NOT ALL EVIDENCE IS EQUAL: ASYMMETRIC REFERENTIAL EVIDENCE MODELING FOR REFERRING REMOTE SENSING IMAGE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Referring Remote Sensing Image Segmentation (RRSIS) aims to identify and segment the target specified by a natural-language expression in a remote-sensing image. A central challenge is to distinguish the intended target from multiple visually and semantically plausible candidates. Existing methods primarily enrich discriminative information, but often overlook how evidence with different functions should influence referent identification. We argue that different forms of referential evidence should be assigned authority according to their functional roles. Based on this principle, we propose Asymmetric Referential Evidence Modeling (AREM), which forms role-specific referential evidence, constructs positive-anchored dense grounding with bounded role-aware correction, and calibrates prompt decoding through discriminative training and controlled refinement. Extensive experiments on RefSegRS and RRSIS-D demonstrate the effectiveness and efficiency of AREM, while cross-architecture and cross-domain evaluations further support its applicability beyond the primary setting.

1 INTRODUCTION

Referring Remote Sensing Image Segmentation (RRSIS) aims to predict the target referred to by a natural-language expression in a remote-sensing image (Yuan et al., 2024; Liu et al., 2024). Compared with natural-image referring segmentation, remote-sensing scenes contain repetitive structures, densely distributed same-category instances and substantial variations in scale and orientation. Consequently, a referring expression may correspond to several visually and semantically plausible candidates. One of the challenges is therefore not merely to locate a region compatible with the expression, but to make the intended referent *identifiable* among competing candidates.

Recent foundation-model-based approaches increasingly adopt decoupled and promptable segmentation paradigms. Vision–language encoders interpret the image–expression pair, while segmentation foundation models such as SAM and SAM2 convert sparse and dense prompts into fine-grained masks (Kirillov et al., 2023; Ravi et al., 2025; Chen et al., 2026; Rong et al., 2026). Although these architectures substantially improve mask generation, stronger segmentation capability alone cannot resolve referential ambiguity when the generated prompts activate multiple plausible candidates. Therefore, referential ambiguity is better addressed before and during prompt construction, rather than being left entirely to the segmentation decoder.

Existing studies mainly address this problem by introducing richer discriminative information through stronger cross-modal interaction, multi-level representations, uncertainty modeling, candidate selection, or negative evidence (Chen et al., 2026; Sun et al., 2026; Gao et al., 2026; Luo et al., 2026). These approaches demonstrate the value of additional contextual and exclusionary cues, but generally focus on what evidence should be incorporated rather than how evidence with different functions should influence referent identification.

In our view, referential evidence is functionally heterogeneous. Some evidence directly characterizes the target and provides the primary basis for identifying it, while other evidence characterizes regions inconsistent with the referent and thereby helps reject plausible alternatives. Although these cues may all contribute to referent identification, they need not play the same role. As illustrated in Fig. 1, treating heterogeneous evidence with comparable authority can lead to ambiguous or over-corrected

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.

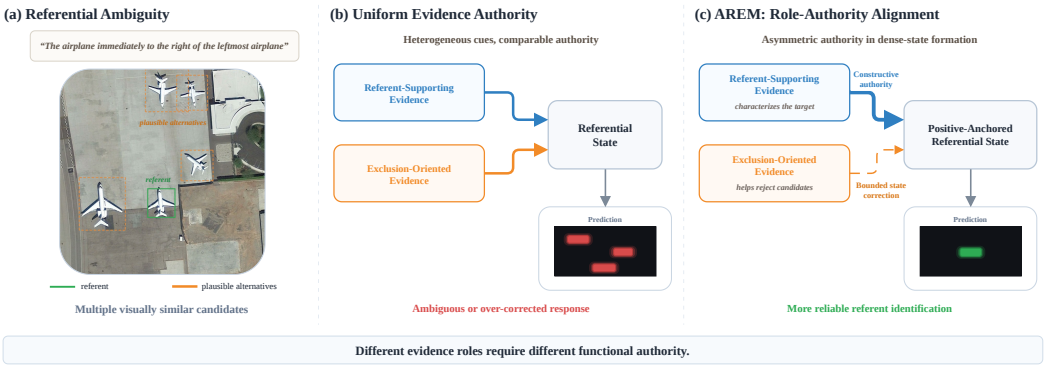


Figure 1: **Motivation of asymmetric evidence authority.** (a) Referential ambiguity with multiple plausible candidates. (b) Uniform evidence authority may yield ambiguous or over-corrected responses. (c) AREM aligns evidence roles with functional authority for more reliable referent identification.

responses, whereas role-aligned authority better preserves the referent while suppressing competing candidates. Our central principle is that *different forms of referential evidence should not be granted identical functional authority*.

Based on these observations, we propose **Asymmetric Referential Evidence Modeling (AREM)**, comprising three functional components. First, **Role-Asymmetric Evidence Formation (RAEF)** forms referent-supporting and exclusion-oriented evidence through role-conditioned queries. Second, **Positive-Anchored Evidence Grounding (PAEG)** constructs the dense referential state through positive-anchored semantic-depth reconciliation and bounded role-aware correction. Third, **Evidence-Calibrated Prompt Decoding (ECPD)** improves the expression discriminability of the initial dense anchor during training and introduces bounded candidate-based refinement during decoding. Overall, AREM coordinates different forms of evidence according to their functional roles and authority, enabling more reliable referent identification in ambiguous scenes.

Our main contributions are summarized as follows:

- We frame referential ambiguity through evidence role and authority, arguing that heterogeneous referential evidence should influence referent identification based on its functional role.
- We propose AREM, which integrates three complementary components to form role-aware evidence, regulate its functional authority, and enhance its use in referent identification.
- We extensively validate AREM across datasets and architectures, demonstrating its effectiveness, efficiency, and transferability.

2 Related Work

2.1 Referring Remote Sensing Image Segmentation

Referring Remote Sensing Image Segmentation (RRSIS) was introduced with the RefSegRS benchmark to study language-guided segmentation in remote-sensing imagery (Yuan et al., 2024). Its difficulty arises partly from dense objects, repetitive structures, and visually similar same-category instances. Existing methods mainly strengthen vision–language interaction and representation adaptation: RMSIN introduces rotated multi-scale interaction (Liu et al., 2024), CroBIM develops cascaded bidirectional cross-modal interaction (Dong et al., 2024), and SBANet combines bidirectional alignment with dynamic feature selection (Li et al., 2025). While these methods improve multimodal grounding, they focus primarily on extracting and integrating richer evidence, with less attention to how evidence serving different functions should influence referent identification. This motivates modeling referential evidence according to both its functional role and authority.

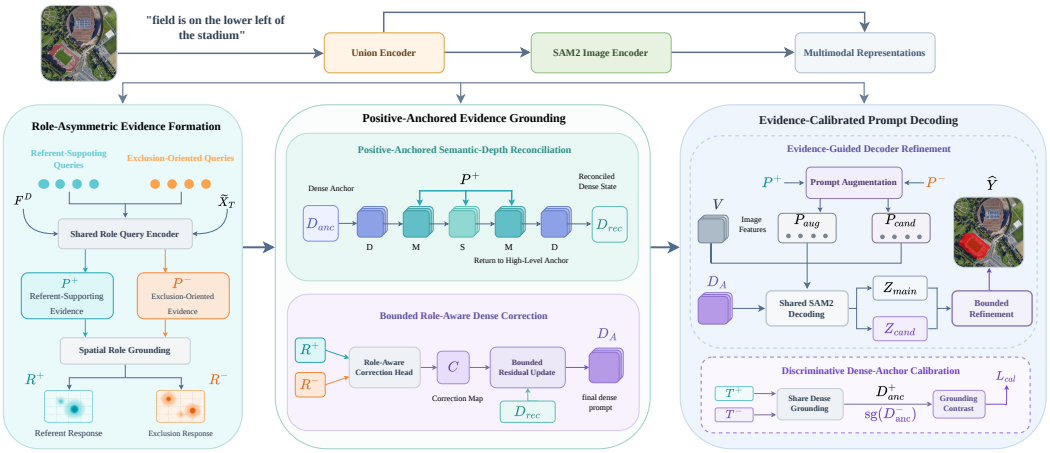


Figure 2: **Overview of the proposed AREM framework.** RAEF forms role-specific referential evidence, PAEG constructs the positive-anchored dense grounding, and ECPD calibrates prompt decoding to obtain the final segmentation mask.

2.2 SAM-BASED RRSIS

The promptable segmentation paradigm introduced by SAM and extended by SAM2 has provided a strong foundation for referring segmentation (Kirillov et al., 2023; Ravi et al., 2025). Recent RRSIS methods accordingly transform vision–language representations into prompts for foundation segmentation models. EVF-SAM explores text-conditioned vision–language fusion (Zhang et al., 2026); RSRefSeg and RSRefSeg 2 combine foundation representations with SAM-based mask generation, with the latter introducing decoupled semantic prompting to alleviate multi-object activation (Chen et al., 2025; 2026); and RS2-SAM2 integrates a unified vision–language encoder with SAM2 (Rong et al., 2026). Subsequent studies further incorporate uncertainty, candidate refinement, and contrastive cues to improve referent discrimination (Sun et al., 2026; Gao et al., 2026; Luo et al., 2026). Nevertheless, these methods primarily improve the acquisition and integration of prompting information, while paying less attention to whether evidence with different functional roles should be granted identical authority in referent formation.

3 METHOD

3.1 OVERVIEW

Given a remote-sensing image I and its referring expression T, we follow RS2-SAM2 (Rong et al., 2026) and employ its union BEiT-3 encoder (Wang et al., 2023) to obtain the textual representation X_T and language-conditioned visual representations F^{S}, F^{M}, and F^{D} from selected shallow, intermediate, and deep layers, respectively. In parallel, the SAM2 image encoder extracts image embeddings for mask decoding. As shown in Fig. 2, AREM organizes the overall pipeline around three functional components. **Role-Asymmetric Evidence Formation (RAEF)** forms role-specific referential evidence; **Positive-Anchored Evidence Grounding (PAEG)** constructs a positive-anchored dense prompt with bounded role-aware correction; and **Evidence-Calibrated Prompt Decoding (ECPD)** improves prompt reliability during training and decoding. Finally, the resulting sparse and dense prompts condition the SAM2 decoder, followed by bounded refinement to produce the final mask $\widehat{Y}$.

3.2 ROLE-ASYMMETRIC EVIDENCE FORMATION

RAEF organizes the shared multimodal representation into referent-supporting evidence P^+ and exclusion-oriented evidence P^-, and derives their spatial role responses R^+ and R^- for subsequent grounding and decoding.

3.2.1 ROLE-CONDITIONED QUERY FORMATION

To separate referential cues by their functional roles, we introduce role-conditioned queries. We initialize eight queries as $Q_{\text{role}} = Q_0 + E_{\text{role}}$, where Q_0 is a learnable query bank and E_{role} encodes their role assignments. The first four queries are referent-supporting, while the remaining four are exclusion-oriented.

Let $\widetilde{X}_T$ and $\widetilde{F}^{\text{D}}$ denote the textual and deep visual representations projected into the shared query space. The queries are progressively conditioned as

$$Q_{\text{role}} \xrightarrow{\text{SelfAttn}} U_{\text{self}} \xrightarrow{\text{CrossAttn}_T(\widetilde{X}_T)} U_T \xrightarrow{\text{CrossAttn}_V(\widetilde{F}^{\text{D}}+E_{\text{pos}},\widetilde{F}^{\text{D}})} U_V \xrightarrow{\text{FFN}} P_{\text{role}}. \tag{1}$$

Here, U_{self}, U_T, and U_V denote intermediate query states. Each block uses residual connections and layer normalization, while E_{pos} is a spatial positional embedding added to the visual keys.

The output preserves the predefined query ordering and is divided as

$$P_{\text{role}} = [P^+; P^-], \qquad P^+, P^- \in \mathbb{R}^{4\times d}, \tag{2}$$

where P^+ and P^- denote referent-supporting and exclusion-oriented sparse evidence, respectively.

3.2.2 SPATIAL ROLE RESPONSES

To associate the role-conditioned sparse evidence with spatial regions, let p_k^+ and p_k^- denote the k-th tokens of P^+ and P^-, respectively. Their spatial responses are

$$R_k^+ = \text{CosMap}\left(p_k^+, \widetilde{F}^{\text{D}}\right), \qquad R_k^- = -\,\text{CosMap}\left(p_k^-, \widetilde{F}^{\text{D}}\right), \tag{3}$$

where $\text{CosMap}(p, F)$ computes the cosine similarity between p and every token in F and restores the resulting scores to the $H_v \times W_v$ visual-token grid. For the exclusion role, we reverse the cosine polarity so that larger R_k^- consistently represents stronger non-referent response. Stacking the four maps in each role gives $R^+, R^- \in \mathbb{R}^{4\times H_v \times W_v}$.

During training, the ground-truth mask is resized to the spatial resolution of the role responses using nearest-neighbor interpolation and denoted by Y_R. We aggregate the strongest response within each role as

$$R^+_{\max} = \text{Max}_{\text{query}}(R^+), \qquad R^-_{\max} = \text{Max}_{\text{query}}(R^-). \tag{4}$$

The RAEF objective is

$$\mathcal{L}_{\text{RAEF}} = \lambda_{\text{reg}}\left[\ell_{\text{bce}}(R^+_{\max}, Y_R) + \eta_- \ell_{\text{bce}}(R^-_{\max}, 1 - Y_R)\right], \tag{5}$$

where ℓ_{bce} denotes spatially averaged binary cross-entropy with logits and η_- balances exclusion-oriented supervision. Thus, R^+ captures referred regions, whereas R^- provides a general non-referent spatial response rather than instance-specific distractor supervision.

3.3 POSITIVE-ANCHORED EVIDENCE GROUNDING

Building on the role-specific evidence formed by RAEF, PAEG governs how it influences dense grounding. Referent-supporting evidence P^+ constructs the dense referential state through semantic-depth reconciliation, while R^+ and R^- refine it through bounded residual correction.

3.3.1 POSITIVE-ANCHORED SEMANTIC-DEPTH RECONCILIATION

We index the selected semantic depths by $\ell \in \{\text{S}, \text{M}, \text{D}\}$, corresponding to the shallow, intermediate, and deep representations, respectively. The RS2-SAM2 grounding pathway provides an initial dense response D_{anc} at the SAM2 dense-prompt resolution, which serves as the deep-semantic referential anchor.

We mean-pool the referent-supporting evidence as $\bar{p}^+ = \text{Mean}_{\text{query}}(P^+)$ and use it to modulate F^ℓ for $\ell \in \{\text{S}, \text{M}\}$:

$$\widetilde{Z}_\ell = \uparrow G_\ell\left(\Pi_V^\ell(F^\ell) \odot \left[1 + \tanh\left(\Pi_P^\ell(\bar{p}^+)\right)\right]\right). \tag{6}$$

Here, Π_V^ℓ and Π_P^ℓ are level-specific projections, G_ℓ is a lightweight response head producing a single-channel spatial response, and $\uparrow$ denotes bilinear resizing to the dense-prompt resolution.

Instead of generating a separate response from F^{D}, we define $\widetilde{Z}_{\mathrm{D}} = \mathrm{logit}(D_{\mathrm{anc}})$ and use it as the initial state $Z^{(0)}$. We then follow the semantic-depth schedule $\boldsymbol{s} = (\mathrm{M}, \mathrm{S}, \mathrm{M}, \mathrm{D})$, where s_t denotes the level selected at step t:

$$Z^{(t)} = Z^{(t-1)} + \phi_{s_t}\left([\widetilde{Z}_{s_t}, Z^{(t-1)}]\right), \qquad t = 1, \ldots, 4. \tag{7}$$

Here, $[\cdot, \cdot]$ denotes channel concatenation, and ϕ_ℓ is a lightweight two-layer convolutional residual unit associated with semantic level ℓ, with shared parameters when the same level is revisited.

The reconciled state and its anchor-preservation constraint are

$$D_{\mathrm{rec}} = \sigma(Z^{(4)}), \qquad \mathcal{L}_{\mathrm{SDR}} = \mathrm{SmoothL1}\left(D_{\mathrm{rec}}, \mathrm{sg}(D_{\mathrm{anc}})\right), \tag{8}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient. This process incorporates complementary multi-level evidence while preserving the initial referential anchor.

3.3.2 Bounded Role-Aware Dense Correction

The spatial role responses refine D_{rec} through a bounded residual. After query-wise mean pooling and resizing, we compute

$$C = \tanh\left(G_C\left([\uparrow \bar{R}^+, \uparrow \bar{R}^-]\right)\right), \qquad \alpha = \alpha_{\max}\sigma(a), \tag{9}$$

where $\bar{R}^\pm = \mathrm{Mean}_{\mathrm{query}}(R^\pm)$, G_C is a lightweight correction head, and a is a learnable scalar.

The final dense prompt is

$$D_{\mathrm{A}} = \mathrm{clip}_{[0,1]}\left(D_{\mathrm{rec}} + \alpha C \odot D_{\mathrm{rec}}\right). \tag{10}$$

Since $C \in [-1, 1]$ and α is bounded, this multiplicative residual can only modulate the established response rather than construct an independent dense state.

We further preserve consistency with the initial dense anchor through

$$\mathcal{L}_A = \mathrm{Mean}\left(|D_{\mathrm{A}} - D_{\mathrm{anc}}|\right). \tag{11}$$

Unlike $\mathcal{L}_{\mathrm{SDR}}$, D_{anc} is not detached in this objective, allowing the initial anchor and the subsequent correction pathway to be jointly optimized.

The complete PAEG objective is

$$\mathcal{L}_{\mathrm{PAEG}} = \lambda_{\mathrm{SDR}}\mathcal{L}_{\mathrm{SDR}} + \lambda_A\mathcal{L}_A. \tag{12}$$

3.4 Evidence-Calibrated Prompt Decoding

Reliable prompt decoding requires both an expression-discriminative dense anchor and controlled utilization of auxiliary prompt evidence. ECPD therefore calibrates the initial dense anchor during training and performs bounded candidate-based refinement during decoding.

3.4.1 Discriminative Dense-Anchor Calibration

During training, we construct a mismatched expression T^- as a discriminative reference for the matched expression $T^+ = T$. Applying the same grounding pathway to (I, T^+) and (I, T^-) produces the dense responses D_{anc}^+ and D_{anc}^-, respectively. The mismatched response is used only as a detached reference and receives no gradient.

Let Y_{anc} denote the ground-truth mask resized by nearest-neighbor interpolation to the spatial resolution of D_{anc}. For a dense probability response D, we define its foreground–background grounding contrast as

$$S(D, Y_{\mathrm{anc}}) = \frac{\langle D, Y_{\mathrm{anc}}\rangle}{\|Y_{\mathrm{anc}}\|_1} - \frac{\langle D, 1 - Y_{\mathrm{anc}}\rangle}{\|1 - Y_{\mathrm{anc}}\|_1}. \tag{13}$$

The discriminative calibration objective is

$$\mathcal{L}_{\mathrm{cal}} = \left[m_d - S(D_{\mathrm{anc}}^+, Y_{\mathrm{anc}}) + \mathrm{sg}\left(S(D_{\mathrm{anc}}^-, Y_{\mathrm{anc}})\right)\right]_+, \tag{14}$$

where m_d is the calibration margin. This objective encourages the matched expression to produce a stronger grounding contrast than the mismatched reference.

3.4.2 AUXILIARY DECODER REFINEMENT

We construct four auxiliary tokens from $\bar{p}^+$ with learnable slot embeddings and jointly refine them before applying bounded exclusion-conditioned attenuation from $\bar{p}^- = \mathrm{Mean}_{\mathrm{query}}(P^-)$. Thus, exclusion evidence can attenuate an established positive-anchored token but cannot independently construct one. Denoting the resulting tokens by $P_{\mathrm{aux}} = \{h_k\}_{k=1}^4$, we form $P_{\mathrm{aug}} = [P^+; P^-; P_{\mathrm{aux}}]$ with $K_{\mathrm{aug}} = 12$.

Let $\bar{P}_{\mathrm{aug}}$ denote the token-wise ℓ_2-normalized representation of P_{aug}. To prevent prompt collapse, we define

$$\mathcal{L}_{\mathrm{div}} = \frac{1}{K_{\mathrm{aug}}^2} \left\| \bar{P}_{\mathrm{aug}} \bar{P}_{\mathrm{aug}}^\top - I_{K_{\mathrm{aug}}} \right\|_F^2 . \tag{15}$$

The mismatched expression T^- is also used for sparse contrastive regularization. We form $u_{\mathrm{mat}} = \mathrm{Mean}(P_{\mathrm{aug}}(T^+))$ and $u_{\mathrm{mis}} = \mathrm{Mean}(P_{\mathrm{aug}}(T^-))$, and apply

$$\mathcal{L}_{\mathrm{ctr}} = \left[m_s - (1 - \cos\left(u_{\mathrm{mat}}, \mathrm{sg}(u_{\mathrm{mis}})\right)) \right]_+ . \tag{16}$$

Using the same image features and dense prompt D_{A}, the shared SAM2 decoder produces Z_{main} from P_{aug} and a candidate logit map Z_k from each auxiliary token h_k. The candidates are weighted by their overlaps with the positive and exclusion role responses:

$$\begin{aligned} o_k^+ &= \mathrm{Mean}\left(\sigma(Z_k) \odot \sigma(\uparrow \bar{R}^+)\right), \qquad o_k^- = \mathrm{Mean}\left(\sigma(Z_k) \odot \sigma(\uparrow \bar{R}^-)\right), \\ \omega_k &= \mathrm{Softmax}_k\left(\gamma o_k^+ - o_k^-\right), \qquad Z_{\mathrm{cand}} = \sum_{k=1}^{4} \omega_k \, \mathrm{sg}(Z_k). \end{aligned} \tag{17}$$

Candidate logits are detached, while their weights remain differentiable through the role responses.

Since candidate weighting depends on the spatial role responses, we further define $\Delta R = R_{\max}^+ - R_{\max}^-$ and $S_R = 2Y_R - 1$, and impose

$$\mathcal{L}_{\mathrm{conf}} = \mathrm{Mean}\left[m_c - S_R \Delta R \right]_+ . \tag{18}$$

The candidate prior refines the main prediction through bounded interpolation:

$$Z_{\mathrm{out}} = Z_{\mathrm{main}} + \alpha_d \left(Z_{\mathrm{cand}} - Z_{\mathrm{main}} \right), \qquad \alpha_d = \alpha_d^{\max} \sigma(a_d). \tag{19}$$

The refined logit map Z_{out} is finally upsampled to the image resolution and converted into the segmentation mask $\widehat{Y}$. The complete ECPD objective is

$$\mathcal{L}_{\mathrm{ECPD}} = \lambda_{\mathrm{cal}}(t)\mathcal{L}_{\mathrm{cal}} + \lambda_{\mathrm{div}}\mathcal{L}_{\mathrm{div}} + \lambda_{\mathrm{ctr}}\mathcal{L}_{\mathrm{ctr}} + \lambda_{\mathrm{conf}}\mathcal{L}_{\mathrm{conf}}. \tag{20}$$

3.5 TRAINING OBJECTIVE AND INFERENCE

Following RS2-SAM2 (Rong et al., 2026), we retain its original segmentation objective $\mathcal{L}_{\mathrm{seg}}$. The complete training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{seg}} + \mathcal{L}_{\mathrm{RAEF}} + \mathcal{L}_{\mathrm{PAEG}} + \mathcal{L}_{\mathrm{ECPD}}. \tag{21}$$

At inference, AREM operates solely on (I, T) and generates all role-specific evidence internally; the mismatched-expression branch is used only during training.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We evaluate AREM on RefSegRS (Yuan et al., 2024) and RRSIS-D (Liu et al., 2024). RefSegRS contains 2,172 training, 431 validation, and 1,817 test samples, with an image resolution of 512×512. RRSIS-D is a larger-scale dataset comprising 12,181 training, 1,740 validation, and 3,481 test samples, with an image resolution of 800×800. Both datasets cover diverse target scales and complex remote-sensing scenes.

Table 1: Quantitative comparison on RefSegRS. Best and second-best results are shown in **bold** and underline, respectively; "–" denotes unreported results.

Method	P@0.5		P@0.6		P@0.7		P@0.8		P@0.9		oIoU		mIoU	
	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test
RRN Li et al. (2018)	55.43	30.26	42.98	23.01	23.11	14.87	13.72	7.17	2.64	0.98	69.24	65.06	50.81	41.88
LSCM Hui et al. (2020)	56.82	31.54	41.24	20.41	21.85	9.51	12.11	5.29	2.51	0.84	62.82	61.27	40.59	35.54
CMPC+ Liu et al. (2022)	56.84	49.19	37.59	28.31	20.42	15.31	10.67	8.12	2.78	2.55	70.62	66.53	47.13	43.65
ETRIS Xu et al. (2023)	54.99	35.77	35.03	23.00	25.60	13.98	12.53	6.44	1.62	1.10	72.89	65.96	54.03	43.11
CRIS Wang et al. (2022)	53.13	35.77	36.19	24.11	24.36	14.36	11.83	6.38	2.55	1.21	72.14	65.87	53.74	43.26
CARIS Liu et al. (2023)	68.45	45.40	47.10	27.19	25.52	15.08	14.62	8.87	3.71	1.98	75.79	69.74	54.30	42.66
RIS-DMMI Hu et al. (2023)	86.17	63.89	74.71	44.30	38.05	19.81	18.10	6.49	3.25	1.00	74.02	68.58	65.72	52.15
RMSIN Liu et al. (2024)	92.28	71.07	84.95	57.26	73.15	34.87	36.34	13.15	8.10	2.75	79.65	72.68	73.04	57.91
CroBIM Dong et al. (2024)	93.04	75.89	87.70	61.42	66.13	34.07	26.91	12.99	5.80	2.75	77.95	72.33	71.93	59.77
Enti-TwistNet Jia et al. (2025)	–	77.92	–	69.29	–	44.82	–	26.83	–	7.58	–	78.85	–	64.02
SBANet Li et al. (2025)	–	77.02	–	–	–	44.15	–	–	–	8.97	–	79.86	–	62.73
DiffRIS Dong et al. (2026)	94.66	74.57	92.11	64.56	84.22	46.51	57.08	24.49	14.62	5.78	85.63	76.18	78.28	62.38
CroBIM-U Sun et al. (2026)	94.11	76.68	88.54	62.54	68.45	34.81	27.22	13.50	6.57	3.26	78.03	73.81	72.55	60.08
ICIPNet Ren et al. (2026)	96.23	82.39	93.93	73.77	87.40	57.47	51.81	22.84	9.54	4.64	83.78	76.55	78.25	66.08
RS2-SAM2 Rong et al. (2026)	95.36	84.31	94.90	79.42	92.58	70.89	83.76	55.70	36.66	21.19	88.03	80.87	85.21	73.90
AREM (ours)	**96.98**	**86.52**	**96.52**	**82.11**	**94.90**	**75.40**	**90.72**	**62.52**	**63.34**	**29.22**	**90.54**	**83.23**	**88.48**	**76.65**

Table 2: Quantitative comparison on RRSIS-D. Best and second-best results are shown in **bold** and underline, respectively; "–" denotes unreported results.

Method	P@0.5		P@0.6		P@0.7		P@0.8		P@0.9		oIoU		mIoU	
	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test
RRN Li et al. (2018)	51.09	51.07	42.47	42.11	33.04	32.77	20.80	21.57	6.14	6.37	66.53	66.43	46.06	45.64
LSCM Hui et al. (2020)	57.12	56.02	48.04	46.25	37.87	37.70	26.37	25.28	7.93	8.27	69.28	69.05	50.36	49.92
BRINet Yang et al. (2020)	58.79	56.90	49.54	48.77	39.65	39.12	28.21	27.03	9.18	8.73	70.73	69.88	51.14	49.65
CMPC+ Liu et al. (2022)	59.19	57.65	49.36	47.51	38.67	36.97	25.91	24.33	8.16	7.78	70.14	68.64	51.41	50.24
LGCE Yuan et al. (2024)	68.10	67.65	60.52	61.53	52.24	51.45	42.24	39.62	23.85	23.33	76.68	76.34	60.16	59.37
RIS-DMMI Hu et al. (2023)	70.40	68.74	63.05	60.96	54.14	50.33	41.95	38.38	23.85	21.63	77.01	76.20	60.72	60.12
LAVT Yang et al. (2022)	69.54	69.52	63.51	63.63	53.16	53.29	43.97	41.60	24.25	24.94	77.59	77.19	61.46	61.04
RMSIN Liu et al. (2024)	74.66	74.26	68.22	67.25	57.41	55.93	45.29	42.55	24.43	24.53	78.27	77.79	65.10	64.20
CroBIM Dong et al. (2024)	74.94	74.58	67.64	67.57	57.18	55.59	44.66	41.63	24.60	23.56	76.94	75.99	65.05	64.46
MAFN Shi & Zhang (2025)	76.32	75.27	69.31	68.14	58.33	56.79	44.54	43.49	24.71	23.76	78.33	77.41	66.03	64.76
FIANet Lei et al. (2025)	–	74.46	–	66.96	–	56.31	–	42.83	–	24.13	–	76.91	–	64.01
ICIPNet Ren et al. (2026)	75.17	73.84	67.83	67.56	56.60	56.11	44.06	43.20	24.99	24.63	77.69	76.87	65.19	63.93
CroBIM-U Sun et al. (2026)	76.31	75.60	68.30	67.68	57.67	56.47	44.97	42.57	24.89	24.16	77.83	76.70	66.07	65.07
RS2-SAM2 Rong et al. (2026)	79.25	77.56	74.08	72.34	63.85	61.76	50.57	47.92	30.40	29.73	**80.16**	**78.99**	68.81	66.72
AREM (ours)	**80.06**	**78.14**	**74.37**	**73.31**	63.85	**63.11**	**51.67**	**47.95**	**30.86**	**30.25**	79.23	78.40	**69.22**	**67.23**

Evaluation Metrics. Following prior RRSIS studies (Yuan et al., 2024; Liu et al., 2024), we report mean Intersection-over-Union (mIoU), overall Intersection-over-Union (oIoU), and Precision@X (P@X). mIoU averages IoU across samples, whereas oIoU measures dataset-level pixel overlap. P@X denotes the percentage of samples whose IoU reaches or exceeds $X \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$.

Implementation Details. Following RS2-SAM2 (Rong et al., 2026), we adopt BEiT-3 Large and SAM2 Hiera-L (Wang et al., 2023; Ravi et al., 2025), with input resolutions of 224×224 and 1024×1024, respectively; dense prompts are resized to 256×256. Models are trained using AdamW (Loshchilov & Hutter, 2019) and BF16 with DeepSpeed ZeRO-2 (Rajbhandari et al., 2020) on four RTX 4090 GPUs, using one sample per GPU. RefSegRS and RRSIS-D are trained for 60 and 40 epochs with base learning rates of 5×10^{-5} and 1×10^{-5}, and prompt/fusion learning rates of 1×10^{-4} and 5×10^{-5}, respectively. Learning rates are decayed by a factor of 0.1 near the end of training. AREM-specific hyperparameters and schedules are detailed in Appendix A.1.

4.2 Comparison with Representative Methods

RefSegRS. As shown in Table 1, AREM achieves the best results across all metrics on both splits, reaching 88.48/90.54 mIoU/oIoU on validation and 76.65/83.23 on test. Compared with RS2-SAM2, it improves test mIoU and oIoU by 2.75 and 2.36 points, respectively, with larger gains of 6.82 and 8.03 points at P@0.8 and P@0.9.

RRSIS-D. As shown in Table 2, AREM obtains the highest mIoU on both splits, reaching 69.22 on validation and 67.23 on test. It also improves RS2-SAM2 across all test precision thresholds, including a 1.35-point gain at P@0.7, while remaining slightly lower in oIoU.

Table 3: Progressive ablation of AREM. Each stage is enabled on top of its preceding stages.

Method	RAEF	PAEG	ECPD	mIoU	oIoU	P@0.5	P@0.7	P@0.9
Baseline				72.02	80.24	84.42	68.19	12.38
AREM	✓			72.89	80.61	84.98	70.23	13.26
AREM	✓	✓		75.45	82.40	85.47	72.65	**29.66**
AREM	✓	✓	✓	**76.65**	**83.23**	**86.52**	**75.40**	29.22

Table 4: Transferability of AREM across architectures and datasets.

Dataset	Split	Model	mIoU	oIoU	P@0.5	P@0.7	P@0.9
RefSegRS	Test	EVF-SAM	27.86	47.47	18.05	4.73	0.28
		+ AREM	**33.26**	**52.08**	**26.25**	**6.27**	**0.33**
RefCOCO	TestA	Baseline	77.96	76.58	87.86	81.85	42.25
		+ AREM	**80.58**	**79.78**	**90.63**	**85.81**	**46.21**
	TestB	Baseline	72.36	71.08	80.33	71.23	39.00
		+ AREM	**75.50**	**74.56**	**83.98**	**75.58**	**43.71**

4.3 Ablation Studies and Analysis

Progressive component ablation. To investigate the contributions of the three components, we construct a baseline comprising the union multimodal encoder and the SAM2 segmentation pipeline, with all AREM components removed. Because RAEF, PAEG, and ECPD have explicit dependencies, we introduce them cumulatively in Table 3. All variants follow the same training and evaluation protocol. RAEF raises mIoU from 72.02 to 72.89, while PAEG provides the largest stage-wise gains of 2.56 mIoU and 1.79 oIoU points. With ECPD, the complete AREM reaches 76.65 mIoU and 83.23 oIoU, outperforming the baseline by 4.63 and 2.99 points, respectively. Although RAEF+PAEG achieves a slightly higher P@0.9, the complete model performs best on all remaining metrics, supporting the complementary contributions of the three components.

Table 5: Inference efficiency of AREM.

Method	Params (M)	GFLOPs	Latency (ms)
Baseline	934.81	2120.22	161.86
AREM	940.99	2138.40	184.07
Change	+0.66%	+0.86%	+13.72%

Mechanism analysis. Figure 3 provides component-wise diagnostics of AREM. RAEF is the only query design that achieves positive selectivity for both evidence roles, indicating that supporting and exclusion queries preferentially respond to their intended regions. Under the same visual state, the correct P^+ improves semantic-response alignment over shuffled and zero controls by 0.0067 and 0.0036, respectively, supporting its contribution to positive-conditioned grounding. ECPD further increases the mean correct–mismatch anchor gap from 0.168 to 0.316, indicating stronger expression discrimination. Additional mechanism analyses are provided in Appendix C.

Transferability. All comparisons use matched training and evaluation settings, with further details provided in Appendix A.3. Table 4 shows that integrating AREM into EVF-SAM (Zhang et al., 2026) improves mIoU and oIoU by 5.40 and 4.61 points on RefSegRS. On RefCOCO (Yu et al., 2016), a natural-image referring segmentation benchmark outside the remote-sensing domain, AREM consistently outperforms the baseline on TestA and TestB, including gains of 3.96 and 4.71 points at P@0.9, respectively. These results support the applicability of AREM across model architectures and application domains.

Efficiency. As shown in Table 5, AREM adds 6.18M parameters (+0.66%) and 18.18 GFLOPs (+0.86%) over the baseline. Its inference latency increases from 161.86 ms to 184.07 ms, corresponding to a 13.72% overhead. Thus, AREM introduces limited model and computational growth, although its evidence formation and refinement operations incur measurable runtime cost. Appendix C.8 provides stage-wise efficiency profiling to quantify the cost of each AREM stage.

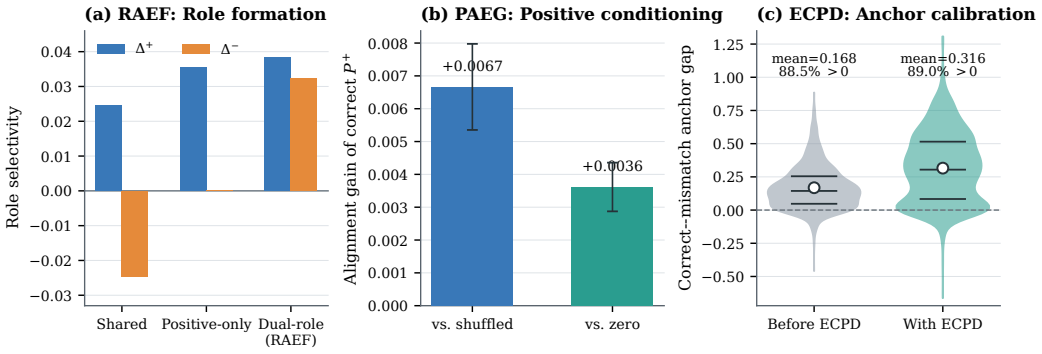


Figure 3: **Mechanistic validation of AREM on RefSegRS.** (a) Role selectivity under different query designs. (b) Paired alignment gains of the correct P^+ over shuffled and zero-conditioning controls. (c) Correct–mismatch dense-anchor gaps before and after ECPD. Error bars indicate 95% bootstrap confidence intervals, and violins show sample distributions.

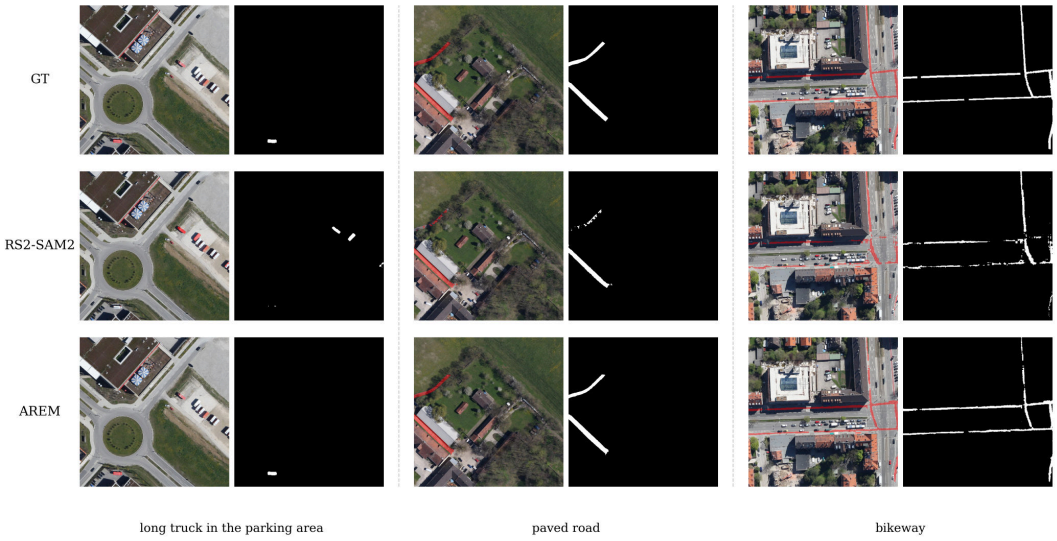


Figure 4: **Qualitative comparison on RefSegRS.** Rows show Ground Truth, RS2-SAM2, and Ours.

4.4 Qualitative Analysis

Figure 4 compares AREM with RS2-SAM2 on representative RefSegRS examples. AREM more consistently isolates the intended referent in cluttered scenes, reducing leakage to neighboring plausible candidates while recovering missing parts of thin or extended structures. These examples show improvements in both referent selection and mask completeness, consistent with the gains at strict overlap thresholds. Additional visualizations and failure cases are provided in Appendix D.

5 Conclusion

We introduced *Asymmetric Referential Evidence Modeling* (AREM), which assigns different functional authority to referent-supporting and exclusion-oriented evidence, thereby separating referent construction from corrective intervention. RAEF forms role-specific evidence, PAEG constructs positive-anchored dense grounding with bounded correction, and ECPD calibrates and refines prompt decoding. Together, these components improve referent identification in ambiguous remote-sensing scenes. Extensive experiments across datasets and architectures validate the effectiveness and practical efficiency of AREM and support the broader applicability of asymmetric evidence authority.

Reproducibility Statement

Implementation details and experimental protocols are provided in Sections 3–4 and Appendix A. An anonymized code snapshot with configurations and training and evaluation scripts is included in the supplementary material.

AI Use Statement

Generative AI tools were used for literature retrieval and language polishing. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the final manuscript.

References

Keyan Chen, Jiafan Zhang, Chenyang Liu, Zhengxia Zou, and Zhenwei Shi. Rsrefseg: Referring remote sensing image segmentation with foundation models. *arXiv preprint arXiv:2501.06809*, 2025.

Keyan Chen, Chenyang Liu, Bowen Chen, Jiafan Zhang, Zhengxia Zou, and Zhenwei Shi. Rsrefseg 2: Decoupling referring remote sensing image segmentation with foundation models. *IEEE Transactions on Geoscience and Remote Sensing*, 64:1–20, 2026.

Zhe Dong, Yuzhe Sun, Yanfeng Gu, and Tianzhu Liu. Cross-modal bidirectional interaction model for referring remote sensing image segmentation. *arXiv preprint arXiv:2410.08613*, 2024.

Zhe Dong, Yu-Zhe Sun, Tian-Zhu Liu, and Yan-Feng Gu. Diffris: Enhancing referring remote sensing image segmentation with pre-trained text-to-image diffusion models. *Fundamental Research*, 6(4): 2571–2586, 2026. ISSN 2667-3258. doi: https://doi.org/10.1016/j.fmre.2025.11.017.

Ziyang Gao, Zhizhuo Jiang, Jingjing Chang, Yixin Yang, Yuwen Pan, Yong-Qiang Mao, Yu Liu, and Hai-Bao Chen. Dicor: Decoupled referent disambiguation and contour recalibration for efficient referring remote sensing image segmentation. *arXiv preprint arXiv:2608.12980*, 2026.

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 *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 4044–4054. IEEE, 2023.

Tianrui Hui, Si Liu, Shaofei Huang, Guanbin Li, Sansi Yu, Faxi Zhang, and Jizhong Han. Linguistic structure guided context modeling for referring image segmentation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020*, pp. 59–75, Cham, 2020. Springer International Publishing.

Yuyu Jia, Qing Zhou, Junyu Gao, and Qi Wang. Entity-guided attention twisting network for referring remote sensing image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 63: 1–10, 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Sen Lei, Xinyu Xiao, Tianlin Zhang, Heng-Chao Li, Zhenwei Shi, and Qing Zhu. Exploring fine-grained image-text alignment for referring remote sensing image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–11, 2025.

Kun Li, George Vosselman, and Michael Ying Yang. Scale-wise bidirectional alignment network for referring remote sensing image segmentation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 226:350–363, 2025.

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 (CVPR)*, June 2018.

Si Liu, Tianrui Hui, Shaofei Huang, Yunchao Wei, Bo Li, and Guanbin Li. Cross-modal progressive comprehension for referring segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(9):4761–4775, 2022.

Sihan Liu, Yiwei Ma, Xiaoqing Zhang, Haowei Wang, Jiayi Ji, Xiaoshuai Sun, and Rongrong Ji. Rotated multi-scale interaction network for referring remote sensing image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26648–26658, 2024. doi: 10.1109/CVPR52733.2024.02517.

Sun'ao Liu, Yiheng Zhang, Zhaofan Qiu, Hongtao Xie, Yongdong Zhang, and Ting Yao. CARIS: context-aware referring image segmentation. In Abdulmotaleb El Saddik, Tao Mei, Rita Cucchiara, Marco Bertini, Diana Patricia Tobon Vallejo, Pradeep K. Atrey, and M. Shamim Hossain (eds.), *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pp. 779–788. ACM, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Tingzhang Luo, Ruizhong Liu, Yichao Liu, Cheng Fan, Yu Liu, and Jianyuan Guo. Cross: Cascaded distillation and dual-constraint grounding for remote sensing referring segmentation. *arXiv preprint arXiv:2608.03147*, 2026.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. ZeRO: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1–16, 2020.

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

Biaoyu Ren, Qingsheng Wang, Cun Xu, Dingkang Yang, and Wenxuan Wang. Image-conditioned instance prompt network for referring remote sensing image segmentation. *CoRR*, abs/2605.24532, 2026. doi: 10.48550/ARXIV.2605.24532.

Fu Rong, Meng Lan, Qian Zhang, and Lefei Zhang. Rs2-sam2: Customized sam2 for referring remote sensing image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 8751–8759, 2026. doi: 10.1609/aaai.v40i11.37828.

Leideng Shi and Juan Zhang. Multimodal-aware fusion network for referring remote sensing image segmentation. *IEEE Geosci. Remote. Sens. Lett.*, 22:1–5, 2025. doi: 10.1109/LGRS.2025.3527485.

Yuzhe Sun, Zhe Dong, Haochen Jiang, Yanfeng Gu, and Tianzhu Liu. Crobim-u: Uncertainty-driven referring remote sensing image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 64:1–16, 2026. doi: 10.1109/TGRS.2026.3665448.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19175–19186, 2023.

Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11676–11685, 2022. doi: 10.1109/CVPR52688.2022.01139.

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 *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 17457–17466. IEEE, 2023.

Xianghui Yang, Bairun Wang, Kaige Chen, Xinchi Zhou, Shuai Yi, Wanli Ouyang, and Luping Zhou. Brinet: Towards bridging the intra-class and inter-class gaps in one-shot segmentation, 2020.

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 (CVPR)*, pp. 18155–18165, June 2022.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, pp. 69–85, 2016.

Zhenghang Yuan, Lichao Mou, Yuansheng Hua, and Xiao Xiang Zhu. Rrsis: Referring remote sensing image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–12, 2024. doi: 10.1109/TGRS.2024.3369720.

Y. Zhang, T. Cheng, L. Zhu, L. Liu, H. Liu, L. Ran, X. Chen, W. Liu, and X. Wang. Evf-sam: Early vision-language fusion for text-prompted segment anything model. *Image and Vision Computing*, 173:106028, 2026.

Table 6: AREM-specific settings used in the reported experiments.

Component	Setting	Value
RAEF	Queries per role	4
	Spatial role weight λ_{reg}	0.05
	Exclusion balance η_-	0.5
PAEG	Correction bound $\alpha_{\max}$	0.35
	Reconciliation weight λ_{SDR}	0.002
	Anchor-preservation weight λ_A	0.016
ECPD	Auxiliary candidate tokens	4
	Decoder interpolation bound $\alpha_d^{\max}$	0.30
	Positive-overlap coefficient γ	0.5
	Maximum calibration weight λ_{cal}	0.03
	Calibration margin m_d	0.05
	Sparse contrast weight λ_{ctr}	0.03
	Sparse contrast margin m_s	0.2
	Diversity weight λ_{div}	0.005
	Role-conflict weight λ_{conf}	0.01
	Role-conflict margin m_c	0.12

Table 7: Operator configurations used in AREM.

Operator	Configuration
Π_V^ℓ / Π_P^ℓ	$1024 \to 256$ / $256 \to 256$
G_ℓ	LN–256–GELU–1
ϕ_ℓ	Conv $2 \to 16 \to 1$
$\mathcal{M}$	$512 \to 256 \to 256$
$\mathcal{T}$	2 layers, 8 heads, FFN 1024
$\mathcal{S}$	$512 \to 256 \to 1$

A Additional Implementation Details

A.1 AREM-Specific Settings

Table 6 summarizes the architecture constants, loss coefficients, and margins used by the complete model. Only settings active in the final configuration are reported; the corresponding loss definitions are provided in the main paper.

The calibration weight is zero during the first 10% of training, increases linearly to 0.03 between 10% and 30%, and remains fixed thereafter. All reported checkpoints are selected using the same joint validation criterion based on mIoU and oIoU within their respective protocols.

Operator configurations. Table 7 records the lightweight operators used by PAEG and ECPD. The final layers of ϕ_ℓ, $\mathcal{M}$, and $\mathcal{S}$ are zero-initialized; $\mathcal{T}$ uses pre-normalization, GELU, and zero dropout, while slot embeddings follow $\mathcal{N}(0, 0.02^2)$. Before the logit transform, D_{anc} is clamped to $[10^{-4}, 1 - 10^{-4}]$; auxiliary candidate masks are decoded without gradient retention.

A.2 Dataset-Specific Mismatched Expressions

Mismatched expressions are used only as detached training references. For RefSegRS, we first attempt a caption permutation when a local batch contains multiple distinct expressions. With a single local sample, an indexed category-, attribute-, or relation-level negative is used when available; otherwise, the expression is sampled from a history buffer containing at most 4096 captions. For RRSIS-D, we construct minimal structured counterfactuals by inverting a relation or comparison, an absolute spatial term, or a single attribute. Samples without a valid transformation are excluded from the negative-reference losses. For RefCOCO (Yu et al., 2016), an expression referring to another

instance in the same image is preferred, followed by the same fallback policy for uncovered samples. None of these negative-expression sources is queried at inference.

A.3 Transfer Evaluation Protocols

Cross-architecture transfer. To evaluate cross-architecture transferability, we integrate AREM into the prompt construction pathway of EVF-SAM while retaining its original backbone, data split, optimization, and evaluation protocol. EVF-SAM and EVF-SAM+AREM are initialized from the same SAM2 Hiera-L and BEiT-3 Large pretrained weights and trained on the RefSegRS training split for 60 epochs. Images are resized to 1024×1024 for the SAM2 branch and 224×224 for the BEiT-3 branch, with ImageNet normalization.

Both models are trained on four GPUs with one sample per GPU using BF16 and AdamW. The base learning rate is 5×10^{-5}, while the prompt, adapter, fusion, and AREM parameters use 1×10^{-4}; weight decay is 0.01. The BEiT-3 encoder, text projections, prompt-generation layers, VM-token modules, SAM2 prompt encoder and mask decoder, and image-encoder adapters are optimized, while the remaining SAM2 image-encoder parameters are frozen. EVF-SAM+AREM uses the complete AREM configuration reported in Table 6, whereas the baseline retains the original EVF-SAM prompt pathway without AREM. The best validation checkpoint is used for final test evaluation.

Cross-dataset transfer. For cross-dataset transfer, the baseline and AREM are initialized from the same pretrained weights and independently fine-tuned on the RefCOCO-UNC training split for five epochs. Images are resized to 1024×1024 for the SAM2 branch and 224×224 for the union BEiT-3 encoder. Both models are trained on four GPUs with one sample per GPU using BF16, DeepSpeed ZeRO-2, and AdamW. The base learning rate is 1×10^{-5}, while the fusion and AREM parameters use 5×10^{-5}; weight decay is 0.01.

The union BEiT-3 encoder, SAM2 prompt encoder, and mask decoder are fine-tuned, while the SAM2 image encoder is frozen except for its MLP and spatial adapters. The baseline and AREM otherwise share the same data, initialization, and optimization settings. AREM uses the same architectural and loss configuration as in the main experiments, together with the RefCOCO-specific mismatched-expression policy described in Appendix A.2. The checkpoint maximizing the sum of validation mIoU and oIoU is evaluated on the official TestA and TestB splits. Within each setting, the baseline and AREM differ only in whether AREM is enabled.

B Theoretical Analysis of Bounded Evidence Authority

This section characterizes the deterministic authority constraints imposed by PAEG and ECPD, showing how the corrective pathways are structurally bounded when modifying an existing referential state.

B.1 Bounded Dense Correction

For a spatial location x, let $d = D_{\mathrm{rec}}(x)$, $c = C(x)$, and $d_A = D_{\mathrm{A}}(x)$. By construction, $d \in [0, 1]$, $c \in [-1, 1]$, and $0 \le \alpha \le \alpha_{\max} < 1$.

Proposition 1 (positive-relative correction). The corrected response satisfies

$$(1-\alpha)d \le d_A \le \min\{1, (1+\alpha)d\}, \qquad |d_A - d| \le \alpha d \le \alpha_{\max} d. \tag{22}$$

Consequently, $d = 0$ implies $d_A = 0$: the correction pathway cannot create a response where the reconciled dense state has no support.

Proof. Before clipping, Eq. (10) gives $d(1 + \alpha c)$. Since $c \in [-1, 1]$, its multiplicative factor lies in $[1 - \alpha, 1 + \alpha]$ and remains nonnegative. Clipping can only replace a value above one by one. If no clipping occurs, the absolute change is $\alpha|c|d \le \alpha d$. If upper clipping occurs, then $(1 + \alpha)d > 1$, which implies $1 - d < \alpha d$. Both cases yield Eq. (22). □

The same bound gives a threshold-stability interpretation. For any $\tau \in (0, 1)$,

$$d \ge \frac{\tau}{1-\alpha} \implies d_A \ge \tau, \qquad d \le \frac{\tau}{1+\alpha} \implies d_A \le \tau. \tag{23}$$

Table 8: Effect of dense evidence authority on the RefSegRS test split.

Dense intervention	mIoU	oIoU	P@0.5	P@0.7	P@0.9
Direct exclusion suppression	73.92	81.06	**86.19**	71.66	15.30
Dual-polarity dense rewrite	72.14	81.14	81.07	68.63	21.68
Positive-preserved correction	**75.63**	**82.81**	86.02	**74.08**	**24.82**

Thus, only responses inside the interval $[\tau/(1+\alpha), \tau/(1-\alpha)]$ can cross the threshold through bounded correction. This formalizes the intended distinction between constructing D_{rec} and correcting it.

B.2 Bounded Candidate Interpolation

The role-aware weights in Eq. (17) satisfy $\omega_k \geq 0$ and $\sum_k \omega_k = 1$. Hence, at every spatial location,

$$\min_k Z_k \leq Z_{\mathrm{cand}} \leq \max_k Z_k, \tag{24}$$

so the candidate prior lies in the pointwise convex hull of the candidate logits.

Proposition 2 (non-extrapolative decoder refinement). Let $m = Z_{\mathrm{main}}(x)$, $q = Z_{\mathrm{cand}}(x)$, and $z = Z_{\mathrm{out}}(x)$. For $0 \leq \alpha_d \leq \alpha_d^{\max} < 1$,

$$z = (1-\alpha_d)m + \alpha_d q, \qquad \min\{m, q\} \leq z \leq \max\{m, q\}, \tag{25}$$

and

$$|z - m| = \alpha_d |q - m| \leq \alpha_d^{\max} |q - m|. \tag{26}$$

Proof. Equation (19) is a convex combination of m and q, which proves the interval bound. Subtracting m gives $z - m = \alpha_d(q - m)$ and yields the displacement bound. □

Since the sigmoid derivative is at most $1/4$, the corresponding probability change is also bounded:

$$|\sigma(z) - \sigma(m)| \leq \frac{\alpha_d^{\max}}{4} |q - m|. \tag{27}$$

Moreover, when m and q have opposite signs, the main binary decision is guaranteed to be preserved whenever

$$|m| > \frac{\alpha_d^{\max}}{1 - \alpha_d^{\max}} |q|. \tag{28}$$

Together, Eqs. (24)– (28) show that candidate aggregation remains inside the candidate ensemble, while its subsequent interpolation exerts only bounded, non-extrapolative influence relative to the main logits.

C Additional Experiments

C.1 Dense Evidence Authority

To examine how role-specific evidence should modify an established dense referential state, we compare three intervention strategies under otherwise identical settings: direct exclusion suppression, unconstrained dual-polarity rewriting, and positive-preserved bounded correction. As shown in Table 8, positive-preserved correction achieves the best overall performance, indicating that preserving the positive anchor while restricting role evidence to bounded correction is more reliable than direct suppression or unrestricted rewriting. This result supports the authority design adopted in PAEG.

Bounded-correction control. We further retrain complete AREM with $\alpha_{\max} = 0$, such that $D_{\mathrm{A}} = D_{\mathrm{rec}}$, while keeping all other settings unchanged. Table 9 shows that bounded correction improves overall and moderate-threshold performance, with a trade-off at P@0.9.

Table 9: Ablation of bounded correction under the complete AREM configuration on the RefSegRS test split.

Configuration	mIoU	oIoU	P@0.5	P@0.7	P@0.9
Without correction ($\alpha_{\max} = 0$)	75.61	82.91	84.31	72.15	**31.70**
AREM ($\alpha_{\max} = 0.35$)	**76.65**	**83.23**	**86.52**	**75.40**	29.22

Table 10: Comparison of alternative semantic-depth grounding strategies on the RefSegRS test split.

Semantic-depth strategy	mIoU	oIoU	P@0.5	P@0.7	P@0.9
Full-range deep-to-shallow cascade	76.05	82.86	**87.01**	74.08	24.44
Mid–high semantic cascade	75.06	82.41	85.25	73.31	21.68
High-semantic cascade	74.13	82.06	83.98	70.45	25.70
Shallow-to-deep semantic maturation	72.94	80.72	84.87	70.17	14.36
Positive-anchored bidirectional reconciliation	**76.65**	**83.23**	86.52	**75.40**	**29.22**

C.2 Semantic-Depth Trajectories

To examine how representations at different semantic depths should contribute to dense grounding, we compare several one-way cascades with the proposed positive-anchored bidirectional reconciliation. The deep-to-shallow strategies start from a high-level referential state and progressively introduce lower-level evidence, with the mid–high and high-semantic variants restricting this process to deeper representation ranges. Shallow-to-deep maturation reverses this direction, whereas our bidirectional strategy acquires complementary intermediate and shallow evidence before reconciling it with the deep referential anchor. As shown in Table 10, the full-range deep-to-shallow cascade provides a strong one-way alternative, while the proposed strategy performs best on most metrics, supporting bidirectional semantic-depth reconciliation for more reliable dense grounding.

C.3 Mismatched-Expression Construction

To examine the sensitivity of expression-conditioned calibration to the construction of T^-, we compare the default mixed sampler with several type-specific counterfactual strategies under the complete AREM configuration. The default sampler combines in-batch permutations, indexed semantic negatives, and caption-history fallback. As shown in Table 11, it provides the strongest overall performance across average-overlap and moderate-threshold metrics. Relation-based counterfactuals achieve higher P@0.9 but slightly lower performance on the remaining metrics, indicating a trade-off between strict discrimination and overall robustness.

C.4 Equal-Capacity Query Designs

To examine whether the benefit of RAEF arises from functional role assignment rather than query capacity, we compare shared, positive-only, and dual-role designs under the same total budget of eight queries. As shown in Table 12, positive-only queries substantially improve over shared queries, confirming the importance of target-oriented evidence. Dividing the queries into referent-supporting and exclusion-oriented roles further improves all reported metrics. These results, together with the selectivity analysis below, support explicitly organizing queries by their referential functions.

For the role-selectivity analysis in the main paper, Δ^+ denotes the foreground–background contrast of the referent-supporting response, whereas Δ^- denotes the background–foreground contrast of the exclusion-oriented response. Their balanced selectivity is the mean of these two quantities.

C.5 Controlled Mechanism Comparisons

To distinguish the proposed mechanisms from simpler alternatives, we construct two matched controls. First, role-symmetric construction replaces the positive-only evidence summary used for semantic-depth reconciliation with $\frac{1}{2}(\bar{p}^+ + \bar{p}^-)$, while retaining the same parameter count, bounded dense correction, ECPD decoding, direct-text pathway, losses, and training protocol. Second,

Table 11: Effect of mismatched-expression construction on the RefSegRS test split. All variants use the complete AREM configuration.

Construction	mIoU	oIoU	P@0.5	P@0.7	P@0.9
Attribute counterfactual	74.88	81.79	85.47	73.69	25.26
Relation counterfactual	76.23	83.13	85.91	73.86	31.98
Attribute–relation balanced	75.80	82.98	85.09	73.53	**33.68**
Relation-first fallback	73.42	80.85	86.13	71.11	14.09
Mixed sampling (default)	**76.65**	**83.23**	**86.52**	**75.40**	29.22

Table 12: Comparison of query-role designs under an equal query budget on the RefSegRS test split.

Query design	mIoU	oIoU	P@0.5	P@0.7	P@0.9
Shared queries (8)	72.92	81.00	85.20	69.68	14.75
Positive-only queries (8)	75.62	82.72	85.80	73.42	27.19
Dual-role queries (4+4; AREM)	**76.65**	**83.23**	**86.52**	**75.40**	**29.22**

uniform candidate weighting replaces the role-conditioned weights ω_k with $1/4$, while preserving the candidate tokens, shared decoder, and bounded interpolation. All variants otherwise use the complete AREM configuration.

As shown in Table 13, replacing positive-anchored construction with symmetric role aggregation reduces mIoU by 6.56 points, supporting the asymmetric assignment of constructive authority in semantic-depth reconciliation while bounded correction is held fixed. Replacing role-conditioned candidate weighting with uniform aggregation reduces mIoU by 7.38 points, showing that retaining the role-aware weighting pathway during optimization is more effective than uniform multi-candidate aggregation.

C.6 Error-Correction Analysis

To examine whether AREM reduces prediction errors rather than merely shifting them across samples, we compare paired baseline and AREM outputs over all 1,817 RefSegRS test expressions using two automatic diagnostics. A severe failure is defined as IoU < 0.4. A remote false-positive activation is an 8-connected predicted component that does not intersect a 9×9 dilation of the ground-truth region and has area at least $\max(16, 0.1|Y|)$ pixels, excluding minor boundary deviations and isolated noise. No manual error annotation is used.

For each diagnostic, *Added* denotes errors introduced by AREM and *Removed* denotes baseline errors corrected by AREM; percentages and *Net reduction* are normalized by the number of baseline failures. As shown in Table 14, AREM removes 67 severe failures while introducing 22, yielding a net reduction of 45 cases (22.17%). For remote false-positive activations, it removes 101 and introduces 48, yielding a net reduction of 53 cases (23.77%). Exact McNemar tests confirm significant paired reductions for both severe failures ($p = 1.90\times10^{-6}$) and remote false-positive activations ($p = 1.69\times10^{-5}$). The severe-failure trend remains significant at IoU thresholds of 0.3 and 0.5. Overall, AREM more often removes than introduces severe segmentation failures and remote false-positive activations.

C.7 Scale-Stratified Analysis

To examine AREM across target scales, we define the foreground-area ratio as $\rho = |Y|/(HW)$. The 33rd and 67th percentiles of ρ on the RefSegRS training split are 0.533% and 3.609%, respectively; these fixed thresholds define small, medium, and large targets on the test split. We compare the baseline and AREM on the same 1,817 test samples and estimate 95% confidence intervals using 10,000 paired bootstrap resamples.

AREM improves mIoU, oIoU, P@0.7, and P@0.9 across all three scale groups. The largest gains occur for small targets, reaching 6.14 mIoU, 8.31 oIoU, and 20.57 points at P@0.9. Substantial improvements remain for medium and large targets, although their P@0.5 differences are not dis-

Table 13: Controlled mechanism comparisons on the RefSegRS test split.

Configuration	mIoU	oIoU	P@0.5	P@0.7	P@0.9
Role-symmetric construction	70.09	78.70	81.12	62.85	15.41
Uniform candidate weighting	69.27	79.37	79.69	62.08	14.80
AREM	**76.65**	**83.23**	**86.52**	**75.40**	**29.22**

Table 14: Automatic error-correction analysis on the RefSegRS test split. Percentages are relative to the corresponding number of baseline failures.

Diagnostic	Baseline failures	AREM failures	Added	Removed	Net reduction
Severe failure (IoU < 0.4)	203	158	22 (10.84%)	67 (33.00%)	45 (22.17%)
Remote false-positive activation	223	170	48 (21.52%)	101 (45.29%)	53 (23.77%)

tinguishable from zero under the bootstrap intervals. These results indicate that AREM particularly benefits small-instance identification while remaining effective across different target scales. The all-sample intervals also exclude zero for every reported metric, supporting the statistical reliability of the overall improvements.

C.8 Inference Efficiency

Table 16 reports cumulative efficiency from the baseline to the complete AREM, measured on NVIDIA RTX 4090 GPUs. Each model is warmed up for 50 iterations and then measured for 500 iterations in each of three repeats. CUDA-event timing excludes data decoding, disk I/O, host-to-device transfer, and metric computation.

D Additional Visual Analysis

D.1 Progressive Prompt Responses

Figure 5 visualizes the cumulative progression from the baseline to the complete AREM. RAEF begins to concentrate activation around expression-compatible regions, PAEG improves the spatial organization of the dense response, and the complete model generally produces a more localized prompt and mask.

D.2 Dense-Prompt Comparison

Figure 6 compares the dense prompts and segmentation masks produced by RS2-SAM2 and AREM across different target scales and geometries. Compared with the relatively diffuse RS2-SAM2 responses, AREM more clearly localizes the small vehicle and better preserves the elongated structures of the bikeway and paved road. The corresponding masks further show that this improved prompt organization translates into more accurate referent segmentation.

D.3 Limitations and Failure Modes

Figure 7 examines representative moderate-quality predictions rather than catastrophic failures. Three recurring limitations remain: incomplete coverage at thin-structure junctions, leakage to nearby same-category instances, and over-expansion to regions that satisfy similar relational cues. These observations suggest topology-aware continuity, instance-level arbitration, and finer spatial-relation grounding as promising directions for future improvement.

Table 15: Scale-stratified performance on the RefSegRS test split.

Scale	Method	N	mIoU	oIoU	P@0.5	P@0.7	P@0.9
All	Baseline	1817	72.02	80.24	84.42	68.19	12.38
	AREM		**76.65**	**83.23**	**86.52**	**75.40**	**29.22**
Small	Baseline	661	68.09	62.41	75.95	60.51	14.83
	AREM		**74.23**	**70.72**	**80.94**	**70.50**	**35.40**
Medium	Baseline	645	71.47	67.88	86.51	69.92	2.02
	AREM		**76.16**	**73.10**	**87.60**	**76.28**	**16.90**
Large	Baseline	511	77.79	82.25	**92.76**	75.93	22.31
	AREM		**80.39**	**84.78**	92.37	**80.63**	**36.79**

Table 16: Cumulative inference efficiency with batch size 1 and BF16 precision.

Method	Params (M)	GFLOPs	Mean (ms)	P50 (ms)	P95 (ms)	FPS	Peak Mem. (MiB)
Baseline	934.81	2120.22	161.86	161.72	168.37	6.18	3856.67
+ RAEF	936.85	2120.52	165.46	165.58	169.85	6.04	3865.84
+ PAEG	938.49	2122.03	174.90	174.44	180.19	5.72	3880.78
+ ECPD	940.99	2138.40	184.07	183.99	189.31	5.43	3879.47
Change	+0.66%	+0.86%	+13.72%	+13.77%	+12.44%	−12.06%	+0.59%

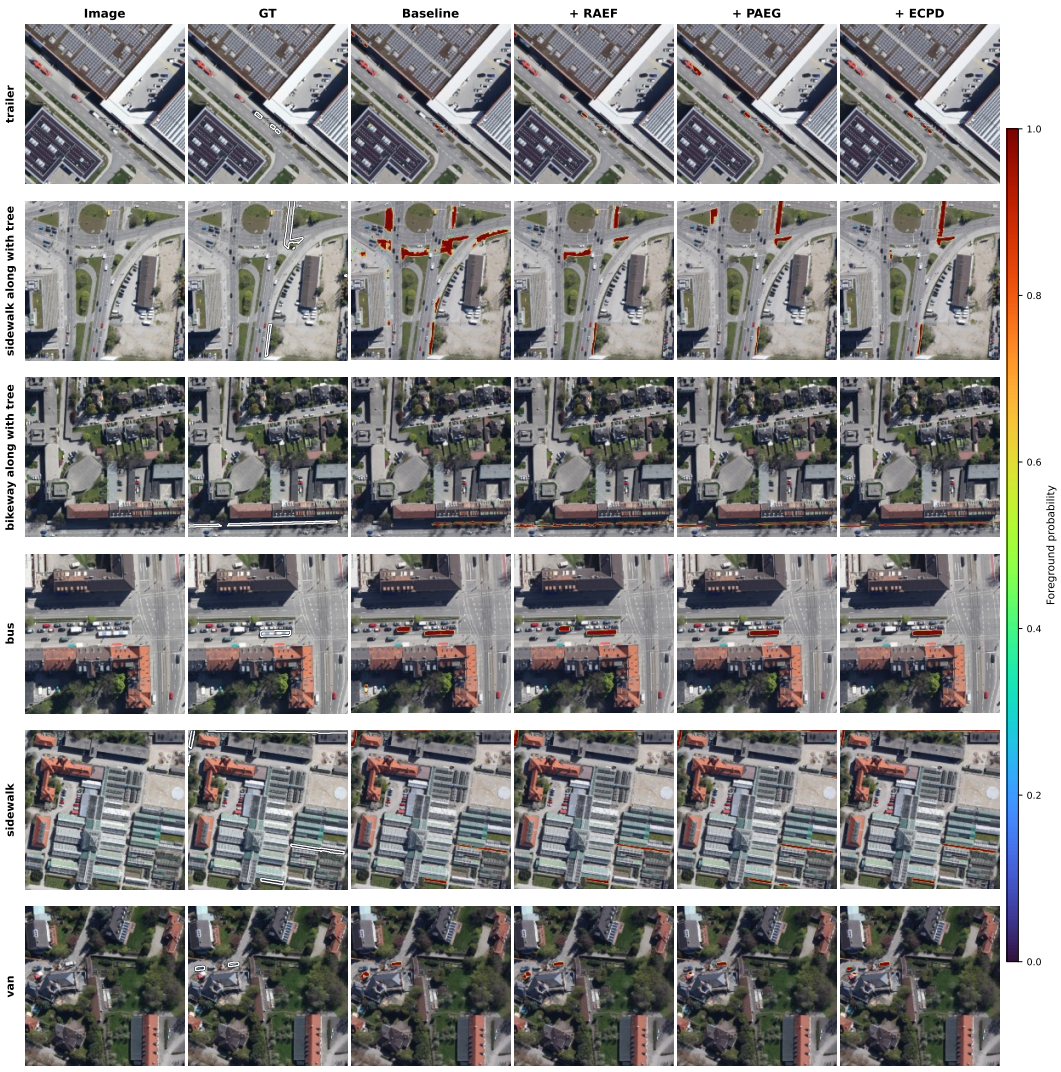


Figure 5: Progressive internal responses from the baseline to the complete AREM model on representative RefSegRS test samples. The columns compare the input image, ground truth, and the dense responses produced by four cumulative configurations: Baseline, +RAEF, +PAEG, and +ECPD.

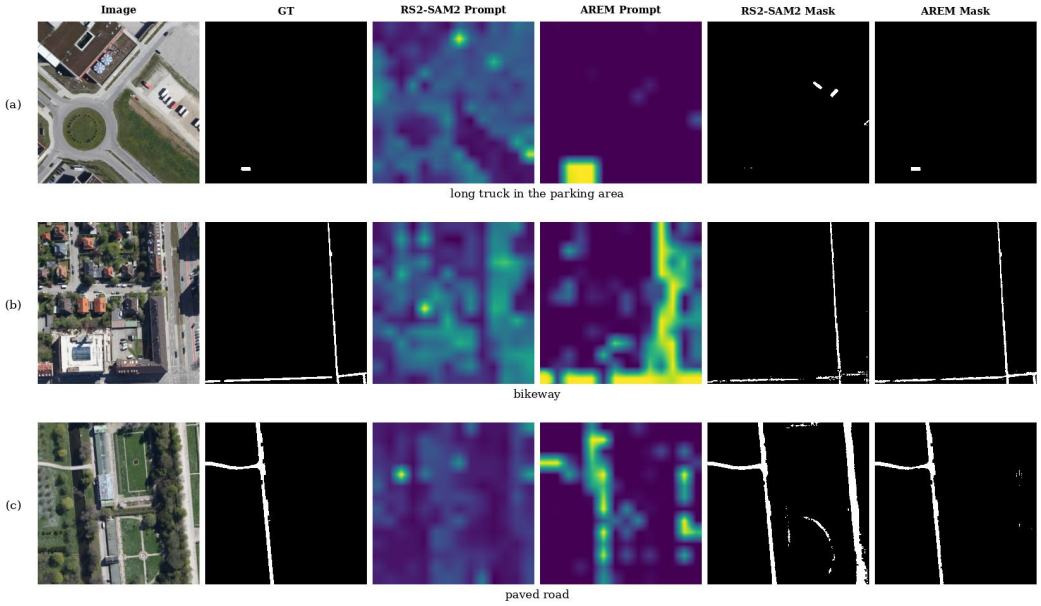


Figure 6: **Dense-prompt comparison between RS2-SAM2 and AREM on RefSegRS.** The rows show a small vehicle, a thin bikeway, and a large paved road. AREM produces more target-aligned prompts across different target scales and geometries. Prompt heatmaps are independently min–max normalized and therefore illustrate spatial response distributions rather than absolute magnitudes across models.

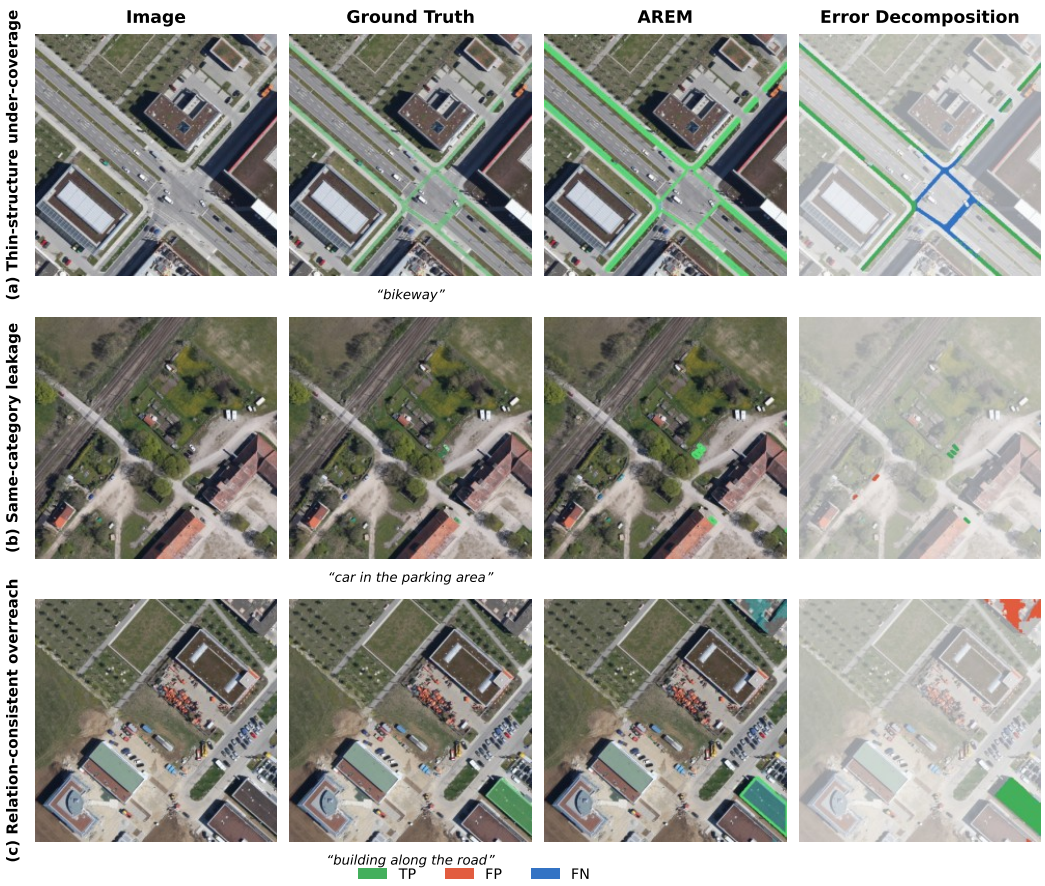


Figure 7: **Representative residual limitations of AREM.** The examples reveal remaining challenges in preserving thin-structure continuity, separating nearby same-category instances, and restricting predictions using fine-grained relational cues. Green, orange, and blue denote true-positive, false-positive, and false-negative regions, respectively.