DIA-RS^2 : DECOUPLED INFORMATION ALIGNMENT FOR REMOTE SENSING REASONING SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing remote sensing image (RSI) reasoning segmentation methods follow the "Embedding-as-Mask" paradigm, directly transforming high-level semantic representations from multimodal large language models into mask prompts. However, we observe that, in complex RSI scenes with high visual load, target-related semantic responses are prone to spatial displacement. We hypothesize that this phenomenon may be related to the structural coupling inherent in the existing paradigm: a single special token simultaneously undertakes target semantic representation, visual evidence grounding, and mask construction, leaving the mapping from language concepts to the visual space insufficiently and implicitly modeled. To examine this hypothesis, we conduct structural intervention experiments by modifying the token coupling scheme. The results show that explicitly modeling the mapping between language concepts and the visual space can significantly alleviate the degradation of vision-language correspondence as visual load increases, further supporting the association between the proposed structural bottleneck and semantic-visual misalignment. Inspired by these observations, we propose **D**ecoupled **I**nformation **A**lignment for **R**emote **S**ensing **R**easoning **S**egmentation (**DIA-RS**^{2}). The method first enhances visual semantic queryability to construct a visual evidence space that is more amenable to concept querying. It then introduces an independent <CON> token together with a Concept-to-Evidence Adapter (CEA) to explicitly establish the correspondence between linguistic concepts and task-relevant visual regions. Subsequently, Evidence-guided Fusion (EGF) injects the retrieved visual evidence into <SEG>, thereby enabling explicit alignment from linguistic concepts and visual evidence to mask prediction. Extensive experiments on multiple public benchmarks demonstrate that DIA-RS^2 achieves more stable and significant performance improvements over existing baseline methods.

1 INTRODUCTION

Transforming high-level semantic embeddings from language models into pixel-level masks has become the de facto mainstream paradigm in reasoning segmentation (Lai et al., 2024; Ren et al., 2024; Wei et al., 2024).We attribute this success to translating language into a semantic embedding directly interpretable by the segmentation model, where target categories, attributes, and reasoning outcomes are aggregated into a special token that serves as a mask prompt for pixel-level decoding-an "Embedding-as-Mask" paradigm. However, this paradigm also entails an often-overlooked structural bottleneck: while target semantics can be progressively constructed through language reasoning, the mapping from semantic concepts to visual evidence remains implicit throughout.In methods such as LISA (Lai et al., 2024), LISAt (Quenum et al., 2026), and FSVLM (Wu et al., 2025a), the <SEG> embedding effectively undertakes a continuous mapping process spanning target semantic representation, visual evidence localization, and mask construction. In other words, the mapping from target semantics to the visual space, together with its subsequent transformation into segmentation prompts, is compressed into the latent representation of a single token, without explicit constraints on the intermediate process.

Unlike natural images, task-relevant visual evidence in remote sensing image (RSI) is typically distributed in a denser and more spatially fragmented manner (Shabbir et al., 2025; Zhang et al., 2026),

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.

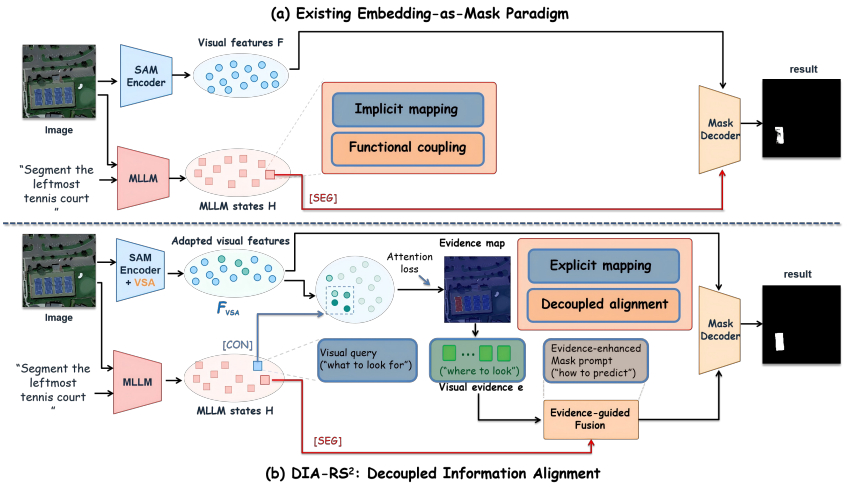


Figure 1: From Implicit Semantic Compression to Explicit Evidence Alignment.

whereas language reasoning provides only a sparse abstraction of target attributes and task intent, creating a more pronounced discrepancy in information density and representational form between the two modalities (Wu et al., 2025b). Consequently, when the correspondence between semantic concepts and visual evidence is still established primarily through the implicit mapping encoded within a single embedding, the model must localize target-relevant evidence under a higher visual discrimination load, placing greater demands on the stability of such spatial correspondence(Chen et al., 2025c; Liu et al., 2025). This leads to a natural hypothesis: as the visual information load increases, can language semantics in existing reasoning segmentation models still be reliably grounded to their corresponding local visual evidence, and is any resulting degradation in this correspondence related to the functional coupling inherent in the Embedding-as-Mask paradigm?

To validate this hypothesis, we conducted special token based decoupling intervention experiments on the Embedding-as-Mask paradigm. The results show that explicitly establishing the correspondence between linguistic concepts and the visual space through dedicated special tokens can mitigate the vision-language alignment shift induced by increasing visual load, while further disentangling the functional roles of the token leads to more stable spatial correspondence. In other words, the semantic-visual misalignment observed under high visual load is closely associated with the implicitly constrained concept-evidence mapping and functional coupling inherent in existing paradigms. Accordingly, we argue that the transition from language semantics to mask prediction should not be mediated solely by a single semantic embedding. Instead, visual evidence should be explicitly introduced as an intermediate state, such that high-level linguistic concepts are first grounded to target-relevant local visual regions, and the resulting visual evidence is then used to construct the subsequent mask condition. This process involves two sequential requirements: (1) the visual representation space should provide sufficient semantic addressability to support stable retrieval of target-relevant evidence; and (2) on this basis, the correspondence between linguistic concepts and visual evidence should be explicitly established, with the retrieved visual evidence further incorporated into mask-condition construction.

Building on these insights, we propose Decoupled Information Alignment for Remote Sensing Reasoning Segmentation (DIA-RS^2). As illustrated in Fig. 1, unlike the conventional Embedding-as-Mask paradigm, DIA-RS^2 reorganizes semantic representation, visual grounding, and mask construction, which are originally coupled within a single <SEG> token, into a sequential process of explicit concept-to-evidence mapping and evidence-enhanced mask construction. Specifically, the model first improves the semantic queryability of the visual representation space, then employs an independent <CON> token to explicitly retrieve target-relevant visual evidence and injects this evidence into <SEG> to form an evidence-enhanced mask prompt. In this way, DIA-RS^2 establishes a progressive alignment process from linguistic concepts to visual evidence and ultimately to pixel-level prediction. Main contributions as follows:

1. We revisit the Embedding-as-Mask paradigm and reveal that the functional coupling within a single <SEG> token leaves the vision-language mapping insufficiently constrained. Through experimental analysis, we further demonstrate its association with semantic-visual misalignment under high visual load.
2. We propose DIA-RS^2, which establishes a progressive alignment pathway from linguistic concepts to visual evidence and ultimately to mask prediction by enhancing the queryability of the visual space, introducing an independent <CON> token to acquire target-relevant evidence, and using this evidence to enhance <SEG>.
3. We systematically validate the effectiveness and generalization ability of DIA-RS^2. DIA-RS^2 achieves consistent improvements across multiple segmentation benchmarks and demonstrates strong generalization across both category and distribution shifts. In addition, ablation studies further provide a closed-loop validation of the effectiveness of explicit visual evidence modeling.

2 Related Work

2.1 Reasoning Segmentation and the Embedding-as-Mask Paradigm

Reasoning segmentation aims to identify target objects through language reasoning and generate the corresponding pixel-level masks. LISA(Lai et al., 2024) was the first to systematically formulate this task and introduced the <SEG> token, directly using semantic states from a multimodal large language model (MLLM) as segmentation prompts, thereby establishing the Embedding-as-Mask paradigm. Subsequent studies have primarily focused on strengthening language reasoning(Xia et al., 2024), improving visual representations(Zhang et al., 2024; Wei et al., 2025), and enhancing mask decoding accuracy. In the remote sensing images (RSIs), LISAt(Quenum et al., 2026) extends this paradigm to geospatial reasoning segmentation and improves complex-scene understanding through RSI data and multi-task training. SegEarth-R1(Li et al., 2025) further integrates hierarchical visual encoding, language understanding, and mask generation to enable language-guided segmentation of high-resolution RSI. SegEarth-R2(Xin et al., 2026) further addresses hierarchical granularity, multiple targets, and complex reasoning by introducing spatial attention supervision to improve target localization under complex language conditions. Despite these advances in reasoning and segmentation capabilities, these methods still largely follow the fundamental pathway of directly driving mask prediction from language semantics(Jia et al., 2026). In contrast, this work focuses on the intermediate correspondence between semantic concepts and local visual evidence within this process and explicitly models this correspondence.

2.2 Vision-Language Alignment for Remote Sensing Segmentation

Vision-language alignment represents another important research direction in language-guided RSI segmentation(Chen et al., 2025b; Chai et al., 2026). FIANet(Lei et al., 2025) investigates fine-grained image-text alignment for referring RSI segmentation, improving target localization by enhancing the interaction between local visual regions and textual semantics. RS2-SAM2(Rong et al., 2026) aligns RSI visual features with linguistic information through joint vision-text encoding and pseudo-mask prompt generation, and then uses the aligned representations to drive SAM2-based segmentation(Ravi et al., 2025). RSRefSeg2(Chen et al., 2025a) further decouples target localization from boundary segmentation, using CLIP(Radford et al., 2021) for cross-modal semantic localization and SAM(Kirillov et al., 2023) for fine-grained mask generation, thereby improving cross-modal localization and segmentation under complex semantic conditions. These studies demonstrate the importance of accurate vision-language correspondence for RSI segmentation(Ding et al., 2025). Existing methods have extensively explored cross-modal feature interaction, target localization, and segmentation prompt construction. However, how high-level semantics derived through reasoning can be further grounded into task-relevant local visual evidence remains insufficiently modeled.

3 Motivating Observation

To validate the hypothesis proposed in the Introduction, we conduct structural intervention experiments based on LISAt-7B, trained on EarthReason, with three configurations: G1 retains the original

coupled architecture with a single <SEG> token; G2 introduces explicit spatial mapping while retaining the single <SEG> token; and G3 further decouples the functions of <SEG> on top of G2. To isolate the effects of structural intervention, all three configurations use the same training budget and evaluation protocol. We compare their segmentation performance and spatial responses across the low, medium, and high load subsets of the EarthReason test set (see the Appendix A.2 for details).

As shown in Table 2(b), the segmentation performance of G1 declines markedly as visual information load increases, indicating pronounced performance degradation under complex visual conditions. With explicit spatial mapping, G2 achieves improvements across all load levels, with more pronounced gains under medium and high loads. The further decoupled G3 consistently yields additional improvements. These observations indicate that explicit mapping and functional decoupling can effectively alleviate segmentation performance degradation under high visual load, with more pronounced benefits in complex visual conditions. To further investigate the spatial correspondence mechanism underlying these performance changes, Fig. 2(a) presents visualizations of target-related spatial responses under different visual load levels, revealing trends consistent with the quantitative results. These findings are consistent with our hypothesis that underconstrained intermediate mapping and functional coupling may exacerbate spatial misalignment under high visual load. They further motivate us to explicitly introduce a visual-evidence alignment process between language concepts and mask prediction, establishing a progressive alignment pathway from language concepts to visual evidence and ultimately to mask prediction.

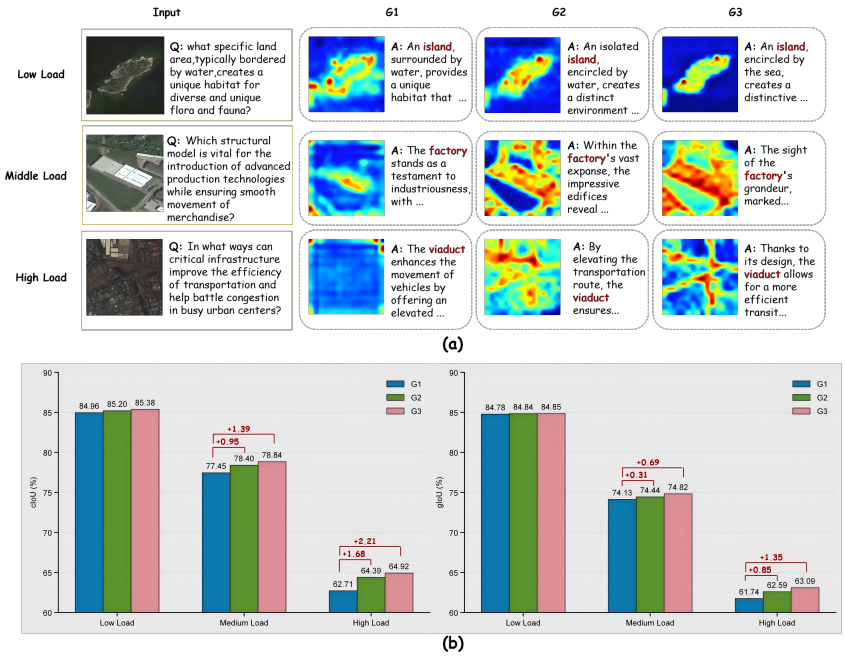


Figure 2: (a) Qualitative observations. The heatmaps visualize the activation responses corresponding to key objects in the answers (highlighted in red). (b) Quantitative observations.

4 METHOD

4.1 OVERVIEW

The proposed DIA-RS^2 is illustrated in Fig. 3. Specifically, given a RSI and its corresponding language instruction, the MLLM first performs joint vision-language understanding and target reasoning, and produces two special tokens with distinct functional roles. The <CON> token represents the target concept formed during reasoning and serves as a semantic query for subsequent visual evidence retrieval, while the <SEG> token preserves the high-level semantic state relevant to the segmentation task and is used to construct the final mask prompt. Meanwhile, the input image is pro-

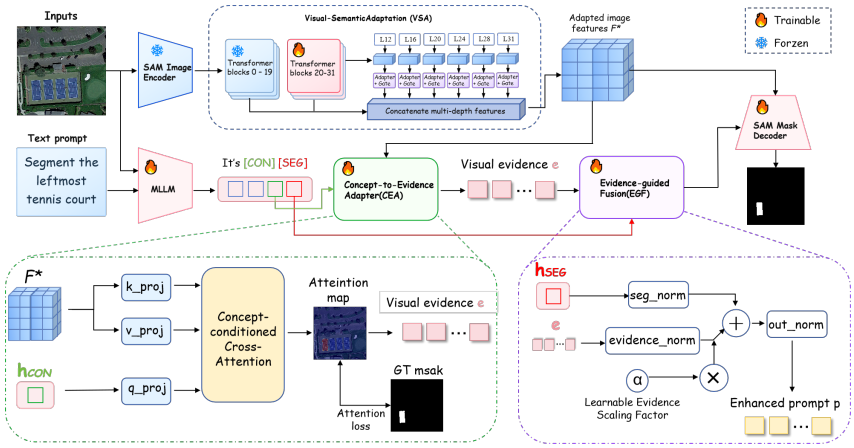


Figure 3: The overall architecture of the network.

cessed by the Visual Semantic Addressing (VSA) module to produce a dense visual representation space with enhanced semantic queryability. Building on this representation, we introduce Decoupled Information Alignment (DIA). Specifically, the Concept-to-Evidence Adapter (CEA) enables <CON> to query the dense visual space and explicitly retrieve spatial responses and visual evidence corresponding to the current target concept. The Evidence-Guided Fusion (EGF) module then injects the queried visual evidence into the <SEG> representation, allowing the final segmentation to jointly encode task semantics and target-relevant local spatial information. Finally, pixel-level prediction is performed by the SAM mask decoder.

4.2 Visual Semantic Addressing

Small-scale targets, fragmented regions, and highly similar background textures in RSI often cause task-relevant information to be distributed across different representation levels. If the visual representation itself lacks sufficient discriminability, it remains difficult to reliably localize the corresponding visual features even when the MLLM provides the correct linguistic concept through reasoning. To this end, we design Visual Semantic Addressing (VSA), which aims to construct visual representations that are more amenable to high-level semantic querying. Specifically, while keeping the main parameters of the SAM image encoder frozen, we introduce low-rank adaptation into the attention projections of its later Transformer layers to adjust the response of the generic visual representation space to remote sensing targets. For an original projection matrix W_l, its adapted form is expressed as:

$$W_l' = W_l + \frac{\gamma}{r} B_l A_l \tag{1}$$

where r denotes the rank of the low-rank adaptation, γ is the scaling factor, and A_l and B_l are learnable low-rank matrices. By optimizing only the low-rank adaptation branches, VSA preserves the generic visual priors of SAM while adapting the visual representations to the semantic distribution of RSI objects. Considering the complementary characteristics of features at different depths in capturing local structures and high-level semantics, we sparsely select multi-level dense features from the encoder to reduce redundancy between adjacent layers while retaining hierarchical information. Let the selected feature set be denoted as $S = \{l_1, l_2, \ldots, l_k\}$, where l_k indexes different encoder depths. For the feature $F^{(l)}$ from the l-th layer, we first project it into a unified feature space through an independent mapping function:

$$\tilde{F}^{(l)} = p_l\left(F^{(l)}\right), \quad l \in S \tag{2}$$

where $p_l(\cdot)$ denotes the feature projection operation associated with the corresponding layer, which is used to align the channel dimensions and representation spaces of features from different depths. On this basis, we jointly fuse the projected multi-level features with the final image embedding F_0 produced by the SAM neck:

$$F^* = F_0 + \Phi\left(\text{Concat}\left[F_0, \tilde{F}^{(l_1)}, \ldots, \tilde{F}^{(l_k)}\right]\right) \tag{3}$$

where $\Phi(\cdot)$ denotes the multi-level feature fusion mapping. Taking F_0 as the visual foundation, intermediate features from different depths are fused to supplement local semantic information, thereby avoiding excessive perturbation of the original spatial priors during cross-layer fusion.

4.3 Decoupled Information Alignment

After obtaining discriminative visual representations, two sequential issues remain to be addressed: how high-level linguistic concepts can localize task-relevant visual evidence, and how such evidence can further contribute to mask generation. To this end, we propose a DIA strategy, which consists of a Concept-to-Evidence Adapter (CEA) and an Evidence-Guided Fusion (EGF).

(1) Concept-to-Evidence Adapter. DIA-RS^2 introduces an independent <CON> token, using the target concept derived through language reasoning as an explicit query to retrieve task-relevant evidence from the visual representation space constructed by VSA. Specifically, the hidden state of <CON> is linearly projected into a concept query q, while the visual representations are projected into keys K and values V, respectively. The concept-conditioned spatial evidence distribution and its corresponding visual evidence representation are defined as:

$$A_e = \mathrm{Softmax}\left(\frac{qK^\top}{\sqrt{d}}\right), \quad e = A_e V \tag{4}$$

where A_e denotes the evidence distribution induced by <CON> over the visual space, and e represents the task-relevant visual evidence aggregated according to this distribution. In this way, visual evidence is explicitly modeled as an intermediate state between language semantics and mask prediction. To prevent this spatial correspondence from being learned solely through implicit supervision from the final segmentation loss, we further impose spatial supervision on the evidence distribution using the ground-truth mask:

$$\mathcal{L}_{\mathrm{attn}} = -\sum_i P_M(i) \log A_e(i) \tag{5}$$

Specifically, the ground-truth mask is resized to the resolution of the evidence map and normalized into a spatial probability distribution P_M. We then constrain the evidence distribution produced by <CON> using a cross-entropy loss.

(2) Evidence-guided Fusion. CEA explicitly establishes the correspondence between the target concept and its associated visual evidence. To further leverage this evidence for mask prediction, we design EGF, which injects the target-relevant visual evidence aggregated by CEA into the <SEG> representation. After projection and normalization, the two representations are fused in a residual manner controlled by a learnable coefficient, yielding an evidence-enhanced mask prompt:

$$p = \mathrm{LN}\left(\bar{S} + \alpha \bar{e}\right) \tag{6}$$

where $\bar{S}$ and $\bar{e}$ denote the normalized <SEG> representation and visual evidence, respectively, α is a learnable coefficient that controls the contribution of visual evidence, and p denotes the resulting evidence-enhanced mask prompt. This prompt is then fed, together with the dense visual features produced by VSA, into the original decoder to generate the prediction mask. DIA-RS^2 leaves the decoder architecture unchanged. Instead, it explicitly establishes the concept-to-evidence mapping before decoding and injects the queried visual evidence into <SEG>, so that visual localization and mask-prompt construction are no longer implicitly entangled within a single <SEG> representation.

4.4 Training Objectives

DIA-RS^2 is trained by jointly optimizing three objectives: language modeling, mask prediction, and concept-to-evidence alignment. The overall training objective is defined as:

$$\mathcal{L} = \mathcal{L}_{\mathrm{txt}} + \mathcal{L}_{\mathrm{mask}} + \lambda_{\mathrm{attn}} \mathcal{L}_{\mathrm{attn}} \tag{7}$$

where $\mathcal{L}_{\mathrm{txt}}$ denotes the autoregressive cross-entropy loss, which preserves the model's language understanding and reasoning capabilities. $\mathcal{L}_{\mathrm{mask}}$ is the mask prediction loss, consisting of binary cross-entropy and Dice losses, for supervising the final mask prediction. $\mathcal{L}_{\mathrm{attn}}$ denotes the concept-to-evidence alignment loss, which constrains the spatial evidence distribution produced by <CON>. And λ_{attn} is the corresponding weighting coefficient. By introducing intermediate evidence supervision, the model directly constrains the spatial correspondence between linguistic concepts and target-relevant visual regions, in addition to supervision from the final mask prediction.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

All baselines are trained on a mixed training set constructed from the training splits of EarthReason(Li et al., 2025) and RRSIS-D(Liu et al., 2024). DIA-RS^2 is initialized from LISAt-7B(Quenum et al., 2026), with RemoteCLIP ViT-L/14 adopted as the MLLM visual encoder and SAM ViT-H used for the pixel-level segmentation branch. We select FIANet(Lei et al., 2025), RS2-SAM2(Rong et al., 2026), SegEarth-R1(Li et al., 2025), SegEarth-R2(Xin et al., 2026), RSRefSeg2(Chen et al., 2025a), LISAt, and LISA(Lai et al., 2024) as the main comparison methods. All models are trained for 10 epochs with an effective batch size of 2 using the AdamW optimizer and an initial learning rate of 1×10^{-5}. Training is conducted with BF16 precision and DeepSpeed ZeRO-2 on four NVIDIA RTX 4090 GPUs. Following prior work, we adopt cIoU and gIoU as the evaluation metrics.

5.2 COMPARISON WITH STATE-OF-THE-ART METHODS

Table 1: Comparison with state-of-the-art methods on RRSIS-D and EarthReason.

	RRSIS-D		EarthReason	
Method	cIoU ↑	gIoU ↑	cIoU ↑	gIoU ↑
FIANet (TGRS'25)	76.24	57.77	77.11	73.90
RS2-SAM2 (AAAI'26)	65.31	51.89	73.42	70.84
SegEarth-R1 (arXiv'25)	66.45	53.63	65.78	70.38
SegEarth-R2 (CVPR'26)	**76.38**	62.46	75.88	73.87
RSRefSeg2 (TGRS'26)	60.33	49.68	67.77	69.11
LISAt (NeurIPS'25)	73.27	57.22	78.28	72.96
LISA (CVPR'24)	72.06	55.52	79.29	73.35
DIA-RS^2 (Ours)	76.20	**62.53**	**81.06**	**76.07**

As shown in Table 1, DIA-RS^2 achieves the best performance on EarthReason in terms of both cIoU and gIoU, outperforming existing methods by 1.77–15.28 and 2.17–6.96 percentage points, respectively. These results indicate that explicitly introducing visual evidence can improve the mapping from high-level semantics to local spatial targets and enhance the stability of semantic information transfer toward pixel-level prediction. On RRSIS-D, DIA-RS^2 achieves 62.53% gIoU, surpassing SegEarth-R2 at 62.46%, while obtaining 76.20% cIoU, only 0.18 percentage points below the best result. Although the cIoU does not reach the top performance, DIA-RS^2 still improves cIoU and gIoU over the direct LISAt baseline by 2.93 and 5.31 percentage points, respectively. This demonstrates that, under the same base model, introducing explicit concept-to-evidence mapping and evidence-enhanced mask construction yields consistent performance gains. The consistent improvements on both EarthReason and RRSIS-D further indicate that the proposed method remains effective under different forms of language conditioning.

5.3 ABLATION STUDIES

(1) Module Ablation. To analyze the contributions of VSA and DIA, we construct four model variants for ablation. As shown in Table 2, introducing either VSA or DIA alone consistently improves segmentation performance on both datasets, validating the effectiveness of enhanced visual representation and explicit concept-to-evidence alignment, respectively. When combined, the two components achieve the best overall performance, with gIoU on RRSIS-D increasing from 57.22% to 62.53%. These results indicate that VSA and DIA are not independent sources of performance gain, but rather complementary components: VSA improves the quality of the visual representation space, while DIA further establishes an explicit correspondence between linguistic concepts and visual evidence. We further provide ablation studies of the relevant mechanisms in the Appendix A.4.

(2) Mechanism Ablation. To further examine the key mechanisms within DIA, we ablate CEA spatial supervision and EGF separately. When CEA spatial supervision is disabled, we set λ_{attn} to 0 while retaining the forward evidence computation of <CON>. When EGF is disabled, the evidence

Table 2: Ablation study of the main components of DIA-RS^2.

Strategy			RRSIS-D		EarthReason	
Baseline	VSA	DIA	cIoU ↑	gIoU ↑	cIoU ↑	gIoU ↑
✓			73.27	57.22	78.28	72.96
✓	✓		75.49	60.86	79.67	75.34
✓		✓	74.33	58.38	79.03	73.26
✓	✓	✓	**76.20**	**62.53**	**81.06**	**76.07**

Table 3: Ablation study of the internal components of DIA.

Strategy		RRSIS-D		EarthReason	
CEA	EGF	cIoU ↑	gIoU ↑	cIoU ↑	gIoU ↑
✓		75.91	62.28	80.57	75.53
	✓	76.04	61.75	80.87	75.69
✓	✓	**76.20**	**62.53**	**81.06**	**76.07**

injection coefficient is fixed to 0, thereby blocking the transfer of visual evidence to <SEG>. As shown in Table 3, the full DIA achieves the best performance on both datasets. Removing EGF decreases cIoU and gIoU on EarthReason by 0.49 and 0.54 percentage points, respectively, while removing CEA spatial supervision reduces gIoU on RRSIS-D by 0.78 percentage points. These results demonstrate the complementary roles of explicit spatial supervision and evidence injection: the former constrains the spatial correspondence between <CON> and target visual regions, while the latter further incorporates the retrieved visual evidence into <SEG> to enhance the mask prompt.

5.4 Generalization and Adaptation under Distribution Changes

In this section, we evaluate the generalization of DIA-RS^2 beyond the training distribution from two perspectives: category shift and distribution shift. For category shift, we directly evaluate the model on LaSeRS(Xin et al., 2026) to assess its generalization to seen and unseen categories without any target-domain adaptation. For distribution shift, we evaluate the model on challenging scenarios in RefSegRS(Yuan et al., 2024) to investigate its ability to rapidly adapt to small targets, fragmented regions, and complex structures under limited target-domain supervision. **See the Appendix A.2 for detailed experimental protocols.**

(1) Category Generalization on LaSeRS. As shown in Table 4, DIA-RS^2 achieves the best performance on both seen and unseen categories. For seen categories, it achieves 48.38% cIoU and 48.52% gIoU, outperforming the strongest baseline by 12.84 and 9.00 percentage points, respectively. For unseen categories, DIA-RS^2 achieves 31.53% cIoU and 40.44% gIoU, with improvements of 2.85 and 10.23 percentage points, respectively. These results demonstrate that the performance gains of DIA-RS^2 are not limited to categories encountered during training, highlighting its strong cross-category generalization and providing further evidence for the transferability of the learned concept-to-evidence mapping.

Table 4: Category generalization on seen and unseen classes of LaSeRS.

	Seen		Unseen	
Method	cIoU ↑	gIoU ↑	cIoU ↑	gIoU ↑
FIANet (TGRS'25)	30.34	33.17	13.86	24.08
RS2-SAM2 (AAAI'26)	26.84	31.34	12.44	23.56
SegEarth-R1 (arXiv'25)	22.39	25.48	13.08	22.01
SegEarth-R2 (CVPR'26)	35.01	39.52	18.76	30.16
RSRefSeg2 (TGRS'26)	29.99	37.94	13.45	28.25
LISAt (NeurIPS'25)	34.37	34.76	28.68	30.12
LISA (CVPR'24)	35.54	38.59	15.39	30.21
DIA-RS^2 (Ours)	**48.38**	**48.52**	**31.53**	**40.44**

(2) Few-shot Adaptation on Challenging RefSegRS Scenes. As shown in Fig. 4, we select LISA, LISAt, and RS2-SAM2 as representative baselines. As the number of target-domain training samples increases, DIA-RS^2 exhibits consistent improvements in segmentation performance, with gIoU and cIoU increasing from 7.20% and 7.41% in the zero-shot setting to 25.89% and 41.84% in the 256-shot setting, respectively. Compared with the base model LISAt, DIA-RS^2 consistently outperforms it across all few-shot settings, achieving improvements of 2.81 and 3.38 percentage points in

gIoU and cIoU at 256 shots, respectively. Moreover, the shaded regions representing standard deviation gradually narrow as the number of training samples increases, indicating reduced performance variability across different random seeds and demonstrating more stable few-shot adaptation.

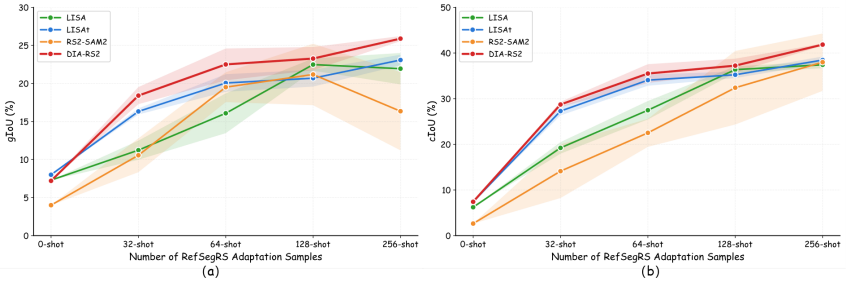


Figure 4: Few-shot adaptation performance. The shaded regions represent the standard deviation.

5.5 Qualitative Results

Figure 5 presents qualitative comparisons of different methods across representative challenging RSI scenarios, including agricultural fields, infrastructure, island regions, and narrow coastlines. Overall, although existing methods can generally identify target regions, they often suffer from spurious activations, missed target regions, and boundary over-expansion in complex backgrounds. In contrast, DIA-RS^2 produces segmentation results that more closely align with the ground truth. These observations suggest that explicitly establishing concept-to-evidence correspondence helps mitigate semantic response diffusion and spatial displacement in complex scenes, leading to more fine-grained segmentation results.

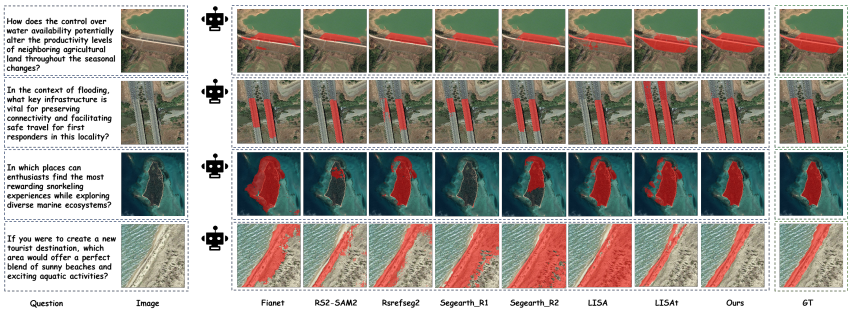


Figure 5: Qualitative comparison on representative challenging remote sensing scenes.

6 Conclusion

We revisit the Embedding-as-Mask paradigm in RSI reasoning segmentation and identify that the functional coupling within a single <SEG> leaves the mapping from linguistic concepts to visual evidence insufficiently constrained. Through experimental analysis, we find that explicitly establishing the correspondence between semantic concepts and the visual space mitigates semantic-visual misalignment under high visual load, while further decoupling concept localization from mask construction improves the stability of spatial correspondence. To this end, we propose DIA-RS^2, which employs VSA to construct a visual representation space more amenable to semantic querying. An independent <CON> is then introduced to explicitly retrieve target-relevant visual evidence, which is subsequently injected into <SEG> to form an evidence-enhanced mask prompt. This establishes a progressive alignment pathway from linguistic concepts to visual evidence and ultimately to mask prediction. Extensive experiments on EarthReason, RRSIS-D, LaSeRS, and RefSegRS validate the effectiveness of DIA-RS^2 and demonstrate its generalization capabilities across both category and distribution shifts.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language editing and improving the clarity and readability of the manuscript, as well as with code implementation and debugging during the development of the experimental pipeline. Generative AI was not used to generate experimental data, fabricate results, or make final scientific conclusions. All AI-assisted text was carefully reviewed and revised by the authors, and all AI-assisted code was manually inspected, tested, and validated against the intended functionality. All experimental results, numerical values, technical claims, and citations were independently verified by the authors. We take full responsibility for the final content of this work, including any text, code, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Jinming Chai, Lingling Li, Licheng Jiao, Xiaoqiang Lu, Long Sun, Xu Liu, Wenping Ma, and Weibin Li. Recs4r: Bridging semantics and geometry for referring remote sensing interpretation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 42213–42224, 2026.

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, 2025a.

Keyan Chen, Jiafan Zhang, Chenyang Liu, Zhengxia Zou, and Zhenwei Shi. Rsrefseg: Referring remote sensing image segmentation with foundation models. In *IGARSS 2025-2025 IEEE International Geoscience and Remote Sensing Symposium*, pp. 1070–1074. IEEE, 2025b.

Wenbo Chen, Zhen Xu, Ruotao Xu, Si Wu, and Hau-San Wong. Task-aware cross-modal feature refinement transformer with large language models for visual grounding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3931–3941. IEEE, 2025c.

Yueli Ding, Di Wang, Ke Li, Xiaohong Zhao, and Yifeng Wang. Visual grounding of remote sensing images with multi-dimensional semantic-guidance. *Pattern Recognition Letters*, 189:85–91, 2025.

Xiangyu Jia, Zhibo Chen, Shengyi Zhang, and Xiaojing Xue. Change-lisa: Language-guided reasoning for remote sensing change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 64:1–15, 2026. doi: 10.1109/TGRS.2026.3684817.

Siyoon Jin, seongchan kim, Jae Lee, Dahyun Chung, Hyunwook Choi, Jisu Nam, Jiyoung Kim, and Seungryong Kim. Matrix: Mask track alignment for interaction-aware video generation. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 119110–119160, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/c15da559b33c0f08a70573804cd5fb22-Paper-Conference.pdf.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

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

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. doi: 10.1109/TGRS.2024.3522293.

Kaiyu Li, Zepeng Xin, Li Pang, Chao Pang, Yupeng Deng, Jing Yao, Guisong Xia, Deyu Meng, Zhi Wang, and Xiangyong Cao. Segearth-r1: Geospatial pixel reasoning via large language model. *arXiv preprint arXiv:2504.09644*, 2025.

Biao Liu, Xu Liu, Lingling Li, Licheng Jiao, Fang Liu, Xinyu Sun, and Youlin Huang. Language-guided hybrid representation learning for visual grounding on remote sensing images. In *IJCAI*, pp. 1557–1566, 2025.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26648–26658. IEEE, 2024.

Jerome Quenum, Wen-Han Hsieh, Tsung-Han Patrick Wu, Ritwik Gupta, Trevor Darrell, and David Chan. Lisat: Language-instructed segmentation assistant for satellite imagery. *Advances in Neural Information Processing Systems*, 38, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

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

Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26364–26373. IEEE, 2024.

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*, volume 40, pp. 8751–8759, 2026.

Akashah Shabbir, Mohammed Zumri, Mohammed Bennamoun, Fahad Shahbaz Khan, and Salman Khan. Geopixel: Pixel grounding large multimodal model in remote sensing. In *Forty-second International Conference on Machine Learning*, 2025. URL `https://openreview.net/forum?id=nF8NxPUd0q`.

Cong Wei, Haoxian Tan, Yujie Zhong, Yujiu Yang, and Lin Ma. Lasagna: Language-based segmentation assistant for complex queries. *arXiv preprint arXiv:2404.08506*, 2024.

Cong Wei, Yujie Zhong, Haoxian Tan, Yingsen Zeng, Yong Liu, Hongfa Wang, and Yujiu Yang. Instructseg: Unifying instructed visual segmentation with multi-modal large language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20193–20203. IEEE, 2025.

Haiyang Wu, Zhuofei Du, Dandan Zhong, Yuze Wang, and Chao Tao. Fsvlm: A vision-language model for remote sensing farmland segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–13, 2025a.

Haiyang Wu, Weiliang Mu, Dandan Zhong, Zhuofei Du, Haifeng Li, and Chao Tao. Farmseg_vlm: A farmland remote sensing image segmentation method considering vision-language alignment. *ISPRS Journal of Photogrammetry and Remote Sensing*, 225:423–439, 2025b.

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

Zepeng Xin, Kaiyu Li, Luodi Chen, Wanchen Li, Xiao Yuchen, Hui Qiao, Weizhan Zhang, Deyu Meng, and Xiangyong Cao. Segearth-r2: Towards comprehensive language-guided segmentation for remote sensing images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13199–13210, 2026.

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.

Peirong Zhang, Yidan Zhang, Luxiao Xu, Jinliang Lin, Zonghao Guo, Fengxiang Wang, Xue Yang, Kaiwen Wei, and Lei Wang. Geovis: Geospatially rewarded visual search for remote sensing visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14335–14345, 2026.

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

A Appendix

A.1 Overview

This supplementary material provides additional details on the datasets, method implementation, experimental settings, and process visualizations presented in the main paper.

- Section A.2 describes the dataset usage and evaluation protocols.
- Section A.3 presents the implementation details.
- Section A.4 provides additional ablation studies and mechanism analyses, including low-rank adaptation, multi-level feature fusion, concept-based evidence localization, evidence injection, and fusion scaling.
- Section A.5 presents qualitative results and failure case analyses.

A.2 Datasets and Evaluation Protocols

We adopt four remote sensing language-guided segmentation datasets, namely EarthReason, RRSIS-D, LaSeRS, and RefSegRS, covering reasoning segmentation, referring segmentation, and comprehensive segmentation tasks involving both explicit and implicit language instructions. The dataset sizes, category counts, and task types are summarized in Table 5.

Table 5: Datasets used in this work and their basic statistics.

Dataset	Training	Validation	Test	Categories	Task Type
EarthReason	14,225	6,802	11,568	28	Reasoning Segmentation
RRSIS-D	12,181	1,740	3,481	20	Referring Segmentation
LaSeRS	18,099	–	1,079	122	Reasoning Segmentation
RefSegRS	2,172	431	1,817	14	Referring Segmentation

The specific usage and evaluation protocols for each dataset are detailed below:

(1) Main Experiments and Model Selection. All methods are trained on a unified training set constructed by combining the training splits of EarthReason and RRSIS-D, comprising 26,406 samples in total. EarthReason provides reasoning segmentation supervision through question-mask pairs, while RRSIS-D provides referring segmentation supervision through referring expression-mask pairs. All comparison methods use identical training data and training budgets, with the best checkpoint selected based on gIoU on the EarthReason validation set. The RRSIS-D validation set is excluded from model selection, and the test sets of both datasets are used exclusively for final performance evaluation.

(2) Cross-Category Generalization. LaSeRS is used exclusively for direct evaluation and is not involved in training, checkpoint selection, or hyperparameter tuning. Based on whether the target categories appear in the combined training categories of EarthReason and RRSIS-D, the 88 categories covered by the test set are divided into 50 seen and 38 unseen categories. All methods are evaluated using the models selected in the main experiments without any target-domain fine-tuning, and segmentation performance is reported separately for seen and unseen categories.

(3) Distribution-Shift Generalization and Few-Shot Adaptation. RefSegRS is used to evaluate direct generalization and few-shot adaptation under different scene distributions. The zero-shot setting directly employs the models selected in the main experiments, while the 32-shot, 64-shot, 128-shot, and 256-shot settings use the corresponding numbers of samples drawn exclusively from

the official training set for adaptation. Experiments are conducted with three random seeds (42, 43, and 44), using nested training subsets within each seed. All methods share identical training samples and training budgets. The official validation set is used only for experimental diagnostics, while the test set is reserved exclusively for final evaluation. We report the mean and standard deviation across the three random seeds.

(4) Ablation Studies and Mechanism Analysis. All component ablations, DIA-specific mechanism ablations, and concept-to-evidence spatial alignment analyses are conducted on EarthReason and RRSIS-D, following the same dataset splits and evaluation protocols as the main experiments.

(5) Visual Information Load Stratification for Motivation Analysis. To investigate the impact of visual information load on the correspondence between language concepts and local visual evidence, we follow prior studies that characterize image segmentation difficulty using target geometry and background visual complexity, and construct a geometry-texture-based visual load score. The underlying rationale is that smaller target scales, greater fragmentation, more dispersed spatial distributions, and more complex boundaries increase the difficulty of localizing and integrating target-related visual evidence. Meanwhile, dense background textures introduce additional distracting responses. Accordingly, we characterize visual information load using five features:

- **Target Scale:** Measured by the negative logarithm of the ratio of target area to image area. Smaller targets correspond to higher visual load.
- **Region Fragmentation:** Measured by the logarithm of the number of valid connected components plus one, reflecting the spatial fragmentation of the target.
- **Spatial Dispersion:** Measured by the average pairwise distance between connected-component centroids, normalized by the image diagonal, reflecting the spatial distribution of target regions.
- **Boundary Complexity:** Computed from the normalized relationship between the squared total outer-contour perimeter and target area, characterizing the irregularity of target shapes.
- **Background Edge Density:** Computed as the proportion of edge pixels within the background region using the Canny edge detector, characterizing background texture density. Edge density has also been employed as a quantitative measure of visual clutter in prior studies.

Specifically, we first binarize the ground-truth target mask, perform eight-connected component analysis, and remove small noisy regions using a fixed area threshold. The resulting five-dimensional feature vector is defined as

$$\mathbf{x}_i = [s_i, f_i, d_i, b_i, e_i] \tag{1}$$

where the first four features characterize target geometric load, while the last feature characterizes background texture load. For samples containing only one valid connected component, spatial dispersion is set to zero. To account for differences in feature scales, we apply robust normalization based on the median and median absolute deviation (MAD), followed by equal-weight aggregation to obtain the overall visual load score:

$$L_i = \frac{1}{5} \sum_{j=1}^{5} \frac{x_{ij} - \mathrm{median}(x_j)}{\mathrm{MAD}(x_j) + \epsilon} \tag{2}$$

where ϵ is a small constant for numerical stability. A higher score indicates a greater visual discrimination burden arising from target geometry and background texture. Finally, we independently calculate the 25th and 75th percentiles of the visual load scores on the EarthReason and RRSIS-D test sets, respectively. Based on these thresholds, samples are divided into three fixed subsets: low load ($L_i \leq Q_{25}$), medium load ($Q_{25} < L_i < Q_{75}$), and high load ($L_i \geq Q_{75}$), yielding six stratified test subsets in total. All visual load scores are computed exclusively from the original images and ground-truth masks, without using model predictions, segmentation performance, or attention responses. All models are evaluated on the same subsets to investigate the relationship between increasing visual load, concept-visual alignment, and final segmentation performance.

(6) Experimental Settings for the Motivating Observation: In the motivating observation, we construct three progressively modified structural configurations based on LISAt-7B:

- G1 (Original Coupled Architecture): Retains the original Embedding-as-Mask pathway, where a single <SEG> representation connects language semantics to mask prediction.
- G2 (Explicit Mapping with Functional Coupling): Retains a single <SEG> token and uses its representation as a query to perform cross-attention with the spatial features produced by the visual encoder. The resulting concept-related spatial response is explicitly supervised using the ground-truth mask. This response is used only to compute an auxiliary loss and does not participate in final mask construction, which remains driven by the original <SEG> prompt.
- G3 (Explicit Mapping with Functional Decoupling): Introduces an independent <CON> token for visual evidence retrieval, applies explicit cross-attention supervision, and injects the retrieved visual evidence into the segmentation prompt through evidence-guided fusion, thereby separating the functions of concept localization and mask construction.

A.3 Implementation Details

DIA-RS^2 is initialized from the pretrained LISAt-7B model, employing RemoteCLIP ViT-L/14 as the multimodal visual encoder, a LLaMA-7B language backbone, and SAM ViT-H for pixel-level segmentation. The multimodal branch processes input images at a resolution of 224×224 and projects the visual features into 256 visual tokens of dimension 4096. The segmentation branch resizes and pads images to 1024×1024, producing spatial features of size $256 \times 64 \times 64$. For Visual Semantic Addressing (VSA), low-rank adaptation is applied to the QKV and output projections of SAM encoder blocks 20–31 (zero-based indexing), while intermediate features from multiple depths are integrated through gated residual fusion. The Concept-to-Evidence Adapter (CEA) employs single-head cross-attention with a hidden dimension of 256, generating a 64×64 spatial evidence distribution. In Evidence-guided Fusion (EGF), the learnable evidence scaling coefficient α is initialized to 0.1, and the concept-evidence alignment loss weight λ_{attn} is set to 0.1. All experiments are conducted on four NVIDIA RTX 4090 GPUs using the AdamW optimizer, BF16 mixed precision, and a random seed of 42. The effective batch size is 2, with 100 warmup steps. All models are trained for 10 epochs. The complete forward computation and training procedure of DIA-RS^2 are summarized in Algorithm 1.

For spatial alignment supervision, the ground-truth mask is downsampled to the evidence resolution using nearest-neighbor interpolation and normalized over valid foreground locations. If the foreground disappears after downsampling, the corresponding alignment loss is skipped, while the mask losses are retained. During inference, the model directly generates the segmentation mask using the evidence-enhanced prompt and addressable visual features, without requiring ground-truth masks or additional supervision.

A.4 Additional Ablation Studies and Mechanism Analysis

This section provides supplementary mechanism analyses of DIA-RS^2, focusing on Visual Semantic Addressing (VSA), Concept-to-Evidence Adapter (CEA), and Evidence-guided Fusion (EGF). Specifically, we present component ablations of VSA, analyze concept-evidence spatial responses and their relationship with segmentation quality, and compare different evidence injection strategies to further clarify the role of each module and its impact on segmentation performance.

A.4.1 Further Analysis of VSA

To analyze the contributions of individual components in Visual Semantic Addressing (VSA), we compare four configurations under identical training and evaluation settings: without VSA (A0), SAM LoRA only (A1), multi-layer feature fusion only (A2), and their combination (A3).

As shown in Table 6, both SAM LoRA and multi-layer feature fusion independently improve segmentation performance, while their combination achieves the best results. Compared with A0, the complete model (A3) improves cIoU and gIoU by 2.93 and 5.31 percentage points on RRSIS-D, and by 2.78 and 3.11 percentage points on EarthReason, respectively. These two components provide complementary support for concept-visual matching and target-relevant evidence retrieval by enhancing the visual representations accessed by CEA through task-specific feature adaptation and cross-layer information integration.

Algorithm 1 Implementation and Training Procedure of DIA-RS^2

Require: Remote sensing image, language instruction, and ground-truth mask (training only).
Ensure: Predicted segmentation mask.

Stage 1: Multimodal Semantic Encoding

Preprocess the input image for RemoteCLIP and SAM.
Encode the image and language instruction using the multimodal language backbone.
Extract the hidden states associated with <CON> and <SEG>.

Stage 2: Visual Semantic Addressing (VSA)

Extract the SAM neck feature F_0 and intermediate features $F^{(l)}$, $l \in \mathcal{S}$.
Apply low-rank adaptation to the selected SAM attention projections.
Project the intermediate features into a unified feature space:

$$\widetilde{F}^{(l)} \leftarrow p_l(F^{(l)}), \quad l \in \mathcal{S}.$$

Construct the addressable visual representation through gated residual fusion:

$$F^* \leftarrow F_0 + \phi\left(\mathrm{Concat}\left[F_0, \widetilde{F}^{(l_1)}, \ldots, \widetilde{F}^{(l_k)}\right]\right).$$

Stage 3: Concept-to-Evidence Adapter (CEA)

Flatten F_a into 4096 spatial feature vectors.
Construct the concept query q from the <CON> representation.
Project the spatial features into keys K and values V.
Compute the concept-conditioned spatial evidence distribution:

$$A_e \leftarrow \mathrm{Softmax}\left(\frac{qK^\top}{\sqrt{d}}\right).$$

Aggregate target-relevant visual evidence:

$$e \leftarrow A_e V.$$

Stage 4: Evidence-guided Fusion (EGF)

Project and normalize the <SEG> representation to obtain $\bar{S}$.
Project and normalize the retrieved evidence to obtain $\bar{e}$.
Construct the evidence-aware mask prompt:

$$p \leftarrow \mathrm{LN}\left(\bar{S} + \alpha\bar{e}\right).$$

Stage 5: Mask Prediction and Optimization

Feed F_a and p into the SAM mask decoder to generate the predicted mask.
if training **then**
    Compute the autoregressive language modeling loss $\mathcal{L}_{\mathrm{txt}}$.
    Compute the mask loss $\mathcal{L}_{\mathrm{mask}}$ using BCE and Dice losses.
    Downsample the ground-truth mask to 64×64 using nearest-neighbor interpolation.
    if the downsampled foreground is non-empty **then**
        Normalize the foreground mask into the spatial target distribution P_M.
        Compute the concept-evidence alignment loss:

$$\mathcal{L}_{\mathrm{attn}} \leftarrow -\sum_i P_M(i) \log A_e(i).$$

    else
        Skip the alignment loss for the current query.
    end if
    Compute the overall training objective:

$$\mathcal{L} \leftarrow \mathcal{L}_{\mathrm{txt}} + \mathcal{L}_{\mathrm{mask}} + \lambda_{\mathrm{attn}}\mathcal{L}_{\mathrm{attn}}.$$

    Update the trainable parameters by backpropagation.
end if
return Predicted segmentation mask.

A.4.2 Supplementary Analysis of Concept-Evidence Alignment

To complement the concept-evidence alignment experiments in the main paper, we conduct additional spatial response comparisons and stratified analyses on the EarthReason and RRSIS-D test sets to examine the target evidence localization capability of the CEA and its relationship with segmentation quality.

Spatial Response Comparison. To examine whether the concept-evidence spatial responses generated by CEA can effectively localize target regions, we compare three settings based on the spatial attention distribution A_e produced by the <CON> concept query. The original response preserves the attention distribution generated by CEA; the uniform response distributes attention mass uniformly across valid spatial locations; and the randomly shuffled response preserves the attention values while randomly rearranging their spatial positions. Random shuffling is repeated with five fixed

Table 6: Ablation study of VSA components on RRSIS-D and EarthReason.

Configuration	RRSIS-D		EarthReason	
	cIoU ↑	gIoU ↑	cIoU ↑	gIoU ↑
A0	73.27	57.22	78.28	72.96
A1	75.63	61.19	80.45	75.53
A2	75.35	61.44	79.90	74.30
A3	**76.20**	**62.53**	**81.06**	**76.07**

seeds, and the results are averaged. All other model parameters, upstream features, and padding regions remain unchanged. Following prior studies that quantify spatial alignment through attention mass within target regions (Jin et al., 2026), we adopt *Target Attention Mass* to measure the concentration of concept-conditioned attention on the ground-truth target:

$$M = \sum_{u \in \Omega} A_e(u) G(u), \tag{3}$$

where Ω denotes the set of valid spatial locations, $A_e(u)$ represents the normalized attention weight at location u, and $G(u) \in \{0, 1\}$ denotes the ground-truth binary mask resized to the same spatial resolution. This metric measures the proportion of attention mass assigned to the target region, with higher values indicating stronger target concentration.

As shown in Fig. 6, on EarthReason, the original response achieves 69.20% target attention mass, substantially exceeding the uniform and randomly shuffled responses (both 12.91%). On RRSIS-D, the original response reaches 53.98%, compared with 4.67% for both control settings. These results indicate that the concept-conditioned spatial responses generated by CEA effectively concentrate on target-relevant visual regions.

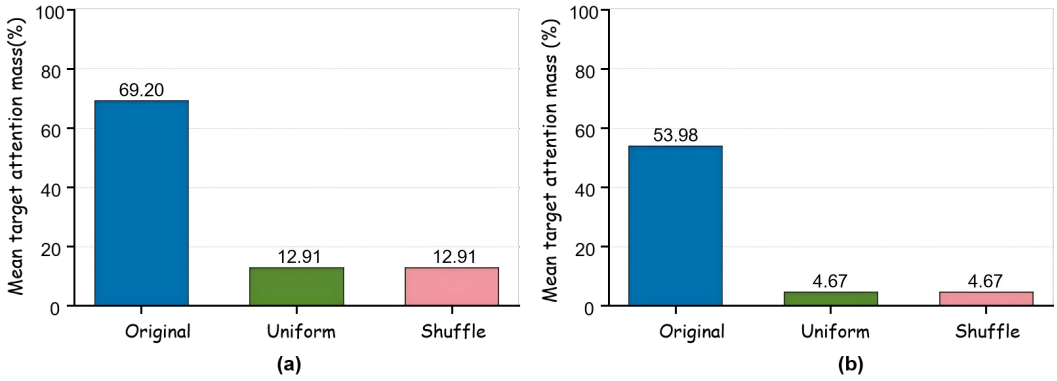


Figure 6: Target attention concentration under different CEA spatial response settings. (a) EarthReason and (b) RRSIS-D. Mean target attention mass is compared across original, uniform, and randomly shuffled responses.

Relationship Between Localization Quality and Segmentation Performance. To further investigate the relationship between concept–evidence localization quality and final segmentation performance, we adopt $E = M - p$ to quantify target evidence localization quality based on the original spatial attention distribution A_e generated by CEA. Here, M denotes the target attention mass defined above, and p represents the ratio of the ground-truth target area to the valid image area. By subtracting the target area baseline, this metric measures the excess attention concentration on the target region. A higher E indicates stronger localization of spatial responses within the ground-truth target. We then divide samples from each dataset into five approximately equal-sized groups (Q1-Q5) in ascending order of E, where Q1 and Q5 represent the lowest and highest localization-quality groups, respectively. The mean intersection-over-union (mIoU) is computed for each group. As

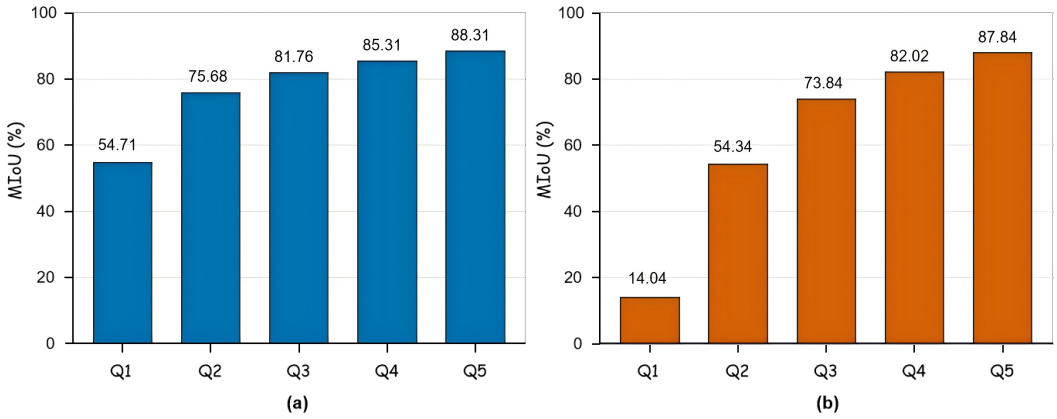


Figure 7: Segmentation performance across different concept-evidence localization quality levels. (a) EarthReason and (b) RRSIS-D. Samples are divided into five groups (Q1-Q5) according to localization quality, and the mean IoU (mIoU) is reported for each group.

shown in Fig. 7, mIoU increases progressively with localization quality on both EarthReason and RRSIS-D. Compared with Q1, Q5 achieves improvements of 33.60 and 73.81 percentage points, respectively. These results reveal a clear positive association between concept-evidence localization quality and final segmentation performance, indicating that samples with attention more concentrated on the ground-truth target generally achieve higher segmentation accuracy. This provides additional empirical evidence supporting the effectiveness of the explicit concept–evidence alignment mechanism presented in the main paper.

Visualization across Localization Quality Levels. To further illustrate the relationship between concept-evidence localization quality and segmentation performance, Figs. 8 presents representative samples from Q1 to Q5 on EarthReason and RRSIS-D, respectively. Each row displays the input image, the spatial attention response A_e generated by CEA, the ground-truth mask, and the predicted mask. Samples with lower localization quality tend to exhibit diffuse attention or insufficient target coverage, often accompanied by missed targets or incorrect region predictions. As localization quality increases, attention generally becomes more concentrated on target-relevant regions, accompanied by improved spatial agreement between predicted and ground-truth masks. These visualizations are consistent with the quantitative trends observed across the five groups, further illustrating the association between concept–evidence alignment quality and segmentation performance.

A.4.3 Further Analysis of EGF

To further investigate the effects of visual evidence injection and the fusion coefficient in EGF, we compare three configurations while keeping VSA, CEA, explicit spatial supervision, and the training and evaluation settings unchanged: a fixed coefficient of $\alpha = 0$ (A), a fixed coefficient of $\alpha = 1$ (B), and a learnable coefficient initialized to 0.1 (C). Specifically, A disables additive evidence injection, B directly incorporates the normalized visual evidence, and C adaptively adjusts the contribution of visual evidence through a learnable coefficient.As shown in Table 7, both B and C improve cIoU and gIoU on both datasets compared with A, indicating that explicit visual evidence injection contributes to segmentation performance. Among the fusion strategies, C achieves the highest cIoU and gIoU on RRSIS-D and the highest gIoU on EarthReason, although its cIoU is slightly lower than that of B. Overall, the learnable fusion coefficient adaptively regulates the contribution of visual evidence while preserving the original segmentation semantics, achieving higher gIoU on both datasets and further supporting the effectiveness of the evidence-guided fusion design.

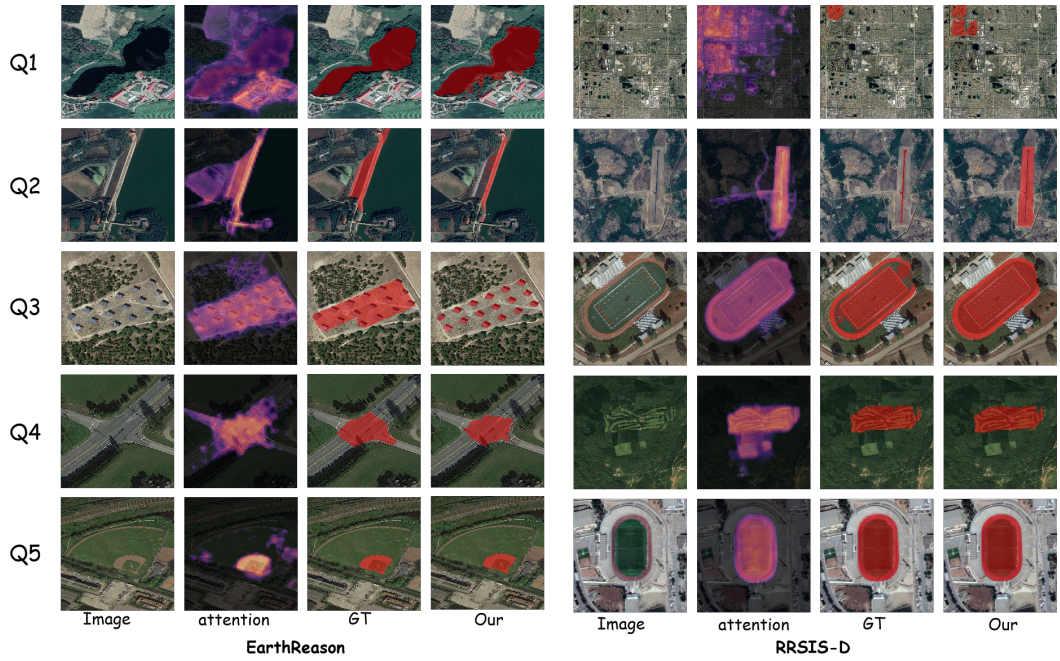


Figure 8: Visualization of concept-evidence spatial responses and segmentation results.

Table 7: Ablation study of different evidence fusion strategies in EGF. A: fixed $\alpha = 0$, B: fixed $\alpha = 1$, C: learnable α initialized to 0.1.

Configuration	RRSIS-D		EarthReason	
	cIoU ↑	gIoU ↑	cIoU ↑	gIoU ↑
A	75.91	62.28	80.57	75.53
B	76.04	62.43	**81.16**	75.69
C	**76.20**	**62.53**	81.06	**76.07**

A.5 Qualitative Results and Failure Case Analysis

A.5.1 Additional Qualitative Results

To further demonstrate the segmentation performance of DIA-RS^2 in complex remote sensing scenes, we provide additional qualitative results covering representative challenging scenarios, including small targets (Fig. 9), fragmented regions (Fig. 10), complex backgrounds (Fig. 11), and elongated structures (Fig. 12). DIA-RS^2 leverages target-relevant visual evidence to improve spatial localization and mask prediction in complex scenes, demonstrating its adaptability to varying target scales, spatial distributions, and structural characteristics.

A.6 Failure Case Analysis

To further analyze the limitations of DIA-RS^2 in complex remote sensing scenes, Fig. 13 presents four representative failure cases. Based on the reasoning outputs and predicted masks, we identify two potential sources of failure.

(1) Limited visual discriminability between targets and backgrounds. Even when the model correctly understands the target semantics, highly similar visual characteristics between the target and surrounding objects may hinder accurate and complete spatial correspondence. As shown in Fig. 13(a), the lawn and pond exhibit similar dark-green tones in the overhead imagery, with poorly distinguishable local boundaries, resulting in incomplete target coverage. In Fig. 13(b), the

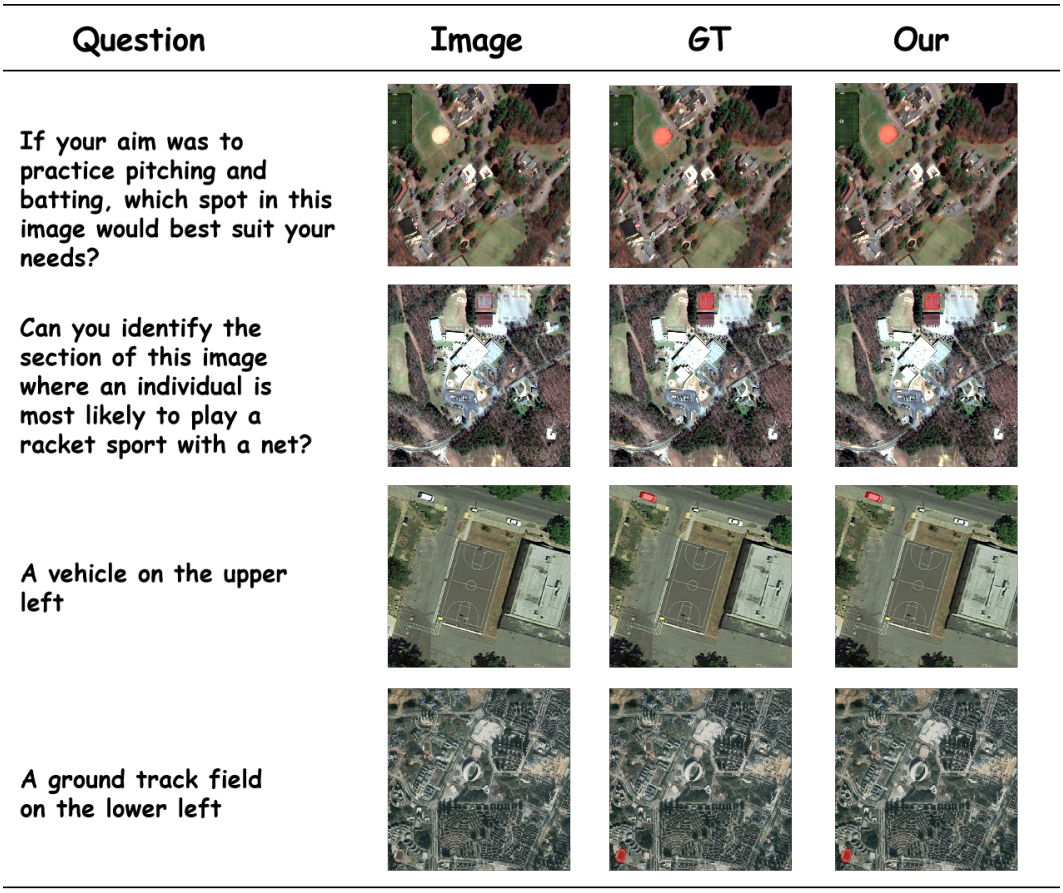


Figure 9: Qualitative segmentation results on small targets.

target railway station and surrounding buildings exhibit highly similar colors, textures, and structures, causing the predicted region to shift toward a building on the opposite side of the railway. These cases suggest that, when local visual features provide insufficient discriminative information, explicit concept–evidence alignment may still be affected by visually similar objects, leading to incomplete segmentation or spatial mislocalization.

(2) Errors in upstream target semantic reasoning. Reasoning segmentation relies not only on visual evidence localization but also on the ability of the multimodal large language model (MLLM) to infer the target category and spatial extent from implicit language instructions. In Fig. 13(c), the correct target is a stadium, whereas the model incorrectly infers a church, causing subsequent segmentation to be conditioned on an incorrect target concept and resulting in the localization of non-target regions. Fig. 13(d) illustrates a more subtle semantic discrepancy: the instruction refers to the core facility area of a substation, but the model infers the broader concept of a substation without specifying its precise spatial extent. Consequently, the predicted mask covers the entire substation site rather than the annotated core area. These cases reveal the impact of incorrect target-category inference and insufficient semantic specificity on final segmentation.

Overall, although DIA-RS^2 improves the spatial correspondence between language semantics and local visual evidence through explicit concept–evidence alignment, its performance remains constrained by visual feature discriminability and the accuracy of upstream semantic reasoning. Future work may incorporate fine-grained visual discriminative learning and enhanced target semantic reasoning to improve evidence discrimination in visually ambiguous scenes and the understanding of implicit target categories and precise spatial extents.

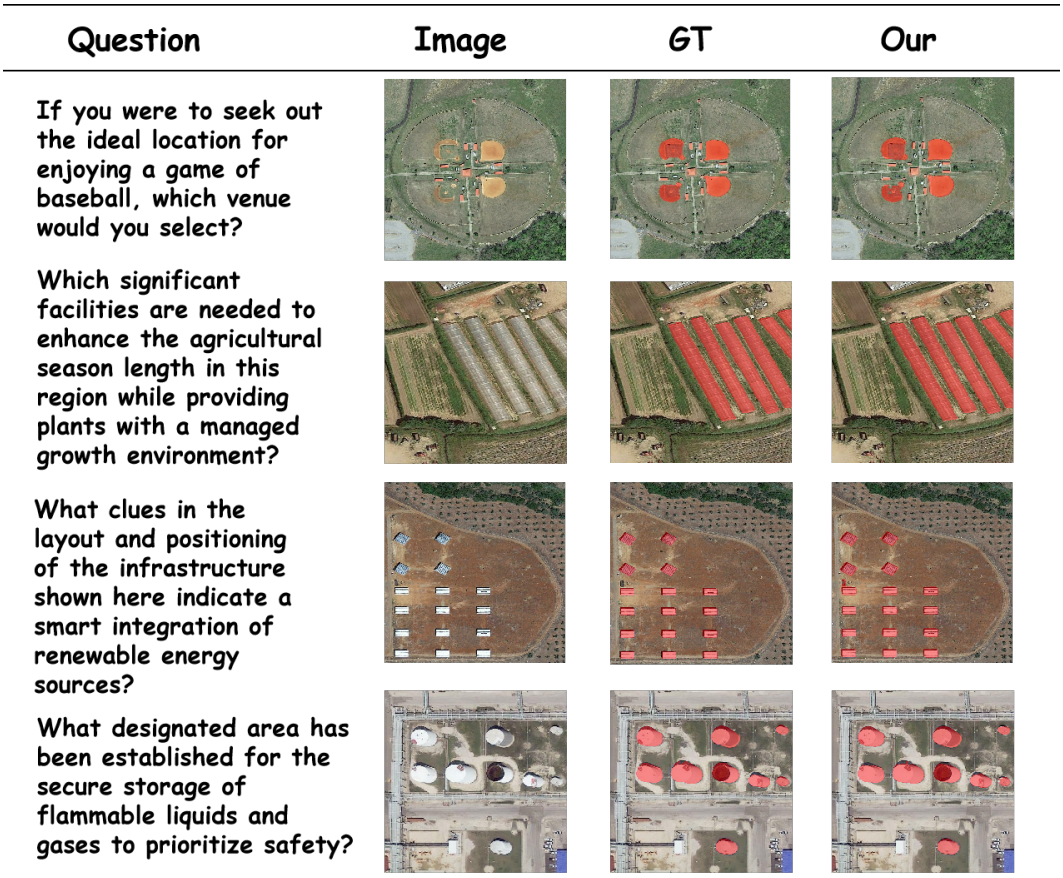


Figure 10: Qualitative segmentation results on fragmented targets.

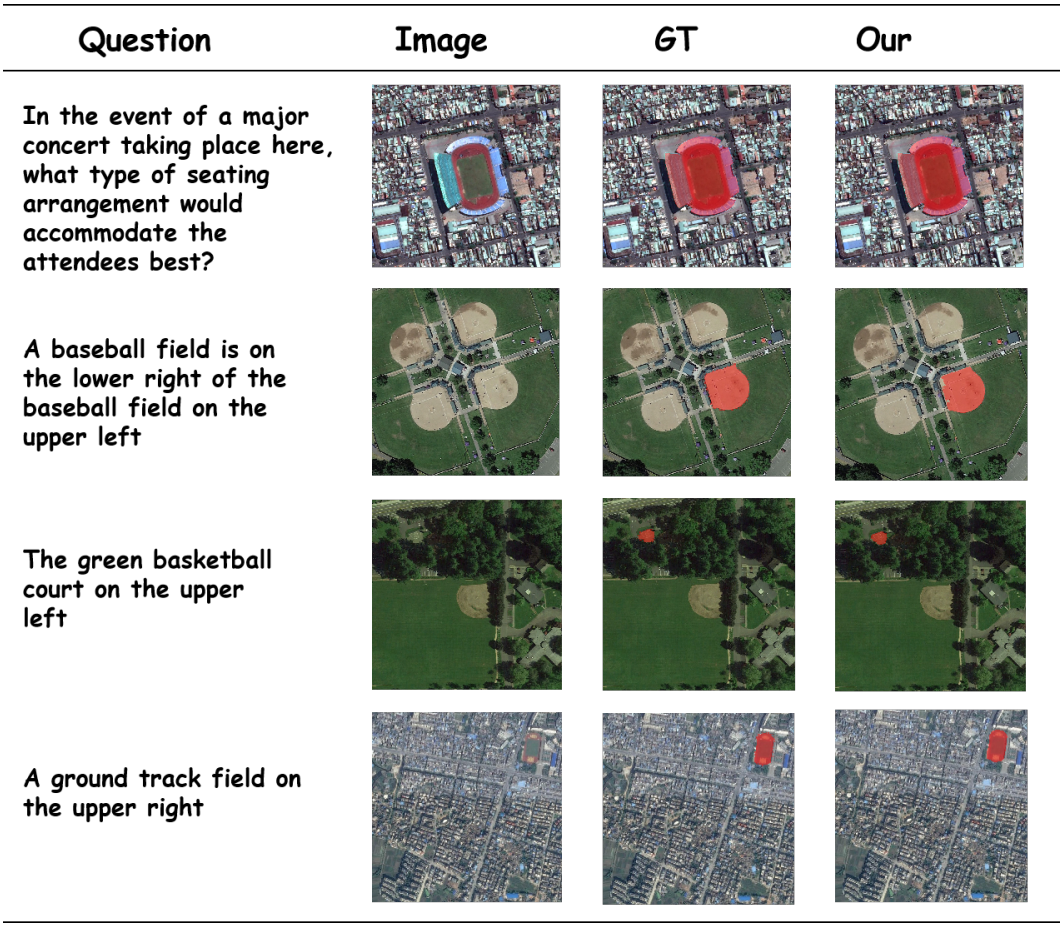


Figure 11: Qualitative segmentation results in complex backgrounds.

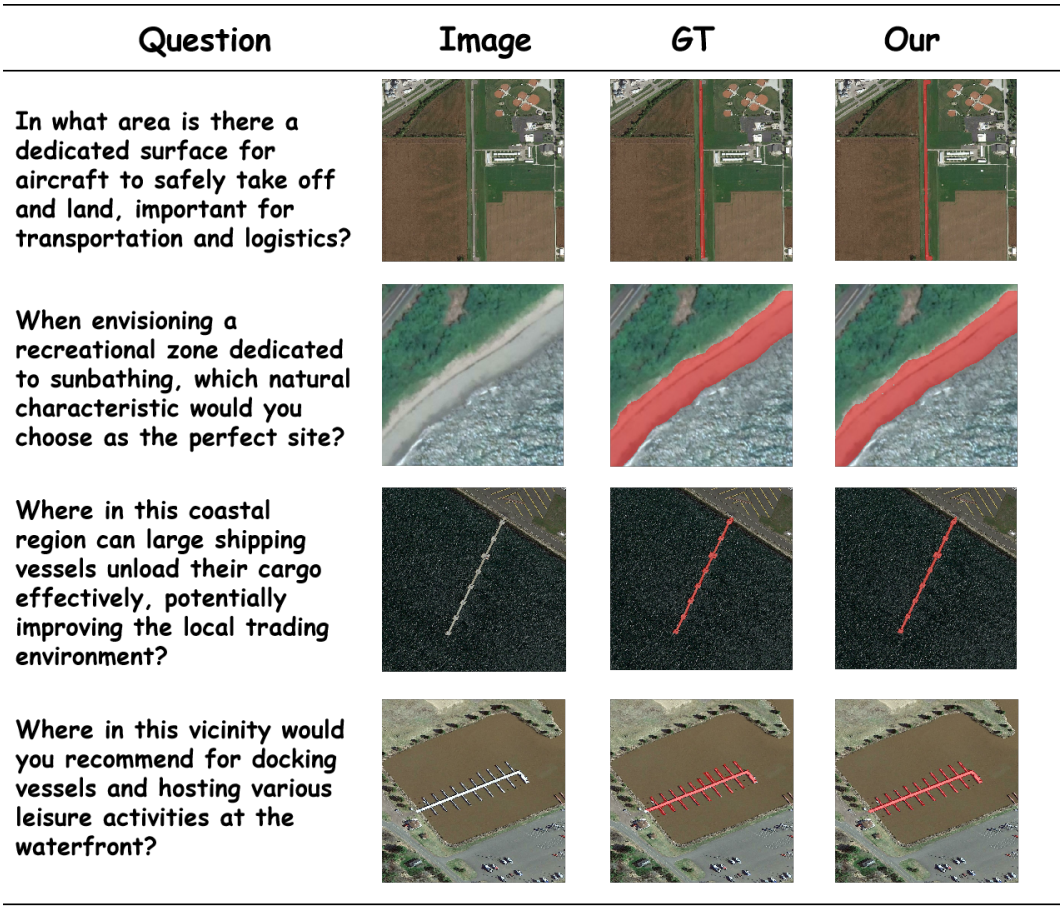


Figure 12: Qualitative segmentation results on elongated structures.

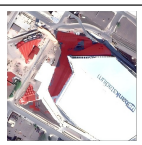
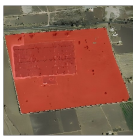
	Question	Image	GT	Our	Answer
(a)	As you prepare for a local picnic event, which picturesque natural attraction would you consider optimal for unwinding and socializing with others?				The lawn is the perfect location for picnic activities.
(b)	The gray slender train station				—
(c)	For someone tasked with planning a community-wide event, which location should be evaluated for its suitability in terms of capacity and ease of access?				The church serves as a vital hub f or community events, offering a s pacious and central location that i s both to the left of the parking l ot and the right of the roadway.
(d)	Which infrastructure category is most important for ensuring reliability in power distribution networks and mitigating blackout threats in metropolitan regions?				The substation plays a crucial role in the electrical grid by converting high-voltage electricity into lower voltages, which is essential for the efficient distribution of power and the prevention of power outages.

Figure 13: Failure Case.