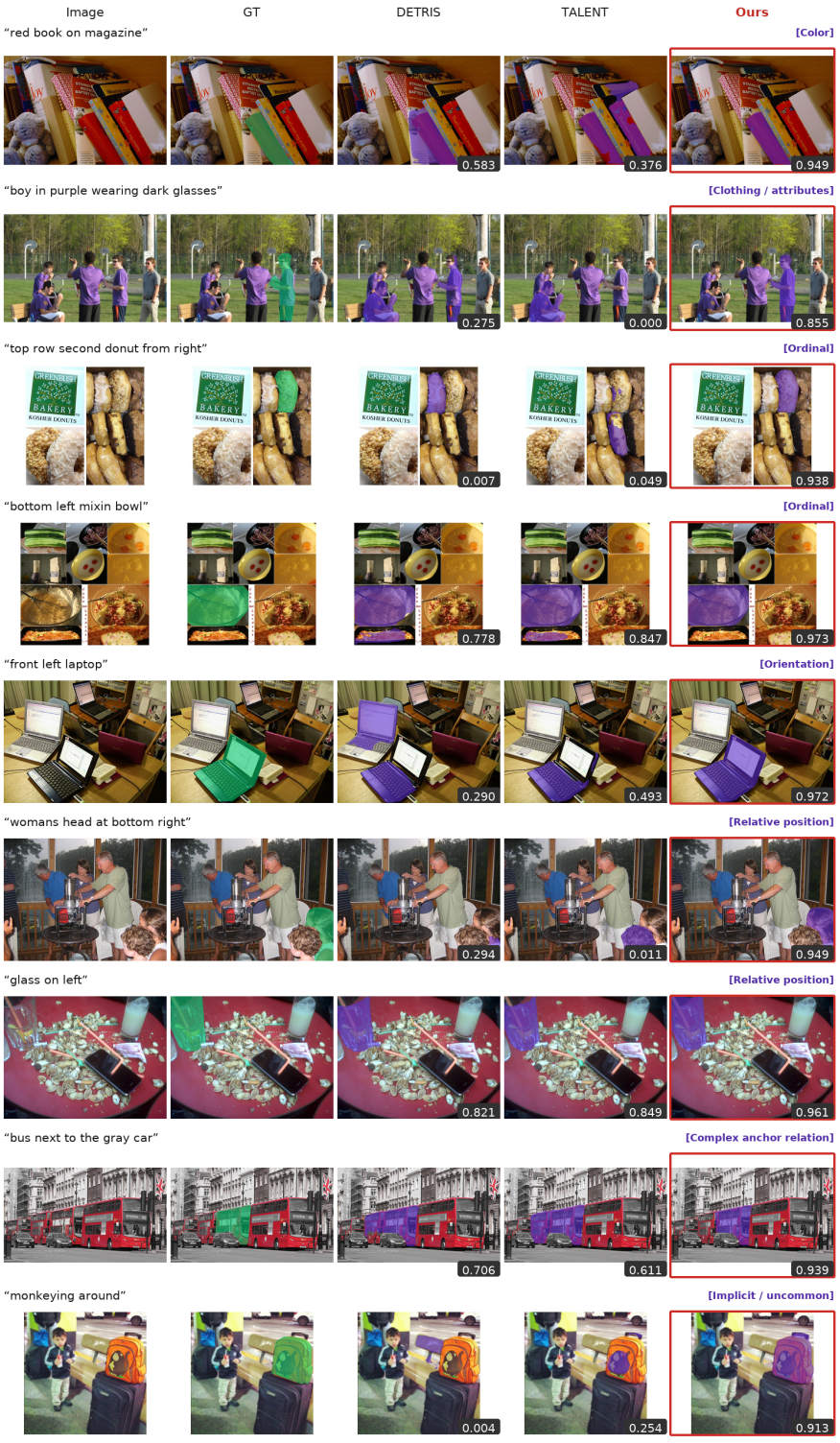

Figure 6: RefCOCO val qualitative cases 1–9 (continued in Figure 7). Green: ground truth; purple: prediction; red frames: ours. Bracketed tags: scene categories; numbers: per-expression IoU. Some images are cropped to highlight segmentation details.

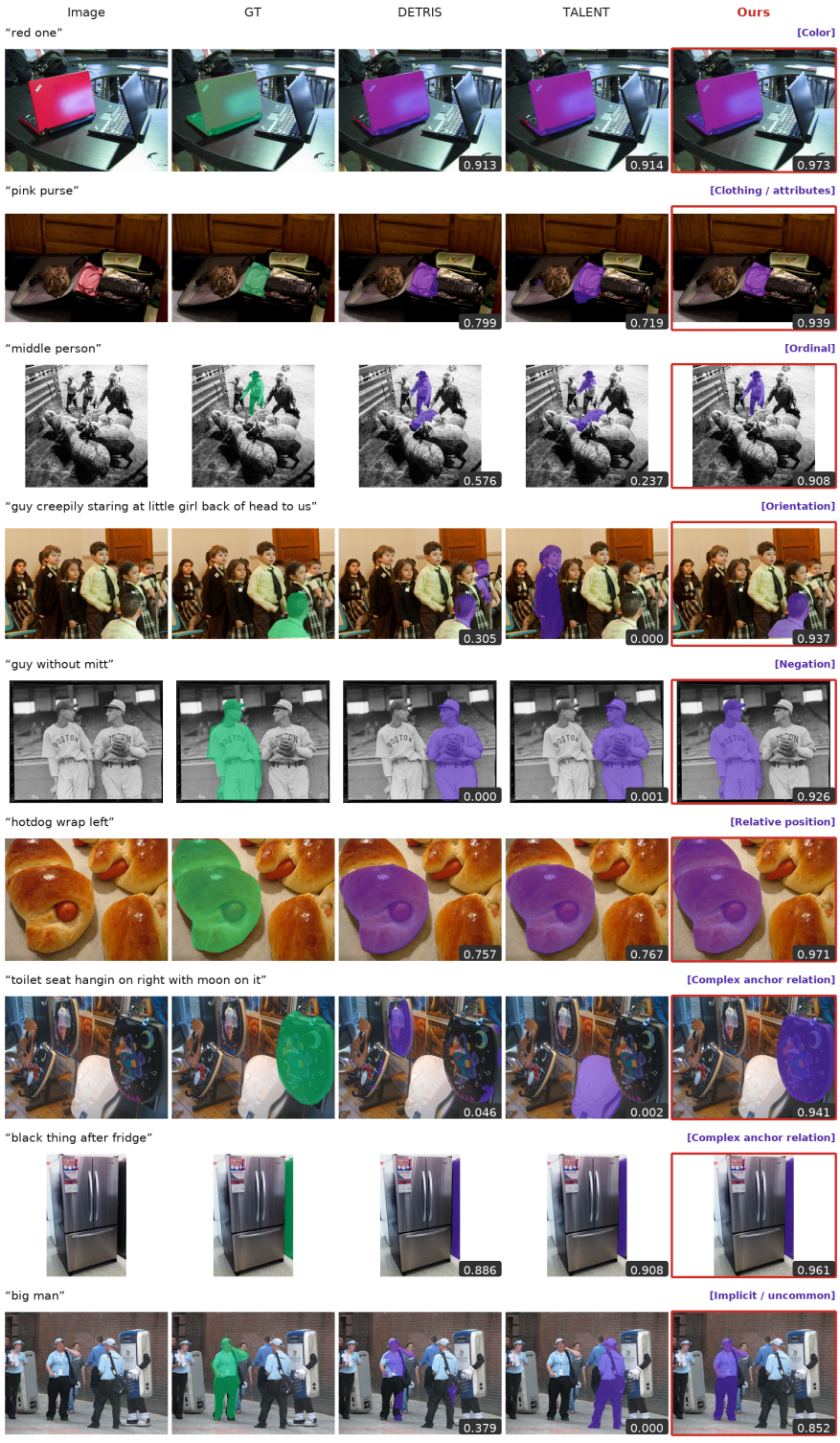

Figure 7: Scene- and difficulty-stratified qualitative comparison on RefCOCO val, cases 10–18 (continued from Figure 6); same conventions. Some images are cropped to highlight segmentation details.

H SAM-BASED COMPARISON SCOPE

Table 21 records the setting of each SAM-driven RIS system cited in the paper. Joint-training systems update a large encoder or a 7B-class LLM together with SAM and are not table rows in Section 4: they do not share the PET trainable-parameter budget. RISAM and Prompt-RIS freeze SAM and publish official RefCOCO numbers; they are the comparison set in Tables 1–2.

RISAM runs on a frozen SAM ViT-H visual backbone (Zhang et al., 2023) and reports 6M trainable weights as about 1% of the full model, so deployment exceeds 0.6B. E3-RIS-B uses frozen DINOv2-B/14, CLIP ViT-B/16, and SAM ViT-B, with 267.40M deployed parameters. Prompt-RIS deploys CLIP ViT-B/16 and SAM ViT-B, already above 240M. Table 21 lists published totals where the source reports them, and otherwise a lower bound from the published backbone or stated trainable fraction.

Table 21: Settings of SAM-driven RIS systems cited in this paper. Train. and Deploy. are published totals, or lower bounds implied by the source when a total is omitted; n/r means not reported. Joint-train: a large encoder or LLM coupled to SAM is updated (SAM's image encoder may still be frozen). Inf. SAM: a SAM decoder is called at inference.

Method	Vision / text	Train.	Deploy.	Joint-train	Inf. SAM	Protocol
LISA (Lai et al., 2024)	SAM / Vicuna-7B	n/r	>7B	yes	yes	instruction
EVF-SAM (Zhang et al., 2024)	SAM ViT-H / BEIT-3-L	673M[§]	1.32B	yes	yes	RES
ReferSAM (Liu et al., 2025)	SAM / BERT	n/r	n/r	yes	yes	pooled referring
E-ReferSAM (Li et al., 2026)	SAM / BERT	n/r	n/r	yes	yes	pooled referring
RISAM (Zhang et al., 2023)	SAM ViT-H / CLIP	6M	>0.6B[¶]	no	yes	single-dataset
Prompt-RIS (Shang et al., 2024)	CLIP+SAM / CLIP	n/r	>240M[¶]	no	yes	three-set union
E3-RIS-B	DINOv2-B+SAM-B / CLIP-B	23.66M	267.40M	no	yes	single or mixed

[§]EVF-SAM freezes the SAM image encoder and trains BEIT-3; the published default also trains SAM's prompt encoder and mask decoder (Zhang et al., 2024). [¶]RISAM reports 6M trainable parameters occupying around 1% of the whole model (Zhang et al., 2023), so the deployed size exceeds 0.6B. Prompt-RIS deploys CLIP ViT-B/16 and SAM ViT-B (Shang et al., 2024); those two backbones already exceed 240M.

I EXTENDED RELATED WORK

Referring image segmentation. Early RIS methods encode an expression with a recurrent language model and fuse it with convolutional features (Hu et al., 2016; Liu et al., 2017; Li et al., 2018; Hu et al., 2020; Huang et al., 2020; Liu et al., 2021), often through an FCN-style decoder (Long et al., 2015). Multi-task variants jointly train comprehension and segmentation (Luo et al., 2020; Yu et al., 2018). Transformers then strengthened cross-modal reasoning: ReSTR removes convolutions (Kim et al., 2022), LAVT injects language into vision-transformer features (Yang et al., 2022), CRIS transfers CLIP alignment to text-to-pixel contrast (Wang et al., 2022; Chen et al., 2022a), VLT uses query generation (Ding et al., 2021), and CGFormer groups object-aware tokens (Tang et al., 2023). Later work adds mask grounding (Chng et al., 2024), context-aware decoding (Liu et al., 2023c; Shah et al., 2024), polygon generation (Liu et al., 2023b), generalized referring expressions (Liu et al., 2023a; Hu et al., 2023; Luo et al., 2025), Mamba-based fusion (Yang et al., 2024), and LLM reasoning (Lai et al., 2024). Related grounded pretraining and detection formulations include MDETR, GLIP, and UNINEXT (Kamath et al., 2021; Li et al., 2022b; Yan et al., 2023; Li et al., 2022a). DeRIS separates perception and cognition branches and reconnects them through loopback synergy (Dai et al., 2025); VATEX conditions text understanding on a CLIP visual prior and separates object from context understanding (Nguyen-Truong et al., 2025); modular and latent-expression methods

decompose language into subject, attribute, or relation representations (Yu et al., 2018; 2025b). These methods stage computation or decompose representation. E3-RIS addresses an orthogonal question: *which source may change which decision*. Its evidence queries pool into one referential code, the external shape prior cannot replace the whole prediction, and the final corrector is spatially confined to direction-specific rings.

Parameter-efficient tuning for RIS. PET adapts large models by updating adapters, prompts, or low-rank residuals (Houlsby et al., 2019; Li & Liang, 2021; Ben Zaken et al., 2022; Hu et al., 2022; Karimi Mahabadi et al., 2021; Chen et al., 2022b; Zhou et al., 2022; Gao et al., 2023; Chen et al., 2023; Jie et al., 2024; Liu et al., 2024a). ETRIS introduces lightweight bridges between frozen encoders (Xu et al., 2023), BarLeRIa couples an intertwined adapter with global tuning (Wang et al., 2024), RISCLIP exploits CLIP geometry (Kim et al., 2024), ETOG studies language-guided grounding under PET (Yu et al., 2025a), and DETRIS densely connects multi-scale adapters on DINOv2 and CLIP (Huang et al., 2025; Oquab et al., 2024). TALENT identifies non-target activation and adds target-aware learning (Jin et al., 2026). We build on a paper-equation-faithful TALENT reconstruction and address the next bottleneck: stronger target discrimination still leaves extent incomplete, and even a useful shape proposal requires a separate trust decision before it may alter the mask.

Promptable segmentation priors. SAM provides a strong class-agnostic shape prior (Kirillov et al., 2023), and two families adapt it to RIS. The large family couples a vision–language encoder to SAM and trains the pair: EVF-SAM encodes image and text jointly with BEIT-3 and feeds the fused embedding to SAM as its prompt (Zhang et al., 2024), ReferSAM fine-tunes SAM with a BERT encoder on pooled referring datasets and E-ReferSAM adds fine-grained alignment and boundary-aware fusion (Liu et al., 2025; Li et al., 2026), and LISA lets a multimodal LLM emit the token that SAM decodes (Lai et al., 2024). The parameter-efficient family freezes SAM and learns its inputs: Prompt-RIS jointly bridges prompt-tuned CLIP and SAM and lets CLIP-derived mask, point, and text prompts drive the SAM decoder (Shang et al., 2024), RISAM adapts frozen SAM with mutual-aware attention features under parameter-efficient tuning (Zhang et al., 2023), and DFAM designs the bridging features for that decoder (Ito, 2025); generalist systems expose broader prompt interfaces (Zou et al., 2023; Liu et al., 2024b). E3-RIS keeps the RIS posterior as the starting identity decision: frozen SAM cannot replace the mask globally, ECBP exposes only posterior-compatible local directions, and BTRA can act only in predeclared outer-addition and inner-removal rings. This restricted interface is also distinct from vision-aware text methods such as RISCLIP and VATEX, which improve cross-modal representation before mask prediction (Kim et al., 2024; Nguyen-Truong et al., 2025); our central intervention governs downstream authority after a posterior already exists.

Boundary refinement, selective prediction, and authority control. PointRend adaptively predicts uncertain boundary points, SegFix transports reliable interior labels to boundary pixels, and Boundary IoU makes contour error measurable (Kirillov et al., 2020; Yuan et al., 2020; Cheng et al., 2021). Mask Scoring R-CNN learns mask-quality scores (Huang et al., 2019); calibration and selective-prediction work instead asks whether confidence matches correctness or whether a model should abstain (Guo et al., 2017; Geifman & El-Yaniv, 2019). BTRA is related but operates at a different granularity: it predicts two pixelwise edit utilities, not a mask-level quality score, and its abstention domain is a pair of direction-specific rings conditioned jointly on the referring posterior, SAM proposal, RGB boundaries, and frozen SAM features. Recurrent and query-mask RIS systems refine inside the predictor (Li et al., 2018; Shah et al., 2024); E3-RIS instead makes the external interface explicit. The interface bounds the action space: edits are direction-specific and confined to the boundary rings. Appendix F.4 measures the per-expression stability of that space on the same-class distractor subset.