DLSA: Dual-Level Semantic Alignment with Adaptive Multi-Scale Modeling for Referring Remote Sensing Image Segmentation

Wei Qu[1,2], Wei Wei[1*], Shihao Zou[1], Zian Cao[1]
[1] School of Computer Science and Technology, Huazhong University of Science and Technology
[2] AVIC Beijing Precision Machinery Institute
{weiqu_hust, weiw, sh_zou}@hust.edu.cn, ziancao_swjtu@126.com

Abstract

Referring Remote Sensing Image Segmentation (RRSIS) aims to segment the corresponding target region in a remote sensing image according to a natural-language expression. Compared with conventional referring image segmentation, RRSIS faces more significant scale variations, complex spatial layouts, and visual distractions, placing higher demands on vision–language alignment and multi-scale modeling. To address these challenges, we propose DLSA, a dual-level semantic alignment framework for RRSIS. DLSA introduces semantic alignment at two complementary levels, namely encoding and decoding. At the encoder level, we design Semantic-guided Adaptive Multi-scale Convolution (SgAMC), which explicitly introduces linguistic semantics into receptive-field selection, enabling language-guided adaptive scale modeling. At the decoder level, we introduce Adaptive Late-stage Text Injection (ALTI), which adaptively injects textual semantics into late decoding stages to refine mask reconstruction while preserving the stability of the decoding pathway. Experiments on RRSIS-D and RefSegRS demonstrate that DLSA achieves strong performance, validating the effectiveness of dual-level semantic alignment in remote sensing image scenarios. The source code is publicly available at https://github.com/young-qkl/DLSA.

1 Introduction

Referring Remote Sensing Image Segmentation (RRSIS) is an important task at the intersection of computer vision and natural language processing (Dong et al., 2024; Liu et al., 2024, 2025; Sun et al., 2026). Similar to conventional referring image segmentation (Hu et al., 2016; Liu et al., 2017; Zheng et al., 2022; Ji et al., 2022; Ma et al., 2023), RRSIS aims to segment a specific target region from a remote sensing image according to a given textual description. This task has important application value in remote sensing interpretation scenarios such as urban planning (Duan and Lafarge, 2016), land resource management (Foody, 2003), and forest protection (Rolnick et al., 2023). Therefore, exploring semantic understanding in remote sensing imagery is of great research significance.

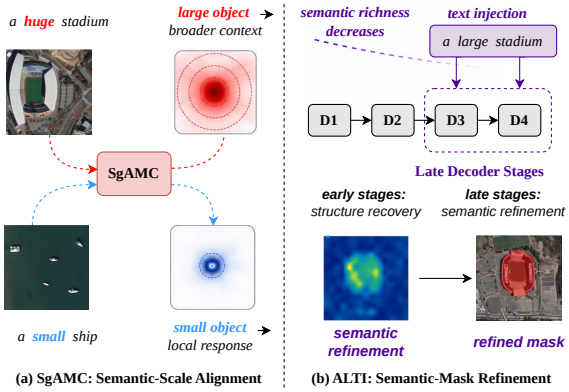


Figure 1: **Motivation of DLSA. (a)** SgAMC uses linguistic semantics to guide adaptive receptive-field selection for semantic-scale alignment. **(b)** ALTI re-injects textual guidance into late decoder stages to alleviate semantic degradation and refine mask reconstruction.

RRSIS is essentially a language-conditioned segmentation task. Its core challenge lies in how to effectively fuse visual representations with linguistic semantics, so that the model can accurately localize and segment the target region according to the textual description. Compared with natural images, remote sensing images usually cover larger geographic areas, where targets often exhibit significant scale variations, complex spatial layouts, and ambiguous boundary structures. Meanwhile, small objects, elongated objects, and visually similar instances are also more common in remote sensing scenes. These characteristics not only increase the difficulty of target segmentation itself, but also impose higher requirements on the model's

*Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 21687–21704
October 24-29, 2026 ©2026 Association for Computational Linguistics

multi-scale modeling capability.

Existing methods have explored this problem from perspectives such as cross-scale feature fusion, multi-level encoding interaction, and language-guided decoding, which have alleviated the influence of scale variations in remote sensing targets to some extent. However, most methods still mainly regard multi-scale modeling as a problem of feature-level fusion or visual representation enhancement, while paying less attention to the role that linguistic semantics can play in receptive-field selection. In other words, existing methods focus more on "how to fuse features of different scales", but rarely answer a more fundamental question: under a given textual description, how should the model select a receptive field that matches the target semantics?

In addition, beyond the insufficient semantic-scale alignment in the encoding stage, semantic preservation during decoding also deserves attention. Although semantic-guided receptive-field selection can enhance vision-language fusion during encoding, textual semantics may still gradually degrade during subsequent mask reconstruction. Recent studies have begun to focus on semantic preservation in the decoding stage, pointing out that introducing language information only during encoding or feature fusion may lead to insufficient semantic constraints and the loss of fine-grained details in mask reconstruction. Directly injecting language information into the decoder can provide additional semantic guidance, but fixed-strength or indiscriminate language injection may disturb spatial structure reconstruction, especially when early decoder features mainly carry low-level structural details.

To address the above issues, we propose DLSA, a Dual-Level Semantic Alignment framework for RRSIS. DLSA performs semantic alignment at two complementary levels. At the encoder level, DLSA introduces linguistic semantics into adaptive multi-scale feature modeling to achieve semantic-scale alignment. At the decoder level, DLSA adaptively re-injects textual guidance into late decoding stages to perform semantic-mask refinement. In this way, DLSA aligns the referring expression with both scale-aware visual representations and the final mask reconstruction process, thereby improving the localization and segmentation quality of referred targets in remote sensing images.

The main contributions of this paper are summarized as follows:

- We propose DLSA, a dual-level semantic alignment framework for RRSIS, which enhances language–vision alignment at both the encoder and decoder levels for scale-aware feature modeling and late-stage mask reconstruction.

- We design Semantic-guided Adaptive Multi-scale Convolution (SgAMC), which introduces linguistic semantics into receptive-field selection and learns semantics-aware routing over multi-scale convolutional branches.

- We introduce Adaptive Late-stage Text Injection (ALTI), which adaptively re-injects textual semantics into late decoding stages through a gated residual formulation to refine mask reconstruction while preserving decoding stability.

2 Related Work

2.1 Referring Remote Sensing Image Segmentation

Referring Remote Sensing Image Segmentation (RRSIS) originates from Referring Image Segmentation (RIS), and both tasks aim to predict pixel-level masks for target regions described by natural language expressions (Hu et al., 2016; Yang et al., 2022). Compared with natural images, remote sensing images usually exhibit more complex spatial layouts, more significant scale variations, and stronger inter-class similarity, making vision–language alignment and target localization more challenging.

Early RRSIS methods mainly transferred RIS techniques to remote sensing scenarios. LGCE (Yuan et al., 2024) constructed the first RRSIS benchmark dataset and proposed a language-guided cross-scale enhancement module for small and scattered targets. RMSIN (Liu et al., 2024) further constructed the large-scale RRSIS-D dataset and modeled scale and orientation variations through intra-scale interaction, cross-scale interaction, and adaptive rotated convolution. Subsequent task-specific methods, such as FIANet (Lei et al., 2025), CroBIM (Dong et al., 2024), and CADFormer (Liu et al., 2025), further improve RRSIS from the perspectives of fine-grained semantic parsing, bidirectional cross-modal interaction, correlation-aware fusion, and text-enhanced decoding.

Recently, foundation-model-based RRSIS methods have emerged as another strong technical route. For example, RS2-SAM2 (Rong et al., 2026) adapts SAM2 (Ravi et al., 2025) to RRSIS by generating dense prompts from multimodal representations. RSRefSeg 2 (Chen et al., 2026) further explores CLIP–SAM-based decoupled pipelines, using foundation models for semantic localization and prompt-guided mask refinement. These methods demonstrate the strong potential of large-scale pretrained segmentation priors for RRSIS.

Unlike foundation-model-based pipelines, this work focuses on improving semantic alignment within a task-specific encoder–decoder framework. Existing methods mainly focus on vision–language feature fusion, cross-scale interaction, or decoder-side enhancement, while the role of textual semantics in receptive-field selection during encoding remains underexplored. We investigate how linguistic semantics can regulate multi-scale convolutional responses, and introduce adaptive late-stage text injection to alleviate semantic degradation during mask reconstruction.

2.2 Multi-scale Modeling and Receptive-field Adaptation

Multi-scale modeling is a fundamental issue in remote sensing image understanding and is particularly important for RRSIS. Different referring expressions often imply different visual observation requirements: small targets rely more on local details, while large facilities and elongated structures usually require broader contextual receptive fields. Therefore, how to adaptively adjust visual receptive fields according to language descriptions is crucial for improving expression understanding and target localization in RRSIS.

Existing methods have explored multi-scale modeling from different perspectives. LGCE fuses shallow and deep features under language guidance to enhance small-target representation. RMSIN models scale and orientation variations through intra-scale interaction, cross-scale interaction, and adaptive rotated convolution. These methods demonstrate the importance of multi-scale feature interaction for RRSIS, while their scale adaptation processes are usually driven mainly by visual features or subsequent fusion mechanisms.

Dynamic receptive-field operators, such as deformable convolution (Dai et al., 2017; Zhu et al., 2019) and selective kernel (Li et al., 2019) mechanisms, also provide useful references for this problem. These methods can adjust sampling positions or fuse different convolutional branches according to visual appearance or feature statistics, but they usually do not fully exploit the semantic, scale, and structural cues contained in referring expressions. Different from the above vision-driven multi-scale modeling methods, the proposed SgAMC explicitly introduces linguistic semantics into adaptive multi-scale convolution. It uses the referring expression to regulate the response contribution of different convolutional branches, thereby achieving more targeted semantic–scale alignment.

3 Method

3.1 Overall Framework

Given an input remote sensing image $I \in R^{H\times W\times 3}$ and a textual expression E, where E is tokenized into a sequence of L tokens, the goal of RRSIS is to predict the binary mask $\hat{M} \in \{0,1\}^{H\times W}$ of the referred target. As shown in Fig. 2, we adopt Swin Transformer as the visual backbone $f_v(\cdot)$ and BERT as the language encoder $f_l(\cdot)$, which extract visual and linguistic features, respectively:

$$F_v = f_v(I), \quad F_l = f_l(E). \tag{1}$$

The proposed method follows an encoder–decoder architecture. In the encoding stage, SgAMC introduces textual semantics into adaptive multi-scale visual response modulation, enhancing language-conditioned scale-aware feature modeling for remote sensing targets. In the decoding stage, ALTI adaptively injects textual residuals into late decoding stages through a sample-wise gating mechanism, refining mask reconstruction while preserving structural stability. Finally, the decoder produces a text-conditioned probability map $P \in [0,1]^{H\times W}$, which is thresholded to obtain the predicted mask $\hat{M}$.

3.2 Semantic-guided Adaptive Multi-scale Convolution

At the i-th encoder stage, the current visual feature F_i^e is updated by two complementary residuals. For cross-modal feature alignment, we keep the original cross-modal alignment branch of RMSIN (Liu et al., 2024) unchanged, which produces a cross-modal residual R_{cm}. Our contribution in the encoder lies in the additional SgAMC branch, which generates a semantic–scale residual R_{vs} by introducing linguistic semantics into adaptive multi-

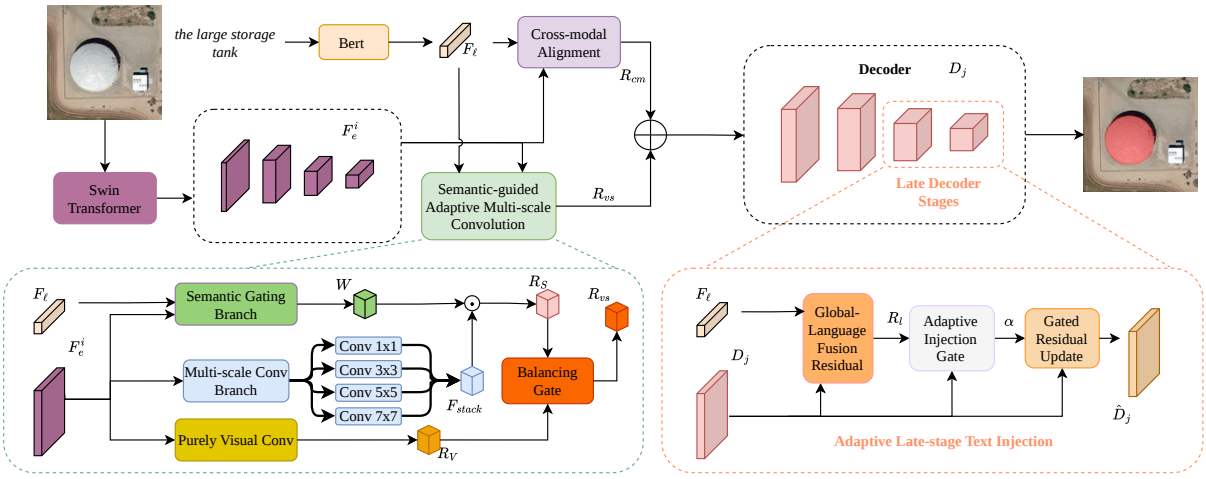


Figure 2: **Overview of DLSA.** Left: SgAMC performs semantic-scale alignment by guiding multi-scale convolution with language semantics. Right: ALTI performs semantic-mask refinement by adaptively injecting textual guidance into late decoding stages.

scale convolution. The stage-wise update is formulated as

$$F_{i+1}^{e} = \phi(F_i^e, R_{cm}, R_{vs}), \tag{2}$$

where $\phi(\cdot)$ denotes a lightweight residual fusion operation. Here, F_0^e is initialized by the backbone visual feature F_v, and higher-stage features are recursively obtained through the above process.

3.2.1 Multi-scale Convolutional Branch

Remote sensing images usually contain targets with significant scale variations. To provide candidate receptive fields for targets at different scales, SgAMC first aggregates the current visual feature along the channel dimension to obtain a channel-averaged spatial response map:

$$Y = \mathrm{Mean}_c(F_i^e) \in R^{H\times W\times 1}. \tag{3}$$

Then, multiple lightweight convolutional branches with different kernel sizes are applied to Y:

$$A_k = \mathrm{Conv}_{k\times k}(Y), \quad k \in \{1,3,5,7\}, \tag{4}$$

where $A_k \in R^{H\times W\times 1}$. The resulting multi-scale responses are concatenated along the channel dimension:

$$F_{\text{stack}} = [A_1; A_3; A_5; A_7] \in R^{H\times W\times 4}. \tag{5}$$

It should be noted that SgAMC does not directly apply costly multi-kernel convolution to the full-channel visual feature. Instead, it performs lightweight multi-scale modeling on the channel-averaged spatial response map, preserving spatial responses under different receptive fields while controlling computational cost.

3.2.2 Semantic Gating Branch

Relying only on visual responses for multi-scale fusion makes it difficult to fully exploit the scale, structure, and spatial relationship cues contained in the referring expression. Therefore, SgAMC introduces a semantic gating branch, allowing linguistic information to participate in the selection of different convolutional branches. Given the visual feature F_i^e and the textual feature sequence $F_\ell = \{F_\ell^t\}_{t=1}^{L}$, we adopt the `[CLS]` representation as the sentence-level textual feature:

$$\bar{F}_\ell = F_\ell^1. \tag{6}$$

Here, F_ℓ^t denotes the feature vector of the t-th token, and F_ℓ^1 corresponds to the `[CLS]` representation. We also compared `[CLS]` pooling with mean pooling over tokens and observed only minor differences. Therefore, for simplicity and consistency with the implementation, we adopt `[CLS]` pooling.

For each spatial position in F_i^e, its visual embedding is concatenated with the global sentence-level textual representation $\bar{F}_\ell$ to predict location-adaptive and branch-adaptive weights over the four convolutional branches $(1\times1, 3\times3, 5\times5, 7\times7)$:

$$\mathcal{W} = \mathrm{Softmax}\left(\mathrm{MLP}\left([\mathrm{Proj}(F_i^e), \mathrm{Proj}(\bar{F}_\ell)]\right)\right). \tag{7}$$

The predicted weights are reshaped back to the spatial grid to obtain the semantic gating map:

$$\mathcal{W} \in R^{H\times W\times 4}. \tag{8}$$

This gating map is used to modulate the multi-scale responses in the subsequent fusion process. Specifically, the stacked multi-scale response

$F_{\text{stack}} \in R^{H\times W\times 4}$ is weighted element-wise to obtain the fused multi-scale representation:

$$F_{\text{fused}} = \mathcal{W} \odot F_{\text{stack}}. \tag{9}$$

The fused response is further projected into a semantic gain map:

$$T = \sigma\left(\text{Conv}_{1\times 1}(F_{\text{fused}})\right) \in [0, 1]. \tag{10}$$

Here, the 1×1 projection maps the four branch channels to a single-channel gain map, $\odot$ denotes element-wise multiplication, and $\sigma(\cdot)$ denotes the sigmoid function. Finally, SgAMC applies residual scaling to the original visual feature:

$$R_s = F_i^e \odot (1 + T). \tag{11}$$

Through this process, SgAMC can adaptively select and fuse spatial responses under different receptive fields according to the target scale, structural attributes, and spatial relationships implied by the referring expression.

3.2.3 Purely Visual Convolutional Branch

Although the semantic gating branch enables textual semantics to participate in receptive-field selection, local texture, spatial structure, and visual saliency remain important for stable segmentation. Therefore, SgAMC also introduces a purely visual convolutional branch to provide text-independent spatial responses.

This branch also takes the channel-averaged response map Y as input and estimates a purely visual spatial weight through a lightweight visual response generator $\mathcal{V}(\cdot)$:

$$A^v = \mathcal{V}(Y). \tag{12}$$

Here, $\mathcal{V}(\cdot)$ denotes a visual response generator composed of lightweight convolutional operations. The visual weight is then used to recalibrate the original visual feature:

$$R_v = F_i^e \odot A^v. \tag{13}$$

Here, R_v denotes the purely visual scale response. This branch does not depend on textual semantics, but preserves language-independent information such as visual appearance, local texture, and spatial structure, providing stable visual evidence for subsequent dynamic fusion.

3.2.4 Dynamic Branch Balancing Gate

The language-guided response R_s and the purely visual response R_v provide complementary information. The former emphasizes receptive-field selection under textual semantic constraints, while the latter preserves language-independent visual spatial evidence. To adaptively balance their contributions, we introduce a dynamic branch balancing gate.

Given the global language representation $\bar{F}_\ell$, the fusion coefficient is predicted as

$$\lambda = \sigma(W_\lambda \bar{F}_\ell). \tag{14}$$

The final semantic–scale residual is computed as

$$R_{vs} = \lambda \cdot \sigma(R_v) + (1 - \lambda) \cdot \sigma(R_s). \tag{15}$$

Here, λ provides sample-level branch balancing between the purely visual and language-guided responses, while position-wise semantic routing weights handle region-level scale adaptivity. Thus, SgAMC combines global branch-level stability with local semantic–scale adaptation, and the resulting R_{vs} is fused with R_{cm} to update F_i^e.

3.3 Adaptive Late-stage Text Injection

Although SgAMC enhances semantic–scale modeling during encoding, textual cues may still weaken during progressive decoding. To alleviate this issue, we propose Adaptive Late-stage Text Injection (ALTI), which adaptively injects textual residuals into late decoding stages for semantic refinement while preserving structural stability.

We denote the progressive decoding process as

$$D_1 \to D_2 \to D_3 \to D_4,$$

where early stages D_1, D_2 mainly recover spatial structures and late stages D_3, D_4 are closer to the final mask prediction. In this work, ALTI is applied only to the last two stages, i.e., $\mathcal{S} = \{3, 4\}$. At each selected stage, ALTI is inserted after skip-feature fusion and the first Conv-BN-ReLU projection, and before the following decoder refinement layer. It acts as a residual plug-in and keeps the original decoder topology and feature resolution unchanged.

For the j-th decoding stage, let $D_j \in R^{H\times W\times d}$ denote the visual feature, where H, W, and d are the spatial size and channel dimension. The token-level textual feature from the language encoder is denoted as $F_l = \{f_t\}_{t=1}^{L}$, where L is the token length.

3.3.1 Global-language Fusion Residual

ALTI first aggregates token-level textual features into a global language representation. Given the textual feature sequence $F_l = \{f_t\}_{t=1}^{L}$ and the padding attention mask $m = \{m_t\}_{t=1}^{L}$, where $m_t = 1$ indicates a valid token and $m_t = 0$ indicates a padded token, we compute

$$\bar{F}_l = \frac{\sum_{t=1}^{L} m_t f_t}{\sum_{t=1}^{L} m_t + \epsilon}. \tag{16}$$

The mask is the standard padding mask from the tokenizer/BERT input pipeline, and no semantic-word mask, object mask, or ground-truth mask is used.

The global language representation and the decoding feature are first projected as $g_l = \mathrm{Proj}_l(\bar{F}_l)$ and $V_j = \mathrm{Proj}_v(D_j)$, respectively. The textual semantic residual is then generated by

$$R_l = \mathcal{F}_{out}(V_j \odot g_l). \tag{17}$$

In implementation, Proj_l is a linear projection with Tanh activation, Proj_v is a 1×1 Conv-BN-Tanh block, and $\mathcal{F}_{out}$ is a 1×1 Conv-BN-GELU block.

3.3.2 Sample-wise Adaptive Injection Gate

Different samples may rely on textual semantics to different degrees. Therefore, ALTI introduces a sample-wise injection coefficient α to control the overall strength of the textual residual. First, the current decoding feature is globally averaged to obtain a visual global representation:

$$\bar{D}_j = \mathrm{AvgPool}(D_j). \tag{18}$$

The global language representation is projected to the same dimension:

$$\tilde{F}_l = \mathrm{Proj}_\alpha(\bar{F}_l). \tag{19}$$

Then, the sample-wise injection coefficient is predicted from the visual global representation, the language global representation, and their interaction:

$$\alpha = \sigma\left(\mathcal{A}\left([\bar{D}_j, \tilde{F}_l, \bar{D}_j \odot \tilde{F}_l]\right)\right). \tag{20}$$

Here, $\alpha \in (0, 1)$ is a sample-wise scalar and broadcast to all spatial locations and channels. ALTI does not introduce additional uncertainty-aware statistics, but performs adaptive control based only on the global visual representation, the global language representation, and their interaction.

3.3.3 Late-stage Selective Injection

ALTI further predicts a local gate to control textual residuals across spatial locations and channels:

$$G_l = \sigma\left(\mathcal{G}([D_j, R_l])\right), \quad G_l \in R^{H\times W\times d}. \tag{21}$$

The operator $\mathcal{G}(\cdot)$ is implemented by a lightweight 1×1 convolution. For the selected late decoding stages $j \in \mathcal{S}$, the decoding feature is updated in a gated residual form:

$$D_j^{out} = D_j + \tanh(s)\cdot\alpha\cdot G_l \odot R_l, \quad j \in \mathcal{S}. \tag{22}$$

Here, s is a learnable residual scale, α controls the sample-wise injection strength, and G_l modulates local spatial-channel responses. For stages outside $\mathcal{S}$, ALTI is skipped and the original decoding feature is kept unchanged.

This selective injection design enables controlled semantic refinement during late-stage mask reconstruction while preserving the stability of the decoding pathway.

4 Experiments

4.1 Datasets and Evaluation Metrics

We evaluate our method on RRSIS-D (Liu et al., 2024) and RefSegRS (Yuan et al., 2024) using the official splits. Following prior RRSIS works, we report oIoU, mIoU, and Precision@X with $X \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$. More details on the datasets, experimental settings, and evaluation metrics are provided in the appendix.

4.2 Comparison with Representative Methods

Tables 1 and 2 compare our method with representative methods on RRSIS-D and RefSegRS, respectively. Overall, our method achieves strong performance on both datasets, particularly in terms of mIoU, demonstrating its effectiveness in improving the average quality of predicted masks.

On RRSIS-D, our method obtains the best mIoU on both validation and test splits, reaching 65.78% and 64.60%, respectively. It also achieves leading results on several precision metrics, including P@0.6 on both splits and P@0.7/P@0.8 on the test split. These results indicate that the proposed modules improve both overall segmentation quality and mask precision under stricter IoU thresholds.

On RefSegRS, our method achieves the best results on all reported metrics, including Pr@0.5–Pr@0.9, oIoU, and mIoU, on both validation and test splits. The mIoU reaches 76.69% and 63.19%,

Table 1: Quantitative comparison with representative methods on the RRSIS-D dataset. The best and second best results are highlighted in **boldface** and underline, respectively.

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
CMPC (Huang et al., 2020)	57.93	55.83	48.85	47.40	38.50	36.94	25.28	25.45	9.31	9.19	70.15	69.22	50.41	49.24
BRINet (Hu et al., 2020)	58.79	56.90	49.54	48.77	39.65	39.12	28.21	27.03	9.19	8.73	70.73	69.88	51.14	49.65
CMPC+ (Liu et al., 2022)	59.19	57.69	49.36	47.51	38.67	36.97	25.91	24.33	8.16	7.78	70.14	68.64	51.41	50.24
CRIS (Wang et al., 2022)	56.44	54.84	47.87	46.77	39.77	38.06	29.31	28.15	11.84	11.52	70.98	70.46	50.75	49.69
ETRIS (Xu et al., 2023)	62.10	61.07	53.73	50.99	43.12	40.94	30.79	29.30	12.90	11.43	72.75	71.06	55.21	54.21
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
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
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	61.70	60.25
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
CADFormer (Liu et al., 2025)	–	74.20	–	67.62	–	55.59	–	42.37	–	23.59	–	77.26	–	63.77
FIANet (Lei et al., 2025)	75.13	74.46	67.45	66.96	56.95	56.31	44.37	42.83	24.48	24.13	77.53	76.91	64.97	64.01
LSCF (Ma et al., 2025)	75.17	74.30	67.93	67.69	**57.99**	56.32	44.94	43.08	**25.98**	**25.67**	78.14	77.42	65.15	64.25
Ours	**75.75**	**74.81**	**69.33**	**68.03**	57.93	**56.88**	45.17	**43.78**	24.89	24.71	78.14	**77.83**	**65.78**	**64.60**

Table 2: Quantitative comparison with representative methods on the RefSegRS dataset. The best and second best results are highlighted in **boldface** and underline, respectively.

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
BRINet (Hu et al., 2020)	36.86	20.72	35.53	14.26	19.93	9.87	10.66	2.98	2.84	1.14	61.59	58.22	38.73	31.51
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 (Huang et al., 2020)	46.09	32.36	26.45	14.14	12.76	6.55	7.42	1.76	1.39	0.22	63.55	55.39	42.08	40.63
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
MAttNet (Yu et al., 2018)	48.56	28.79	40.26	22.51	20.59	11.32	12.98	3.62	2.02	0.79	66.84	64.28	41.73	33.42
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
BKINet (Ding et al., 2024)	52.04	36.12	35.31	20.62	18.35	15.22	12.78	6.26	1.23	1.33	75.37	63.37	56.12	40.41
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
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
CrossVLT (Cho et al., 2024)	67.52	41.94	43.85	25.43	25.99	15.19	14.62	3.71	1.87	1.76	76.12	69.73	55.27	42.81
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
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-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
Ours	**94.90**	**78.92**	**90.26**	**67.69**	**84.45**	**47.66**	**46.64**	**19.22**	**9.28**	**3.96**	**82.00**	**74.90**	**76.69**	**63.19**

Table 3: Ablation study on modules SgAMC and ALTI.

SgAMC	ALTI	P@0.5	P@0.7	P@0.9	mIoU
✗	✗	92.11	69.61	7.19	72.73
✓	✗	93.74	80.28	9.05	75.38
✗	✓	92.34	74.25	8.12	73.39
✓	✓	**94.90**	**84.45**	**9.28**	**76.69**

respectively, further confirming the generalization ability of the proposed method across different remote sensing benchmarks.

These results show that SgAMC and ALTI are complementary: SgAMC strengthens language-aware multi-scale encoding, while ALTI improves late-stage semantic refinement during mask reconstruction.

We conduct module-level ablation experiments to evaluate the effectiveness of the proposed SgAMC and ALTI. More detailed component-level analyses and design studies of both modules are provided in the appendix.

Effect of SgAMC and ALTI. As shown in Table 3, each module alone improves the baseline. The baseline model without SgAMC and ALTI obtains 72.73% mIoU. Introducing SgAMC boosts mIoU to 75.38% (+2.65) and P@0.7 from 69.61 to 80.28, validating the effectiveness of semantic-guided adaptive multi-scale encoding. Adding ALTI alone also improves mIoU to 73.39% and raises P@0.7 to 74.25, showing that late-stage text injection provides useful semantic refinement during mask reconstruction. When SgAMC and ALTI are combined, the model achieves the best performance across all metrics, reaching 76.69% mIoU, 94.90 P@0.5, 84.45 P@0.7, and 9.28 P@0.9. Compared with the baseline, the full model improves mIoU by 3.96 percentage points and P@0.7 by

Table 4: Scale-aware kernel routing statistics of SgAMC on target-region positions. Targets are divided into large and small groups according to mask area. We report the mean±std of branch weights over $\{1\times1,\ 3\times3,\ 5\times5,\ 7\times7\}$, together with the aggregated small-kernel ratio W_{1+3} and large-kernel ratio W_{5+7}.

Stage	Size	1×1	3×3	5×5	7×7	W_{1+3}	W_{5+7}	N
0	Large	0.177 ± 0.028	0.093 ± 0.017	0.434 ± 0.076	0.296 ± 0.039	0.270 ± 0.044	0.730 ± 0.044	580
0	Small	0.185 ± 0.031	0.094 ± 0.018	0.426 ± 0.074	0.295 ± 0.038	0.279 ± 0.044	0.721 ± 0.044	581
1	Large	0.229 ± 0.040	0.331 ± 0.037	0.231 ± 0.032	0.209 ± 0.030	0.560 ± 0.045	0.440 ± 0.045	580
1	Small	0.202 ± 0.047	0.338 ± 0.047	0.249 ± 0.043	0.211 ± 0.039	0.540 ± 0.051	0.460 ± 0.051	581
2	Large	0.024 ± 0.038	0.541 ± 0.223	0.241 ± 0.198	0.194 ± 0.160	0.565 ± 0.217	0.435 ± 0.217	580
2	Small	0.013 ± 0.028	0.624 ± 0.247	0.271 ± 0.218	0.091 ± 0.114	0.637 ± 0.243	0.363 ± 0.243	581
3	Large	0.261 ± 0.142	0.219 ± 0.114	0.185 ± 0.102	0.335 ± 0.150	0.480 ± 0.151	0.520 ± 0.151	580
3	Small	0.294 ± 0.172	0.210 ± 0.118	0.200 ± 0.143	0.296 ± 0.150	0.504 ± 0.188	0.496 ± 0.188	581

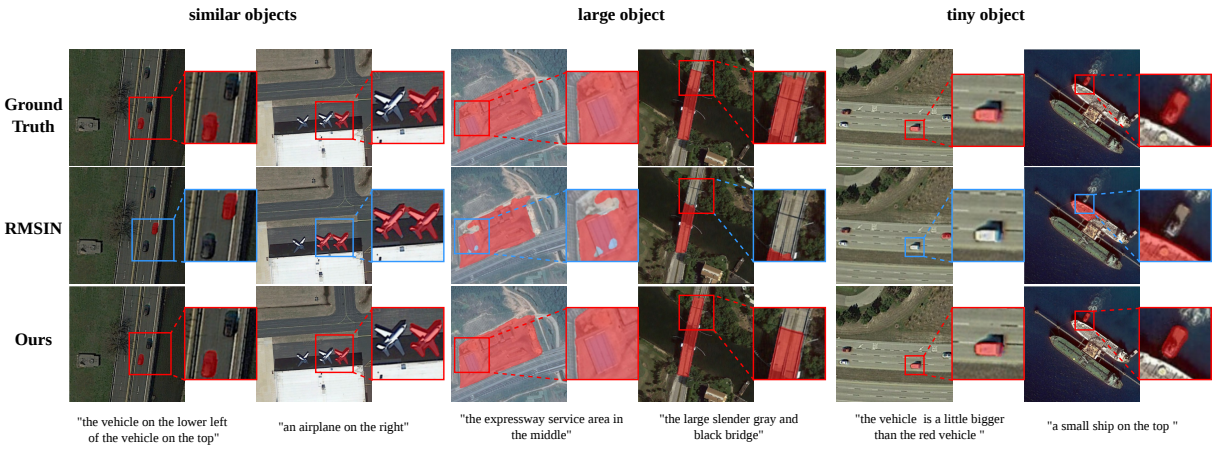


Figure 3: **Qualitative comparison on RRSIS.** From top to bottom: Ground Truth, RMSIN, and Ours.

14.84 points, demonstrating the complementary roles of SgAMC and ALTI in semantic-guided multi-scale feature modeling and adaptive late-stage mask refinement.

Scale-Aware Kernel Routing. To examine whether SgAMC learns scale-dependent receptive-field selection, we sort validation samples by mask area and use the top and bottom thirds as Large and Small groups. We report position-wise routing statistics in Table 4. The results show clear stage-dependent and scale-aware patterns. At Stage 0, both groups assign larger weights to 5×5 and 7×7, suggesting the need for broad contextual aggregation in early features. At deeper stages, the routing becomes more scale-sensitive: small targets rely more on the 3×3 branch at Stage 2, while large targets assign higher weights to the 7×7 branch at Stage 3. These results indicate that SgAMC learns non-uniform kernel routing rather than fixed multi-scale fusion, supporting the effectiveness of adaptive receptive-field selection.

Qualitative Visualization. Figure 3 compares DLSA with RMSIN on representative RRSIS-D cases, including similar objects, large targets, and tiny objects. DLSA produces more coherent masks, better aligns with the referring expressions, and reduces false positives and missing regions, visually supporting the quantitative improvements.

5 Conclusion

This paper proposes DLSA, a dual-level semantic alignment framework for Referring Remote Sensing Image Segmentation (RRSIS). To address scale variations and semantic degradation during decoding, DLSA performs semantic alignment at both encoding and decoding levels. Specifically, SgAMC introduces linguistic semantics into receptive-field selection for adaptive multi-scale modeling, while ALTI injects gated textual residuals into late decoding stages to refine mask reconstruction. Together, they align the referring expression with scale-aware encoding and late-stage reconstruction, producing more semantically consistent segmentation. Extensive experiments on benchmark datasets demonstrate the effectiveness of the proposed strategy.

6 Limitations

Although DLSA achieves effective dual-level semantic alignment for referring remote sensing image segmentation, there is still room for further improvement. First, SgAMC adopts a compact multi-branch convolutional design to model different receptive fields. This design provides an efficient way to perform semantic-guided multi-scale modeling, but more diverse receptive-field patterns, such as deformable, shape-adaptive, or orientation-aware receptive fields, may further enhance the representation of irregular and elongated remote sensing targets.

Second, remote sensing images often contain complex spatial layouts, dense instances, and large intra-class variations. Although the proposed method improves semantic-scale alignment and mask refinement, challenging cases with highly similar objects or fine-grained boundaries may still require stronger structural modeling. Future work will explore more flexible receptive-field modeling and structure-aware refinement strategies to further improve the robustness of referring segmentation in complex remote sensing scenes. Potential risks may arise when fine-grained remote sensing segmentation is applied to sensitive surveillance or location-analysis scenarios.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62276110 and in part by the fund of Joint Laboratory of HUST and Pingan Property Casualty Research (HPL). The authors would also like to thank the anonymous reviewers for their comments on improving the quality of this paper.

References

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

Yubin Cho, Hyunwoo Yu, and Suk-Ju Kang. 2024. Cross-aware early fusion with stage-divided vision and language transformer encoders for referring image segmentation. *IEEE Trans. Multim.*, 26:5823–5833.

Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. 2017. Deformable convolutional networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 764–773.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Haixin Ding, Shengchuan Zhang, Qiong Wu, Songlin Yu, Jie Hu, Liujuan Cao, and Rongrong Ji. 2024. Bilateral knowledge interaction network for referring image segmentation. *IEEE Trans. Multim.*, 26:2966–2977.

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

Liuyun Duan and Florent Lafarge. 2016. Towards large-scale city reconstruction from satellites. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*, volume 9909 of *Lecture Notes in Computer Science*, pages 89–104. Springer.

G. M. Foody. 2003. Remote sensing of tropical forest environments: towards the monitoring of environmental resources for sustainable development. *International Journal of Remote Sensing*, 24(20):4035–4046.

Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. 2016. Segmentation from natural language expressions. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, volume 9905 of *Lecture Notes in Computer Science*, pages 108–124. Springer.

Yutao Hu, Qixiong Wang, Wenqi Shao, Enze Xie, Zhenguo Li, Jungong Han, and Ping Luo. 2023. 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*, pages 4044–4054. IEEE.

Zhiwei Hu, Guang Feng, Jiayu Sun, Lihe Zhang, and Huchuan Lu. 2020. Bi-directional relationship inferring network for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shaofei Huang, Tianrui Hui, Si Liu, Guanbin Li, Yunchao Wei, Jizhong Han, Luoqi Liu, and Bo Li. 2020.

Referring image segmentation via cross-modal progressive comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tianrui Hui, Si Liu, Shaofei Huang, Guanbin Li, Sansi Yu, Faxi Zhang, and Jizhong Han. 2020. Linguistic structure guided context modeling for referring image segmentation. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part X*, Lecture Notes in Computer Science, pages 59–75. Springer.

Jiayi Ji, Yiwei Ma, Xiaoshuai Sun, Yiyi Zhou, Yongjian Wu, and Rongrong Ji. 2022. Knowing what to learn: A metric-oriented focal mechanism for image captioning. *IEEE Trans. Image Process.*, 31:4321–4335.

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

Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. 2019. Selective kernel networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 510–519.

Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan L. Yuille. 2017. Recurrent multimodal interaction for referring image segmentation. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1280–1289. IEEE Computer Society.

Maofu Liu, Xin Jiang, and Xiaokang Zhang. 2025. Cadformer: Fine-grained cross-modal alignment and decoding transformer for referring remote sensing image segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.

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

Sihan Liu, Yiwei Ma, Xiaoqing Zhang, Haowei Wang, Jiayi Ji, Xiaoshuai Sun, and Rongrong Ji. 2024. Rotated multi-scale interaction network for referring remote sensing image segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sun'ao Liu, Yiheng Zhang, Zhaofan Qiu, Hongtao Xie, Yongdong Zhang, and Ting Yao. 2023. CARIS: context-aware referring image segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pages 779–788. ACM.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Qin Ma, Lingling Li, Xiaoqiang Lu, Licheng Jiao, Fang Liu, Wenping Ma, Xu Liu, and Long Sun. 2025. LSCF: long-term semantic-guidance convformer for referring remote sensing image segmentation. *IEEE Trans. Geosci. Remote. Sens.*, 63:1–13.

Yiwei Ma, Jiayi Ji, Xiaoshuai Sun, Yiyi Zhou, and Rongrong Ji. 2023. Towards local visual modeling for image captioning. *Pattern Recognit.*, 138:109420.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloé Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross B. Girshick, Piotr Dollár, and Christoph Feichtenhofer. 2025. SAM 2: Segment anything in images and videos. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

David Rolnick, Priya L. Donti, Lynn H. Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, Alexandra Sasha Luccioni, Tegan Maharaj, Evan D. Sherwin, S. Karthik Mukkavilli, Konrad P. Kording, Carla P. Gomes, Andrew Y. Ng, Demis Hassabis, John C. Platt, and 3 others. 2023. Tackling climate change with machine learning. *ACM Comput. Surv.*, 55(2):42:1–42:96.

Fu Rong, Meng Lan, Qian Zhang, and Lefei Zhang. 2026. RS2-SAM2: customized SAM2 for referring remote sensing image segmentation. In *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pages 8751–8759. AAAI Press.

Yuzhe Sun, Zhe Dong, Haochen Jiang, Tianzhu Liu, and Yanfeng Gu. 2026. Crobim-u: Uncertainty-driven referring remote sensing image segmentation. *CoRR*, abs/2601.03490.

Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. 2022. CRIS: clip-driven referring image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 11676–11685. IEEE.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara

Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics.

Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. 2020. Phrasecut: Language-based image segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zunnan Xu, Zhihong Chen, Yong Zhang, Yibing Song, Xiang Wan, and Guanbin Li. 2023. 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*, pages 17457–17466. IEEE.

Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip H. S. Torr. 2022. LAVT: language-aware vision transformer for referring image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18134–18144. IEEE.

Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg. 2018. Mattnet: Modular attention network for referring expression comprehension. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1307–1315. Computer Vision Foundation / IEEE Computer Society.

Zhenghang Yuan, Lichao Mou, Yuansheng Hua, and Xiao Xiang Zhu. 2024. Rrsis: Referring remote sensing image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–12. Art. no. 5613312.

Yang Zhan, Zhitong Xiong, and Yuan Yuan. 2023. Rsvg: Exploring data and models for visual grounding on remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13.

Duo Zheng, Tao Kong, Ya Jing, Jiaan Wang, and Xiaojie Wang. 2022. Towards unifying reference expression generation and comprehension. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 6598–6611. Association for Computational Linguistics.

Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. 2019. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9308–9316.

Table 5: Benchmark datasets for referring remote sensing image segmentation.

Dataset	Source	Resolution	Size	Triplets	Categories	Attributes
RefSegRS	SkyScapes	0.13 m	512^2	4,420	14	5
RRSIS-D	RSVGD	0.5–30 m	800^2	17,402	20	7

A Experimental Setup

A.1 Dataset Description

We evaluate DLSA on two benchmark datasets for referring remote sensing image segmentation: RefSegRS (Yuan et al., 2024) and RRSIS-D (Liu et al., 2024). Table 5 summarizes their basic statistics.

RefSegRS is built upon the SkyScapes dataset and provides pixel-level object masks paired with referring expressions. It contains 4,420 image–description–mask triplets from 285 scenes, which are split into training, validation, and test subsets with 2,172, 431, and 1,817 triplets, respectively. The annotations cover 14 hierarchical semantic categories, including roads, buildings, cars, and vans, and each object is described with five attributes. All images are provided at a spatial size of 512×512 pixels with a spatial resolution of 0.13 m/pixel.

RRSIS-D is constructed from the RSVGD dataset. It uses the original bounding-box annotations and further processes them with SAM to obtain instance-level masks for RRSIS. The dataset contains 17,402 image–description–mask triplets, covering 20 semantic categories such as aircraft, golf courses, highway service areas, baseball fields, and stadiums, with seven additional attributes for richer semantic descriptions. It is split into 12,181 training triplets, 1,740 validation triplets, and 3,481 test triplets. The released images are resized to 800×800 pixels, and their spatial resolutions range from 0.5 to 30 m/pixel, leading to large target-scale variation across samples. During model training and inference, we further resize all input images to 512×512 pixels for a unified network input size.

A.2 Evaluation Metrics

Following prior work in referring segmentation (Zhan et al., 2023; Wu et al., 2020; Liu et al., 2024), we evaluate model performance using three metrics: Overall Intersection-over-Union (oIoU), Mean Intersection-over-Union (mIoU), and Precision@X. oIoU computes IoU over all pixels, mIoU averages IoU across instances, and Precision@X measures the percentage of samples whose predicted mask achieves an IoU above a threshold $X \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$ with the ground truth. Higher values indicate better text-conditioned segmentation quality.

A.3 Implementation Details

Following RMSIN (Liu et al., 2024), we use Swin Transformer (Liu et al., 2021) as the visual backbone and BERT-base (Devlin et al., 2019) as the language encoder. The visual backbone is initialized with ImageNet22K pre-trained weights (Deng et al., 2009), and the language encoder adopts the base BERT model from HuggingFace's library (Wolf et al., 2020). We also follow RMSIN in the main optimization setting, using AdamW (Loshchilov and Hutter, 2017) with a learning rate of 3×10^{-5}, a weight decay of 0.01, and a polynomial learning-rate decay schedule.

For the language encoder, we adopt partial fine-tuning. Specifically, the first ten BERT encoder layers are included in the optimizer, while the remaining language-encoder parameters are kept unchanged during optimization. All trainable parameter groups share the same base learning rate, and no discriminative layer-wise learning rate is used.

The segmentation objective is a weighted combination of pixel-wise cross-entropy loss and Dice loss:

$$\mathcal{L} = 0.9\mathcal{L}_{\text{CE}} + 0.1\mathcal{L}_{\text{Dice}}, \tag{23}$$

where the cross-entropy term uses class weights $[0.9, 1.1]$. Focal loss is not used.

All modules are implemented in PyTorch. We train the model on four RTX 4090 GPUs with a global batch size of 8. For RRSIS-D, we follow the training and evaluation protocol of RMSIN (Liu et al., 2024), including the official data split and evaluation metrics. During training and inference, all images are resized to 512×512 as network inputs, converted to tensors, and normalized with ImageNet mean and standard deviation. No additional random data augmentation is applied.

B Additional Ablation Studies

B.1 Design of SgAMC

Table 6 analyzes the contribution of different SgAMC components on RRSIS-D and RefSegRS. Overall, the full SgAMC achieves the best mIoU and P@0.7 on both datasets, demonstrating the effectiveness of combining semantics-aware routing, the purely visual branch, and the dynamic branch balancing gate.

On RRSIS-D, removing the dynamic branch balancing gate causes the largest mIoU drop, from

Table 6: Ablation study on the components of SgAMC on RRSIS-D and RefSegRS.

Dataset	Setting	P@0.5	P@0.7	P@0.9	mIoU
RRSIS-D	w/o balancing gate	71.93	53.22	22.81	62.69
	w/o pure conv	74.43	56.13	**24.96**	64.33
	w/o semantics-aware routing	72.85	55.39	23.56	64.20
	SgAMC	**74.52**	**56.33**	24.65	**64.61**
RefSegRS	w/o balancing gate	90.95	70.07	6.73	71.98
	w/o pure conv	91.42	70.53	6.96	72.22
	w/o semantics-aware routing	90.49	71.93	6.73	72.19
	SgAMC	**93.74**	**80.28**	**9.05**	**75.38**

Table 7: Ablation study on the kernel set used in SgAMC on RRSIS-D and RefSegRS. We start from the full multi-branch configuration with four parallel kernels $\{1, 3, 5, 7\}$ and remove one kernel size at a time.

Dataset	Setting	Kernels used	mIoU	oIoU	P@0.7
RRSIS-D	w/o 1×1	$\{3, 5, 7\}$	63.60	76.95	55.47
	w/o 3×3	$\{1, 5, 7\}$	64.26	76.81	55.87
	w/o 5×5	$\{1, 3, 7\}$	64.04	77.18	**56.33**
	w/o 7×7	$\{1, 3, 5\}$	64.08	77.40	56.26
	SgAMC	$\{1, 3, 5, 7\}$	**64.61**	**77.49**	**56.33**
RefSegRS	w/o 1×1	$\{3, 5, 7\}$	72.87	79.77	71.46
	w/o 3×3	$\{1, 5, 7\}$	71.09	79.36	68.21
	w/o 5×5	$\{1, 3, 7\}$	70.41	79.26	67.29
	w/o 7×7	$\{1, 3, 5\}$	73.40	79.75	74.01
	SgAMC	$\{1, 3, 5, 7\}$	**75.38**	**81.60**	**80.28**

64.61% to 62.69%, indicating the importance of adaptively balancing visual and language-guided responses. Removing the purely visual branch or the semantics-aware routing branch also degrades mIoU, showing that both language-independent spatial evidence and semantic guidance contribute to stable semantic–scale alignment. On RefSegRS, all ablated variants also underperform the full SgAMC in terms of mIoU and P@0.7. In particular, removing the semantics-aware routing branch reduces mIoU from 75.38% to 72.19%, suggesting that semantic guidance remains beneficial even when the visual features already provide relatively strong cues. Although some ablated variants obtain higher P@0.9 on RRSIS-D, their overall mIoU and medium-threshold precision are lower, indicating that the complete SgAMC provides more stable segmentation quality across samples.

B.2 Kernel Set Analysis

Table 7 studies the influence of different kernel sets in SgAMC on RRSIS-D and RefSegRS. We start from the full multi-branch configuration with four parallel kernels $\{1, 3, 5, 7\}$ and remove one kernel size at a time. Overall, the complete kernel set achieves the best performance on both datasets, indicating that different receptive-field branches provide complementary information for semantic–scale alignment.

On RRSIS-D, the full kernel set obtains the highest mIoU and oIoU, while removing any single branch leads to a performance drop. In particular, removing the 1×1 branch causes the largest mIoU degradation, suggesting that the local branch is important for preserving fine details and small structures. On RefSegRS, the full kernel set also consistently outperforms all reduced variants by a clear margin, further confirming the robustness of multi-scale receptive-field modeling. These results show that SgAMC benefits from combining local, middle-range, and broader contextual responses, rather than relying on a single or incomplete kernel configuration.

B.3 Effect of Token-aware Routing Variant

We further evaluate a token-aware routing variant to investigate whether explicitly modeling word-level semantic cues can further improve the receptive-

Table 8: Effect of the token-aware routing variant on RRSIS-D and RefSegRS.

Dataset	Variant	P@0.5	P@0.7	P@0.9	oIoU	mIoU
RRSIS-D	Original SgAMC	74.52	**56.33**	24.65	**77.49**	**64.61**
	Token-aware SgAMC	**74.58**	56.10	**24.85**	77.12	64.56
RefSegRS	Original SgAMC	**93.74**	**80.28**	**9.05**	**81.60**	**75.38**
	Token-aware SgAMC	90.95	79.35	7.89	81.08	73.72

Table 9: Ablation study on the injection stages of ALTI on RRSIS-D and RefSegRS.

Dataset	Variant	P@0.5	P@0.6	P@0.7	P@0.8	P@0.9	oIoU	mIoU
RRSIS-D	Early-stage injection	75.17	68.16	57.01	44.37	24.60	77.93	65.53
	All-stage injection	75.23	68.39	57.41	44.60	24.54	77.99	65.48
	Late-stage injection	**75.75**	**69.33**	**57.93**	**45.17**	**24.89**	**78.14**	**65.78**
RefSegRS	Early-stage injection	94.66	**90.26**	83.29	45.01	9.28	81.94	76.39
	All-stage injection	**94.90**	89.33	82.37	43.16	**9.51**	81.95	76.14
	Late-stage injection	**94.90**	**90.26**	**84.45**	**46.64**	9.28	**82.00**	**76.69**

field selection of SgAMC. As reported in Table 8, the token-aware variant achieves comparable but not consistently better results than the original SgAMC. On RRSIS-D, it obtains slightly higher P@0.5 and P@0.9, but its mIoU and oIoU decrease from 64.61% and 77.49% to 64.56% and 77.12%, respectively. On RefSegRS, the original SgAMC performs better across all reported metrics, with mIoU decreasing from 75.38% to 73.72% and oIoU decreasing from 81.60% to 81.08% after introducing token-aware routing.

These results indicate that explicitly introducing token-level routing does not necessarily lead to more reliable semantic-scale selection. A possible reason is that the original SgAMC already captures sufficient sentence-level semantic guidance for adaptive multi-scale modeling, while finer token-level routing may introduce additional sensitivity to noisy or weakly discriminative words in referring expressions. Therefore, we adopt the original SgAMC as the default design in our final model, and treat the token-aware routing scheme as an exploratory variant.

B.4 Effect of Text Injection Stages in ALTI

To analyze the influence of text injection stages in ALTI, we compare three variants on RRSIS-D and RefSegRS: early-stage injection, all-stage injection, and the proposed late-stage injection. As shown in Table 9, late-stage injection achieves the best overall performance on both datasets. On RRSIS-D, late-stage injection obtains the highest mIoU, oIoU, and precision scores across all thresholds, outperforming both early-stage and all-stage injection. On RefSegRS, late-stage injection also achieves the highest mIoU, oIoU, P@0.7, and P@0.8, while obtaining competitive results on P@0.5 and P@0.6. Although all-stage injection achieves a slightly higher P@0.9 on RefSegRS, its mIoU and medium-threshold precision are lower than those of late-stage injection.

These results indicate that applying textual guidance to late decoding stages provides a better trade-off between semantic refinement and structural stability. Early-stage injection may interfere with structural recovery, while all-stage injection introduces textual residuals throughout the decoder and may cause unnecessary perturbation. In contrast, late-stage injection better matches the role of ALTI: supplementing semantic information during mask refinement while preserving the stability of early structural reconstruction.

B.5 Component Analysis of ALTI

To further analyze the design of ALTI, we conduct component-level ablation experiments on RRSIS-D and RefSegRS. As shown in Table 10, the full ALTI achieves the highest mIoU on both datasets, indicating that its components jointly contribute to stable semantic refinement during mask reconstruction.

On RRSIS-D, removing the sample-wise gate, local gate, or residual scale reduces mIoU from 65.78% to 65.55%, 65.31%, and 65.64%, respec-

Table 10: Component ablation of ALTI on RRSIS-D and RefSegRS.

Dataset	Variant	P@0.5	P@0.6	P@0.7	P@0.8	P@0.9	oIoU	mIoU
RRSIS-D	w/o sample-wise gate	75.06	67.47	56.38	44.77	24.60	77.99	65.55
	w/o local gate	75.17	67.99	57.87	44.94	24.66	77.92	65.31
	w/o residual scale	75.17	68.45	57.24	43.68	24.60	77.71	65.64
	ALTI	**75.75**	**69.33**	**57.93**	**45.17**	**24.89**	**78.14**	**65.78**
RefSegRS	w/o sample-wise gate	93.74	89.79	81.21	40.37	**9.98**	81.66	75.68
	w/o local gate	93.74	89.79	81.90	39.91	9.74	81.71	75.76
	w/o residual scale	93.50	89.33	79.81	40.37	**9.98**	81.56	75.54
	ALTI	**94.90**	**90.26**	**84.45**	**46.64**	9.28	**82.00**	**76.69**

Table 11: Computational cost comparison of different variants. FLOPs are estimated by fvcore. Latency is measured with token length 20, batch size 1, and AMP enabled.

Variant	Params (M)	ΔParams (M)	Param. Ratio (%)	FLOPs (G)	ΔFLOPs (G)	FLOPs Ratio (%)	Latency (ms)	mIoU (%)
Baseline	240.04	–	–	202.10	–	–	55.98 ± 0.29	72.73
+SgAMC	242.51	2.47	1.03	206.24	4.14	2.05	62.51 ± 0.52	75.38
+ALTI	244.11	4.07	1.70	221.01	18.91	9.36	60.94 ± 0.28	73.39
DLSA	246.58	6.54	2.72	225.15	23.05	11.41	65.73 ± 0.41	76.69

tively. The degradation is also observed on most precision metrics, especially P@0.7 and P@0.8, where the full ALTI obtains the best results. These results show that the sample-wise gate helps control the global strength of textual residuals, the local gate provides spatial-channel modulation, and the residual scale stabilizes residual injection during late-stage decoding.

On RefSegRS, similar trends can be observed. Removing the sample-wise gate, local gate, or residual scale decreases mIoU from 76.69% to 75.68%, 75.76%, and 75.54%, respectively. The full ALTI also achieves the best results on P@0.5, P@0.6, P@0.7, P@0.8, oIoU, and mIoU, demonstrating its overall effectiveness for controlled semantic refinement. Although some ablated variants obtain slightly higher P@0.9, their mIoU and medium-threshold precision are consistently lower, suggesting that the complete ALTI provides more stable overall segmentation quality rather than optimizing only a single strict-threshold metric.

C Additional Analysis

C.1 Computational Cost

We further analyze the computational cost of different variants. As shown in Table 11, SgAMC introduces 2.47M additional parameters, accounting for only 1.03% of the baseline model size, with 4.14G additional FLOPs. ALTI introduces 4.07M parameters, corresponding to 1.70% of the baseline parameters, and brings 18.91G additional FLOPs due to adaptive textual residual modeling in late decoding stages. The full DLSA adds 6.54M parameters in total, only 2.72% over the baseline, while achieving the highest mIoU. These results indicate that DLSA improves segmentation performance with moderate computational overhead and without substantially increasing the model size.

C.2 Scale-based Subset Analysis

To complement the semantic–scale routing analysis, we further conduct a scale-based subset evaluation on RRSIS-D and RefSegRS. Samples are divided into Small, Medium, and Large groups according to the ground-truth mask area, using the bottom 30%, middle 40%, and top 30% of the area distribution, respectively. As shown in Table 12, DLSA generally improves the baseline across different target scales on both datasets. On RefSegRS, DLSA brings consistent gains on all scale groups, with particularly large improvements on small and medium targets, improving mIoU by 6.43% and 3.60%, respectively. On RRSIS-D, the improvements are more moderate, where DLSA improves mIoU by 1.59% on small targets and 0.76% on medium targets, while maintaining comparable performance on large targets. These results indicate that DLSA improves segmentation robustness un-

Table 12: Scale-based subset analysis on RRSIS-D and RefSegRS. Targets are divided into Small, Medium, and Large groups according to the ground-truth mask area, using the bottom 30%, middle 40%, and top 30% of the area distribution, respectively.

Dataset	Subset	Method	N	P@0.5	P@0.7	P@0.9	oIoU	mIoU
RRSIS-D	Small	Baseline	522	53.26	**21.84**	0.00	37.49	42.75
		DLSA	522	**55.56**	21.46	0.00	**38.99**	**44.34**
	Medium	Baseline	696	78.02	64.08	**23.85**	61.98	68.34
		DLSA	696	**78.88**	**65.95**	23.13	**62.00**	**69.10**
	Large	Baseline	522	91.57	83.33	51.53	83.13	**82.83**
		DLSA	522	**91.76**	**83.91**	**52.11**	**83.20**	82.81
	All	Baseline	1740	74.66	57.18	**25.00**	78.10	65.01
		DLSA	1740	**75.75**	**57.93**	24.89	**78.14**	**65.78**
RefSegRS	Small	Baseline	129	88.37	66.67	3.88	65.48	70.10
		DLSA	129	**92.25**	**84.50**	**4.65**	**69.51**	**76.53**
	Medium	Baseline	173	91.91	65.32	2.31	68.62	70.49
		DLSA	173	**95.38**	**84.39**	**2.89**	**72.45**	**74.09**
	Large	Baseline	129	96.12	78.29	17.05	81.75	78.37
		DLSA	129	**96.90**	**84.50**	**22.48**	**83.46**	**80.33**
	All	Baseline	431	92.11	69.61	7.19	80.01	72.73
		DLSA	431	**94.90**	**84.45**	**9.28**	**82.00**	**76.69**

Table 13: Performance of SgAMC on different semantic–scale correlation groups on RRSIS-D.

Group	P@0.5	P@0.6	P@0.7	P@0.8	P@0.9	oIoU	mIoU
Overall	74.52	67.57	56.33	42.75	24.65	77.49	64.61
S-SR	**75.23**	**69.70**	**61.40**	**48.09**	**29.64**	**79.61**	**65.97**
W-SR	74.32	66.97	54.89	41.26	23.25	76.22	64.24

der scale variations, and the gains are not limited to a single target-size group.

C.3 Semantic–Scale Correlation

We further analyze whether SgAMC benefits from explicit semantic–scale correlation. Following the automatic grouping rule in our evaluation code, evaluation samples are divided into two groups according to the relation between scale-related expressions and target geometry. The strong semantic–scale relation group (S-SR) contains samples whose referring expressions include explicit scale- or shape-related cues that are consistent with the ground-truth mask geometry, while the weak semantic–scale relation group (W-SR) contains the remaining samples.

As shown in Table 13, S-SR consistently outperforms W-SR on RRSIS-D across all evaluation metrics. Specifically, S-SR improves mIoU from 64.24% to 65.97% and oIoU from 76.22% to 79.61%. The improvement is more evident under stricter precision thresholds, where P@0.7, P@0.8, and P@0.9 increase from 54.89%, 41.26%, and 23.25% to 61.40%, 48.09%, and 29.64%, respectively. These results indicate that explicit scale-related linguistic cues help SgAMC perform more reliable scale-aware feature adaptation, especially for high-quality mask prediction.

C.4 Text Perturbation Analysis.

We further analyze the behavior of ALTI under incomplete or noisy referring expressions. Specifically, we compare the original expressions with three perturbed settings: Drop30, Mask30, and Mask50. As shown in Table 14, the performance consistently decreases when textual information becomes less reliable. For example, Mask50 reduces mIoU from 65.78 to 52.44 on RRSIS-D and from 76.69 to 58.28 on RefSegRS, indicating that sufficient discriminative language cues remain essential

Dataset	Text setting	mIoU	Δ mIoU	oIoU	Δ oIoU	P@0.5	P@0.7	P@0.9	Mean α	Std α
RRSIS-D	Clean	65.78	–	78.14	–	75.75	57.93	24.89	0.2673	0.0002
	Drop30	56.94	-8.84	71.18	-6.96	64.08	48.68	21.61	0.2674	0.0002
	Mask30	58.84	-6.94	71.88	-6.26	66.72	51.26	22.47	0.2673	0.0001
	Mask50	52.44	-13.34	63.79	-14.34	58.62	44.08	19.37	0.2674	0.0001
RefSegRS	Clean	76.69	–	82.00	–	94.90	84.45	9.28	0.2407	0.0028
	Drop30	64.23	-12.46	66.24	-15.77	79.58	65.43	5.80	0.2412	0.0028
	Mask30	65.40	-11.28	70.46	-11.54	80.97	67.29	6.73	0.2412	0.0027
	Mask50	58.28	-18.41	62.74	-19.26	71.93	58.24	5.80	0.2415	0.0026

Table 14: Text perturbation analysis on RRSIS-D and RefSegRS. Drop30 removes 30% valid tokens, while Mask30 and Mask50 replace 30% and 50% valid tokens with [MASK], respectively. Δ denotes the performance change compared with the clean expression setting.

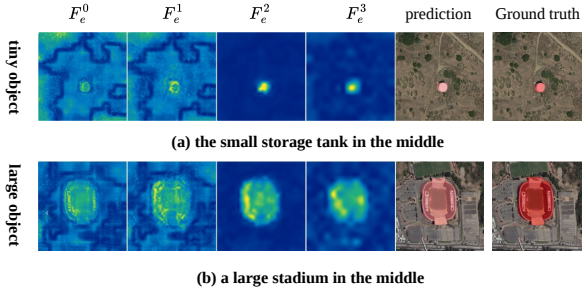


Figure 4: **Stage-wise encoder response visualization with SgAMC.** We visualize encoder activation heatmaps under queries describing small and large targets.

for accurate target localization and mask prediction. Meanwhile, the mean injection coefficient of ALTI remains stable across perturbation settings. This suggests that ALTI performs bounded residual text injection with a relatively stable injection strength, rather than aggressively amplifying noisy textual inputs. Therefore, ALTI provides controlled semantic refinement when reliable textual cues are available, while missing or corrupted referring information remains a challenging failure condition.

D Additional Qualitative Results

D.1 Encoder Feature Visualization

To qualitatively illustrate the effect of semantic guidance, we visualize stage-wise encoder response heatmaps for queries involving small and large targets. As shown in Fig. 4, the activation patterns vary across stages and reflect differences in the semantic scale implied by the referring expressions.

D.2 Failure Cases

Despite overall improvements in coherence and semantic alignment, DLSA still exhibits two typical failure modes. **(1) Semantic confusion** among visually similar instances: when the referring expression contains ambiguous or fine-grained cues, the semantic routing branch may mistakenly attend to a non-target instance, as shown in Fig. 5(a). **(2) Boundary degradation** on fine structures: scale-adaptive modulation in SgAMC may smooth high-frequency details, leading to noticeable IoU loss around object boundaries, especially for tiny instances, as shown in Fig. 5(b,c). Future work will focus on disambiguating subtle semantic differences and enhancing scale/shape modeling to mitigate these issues.

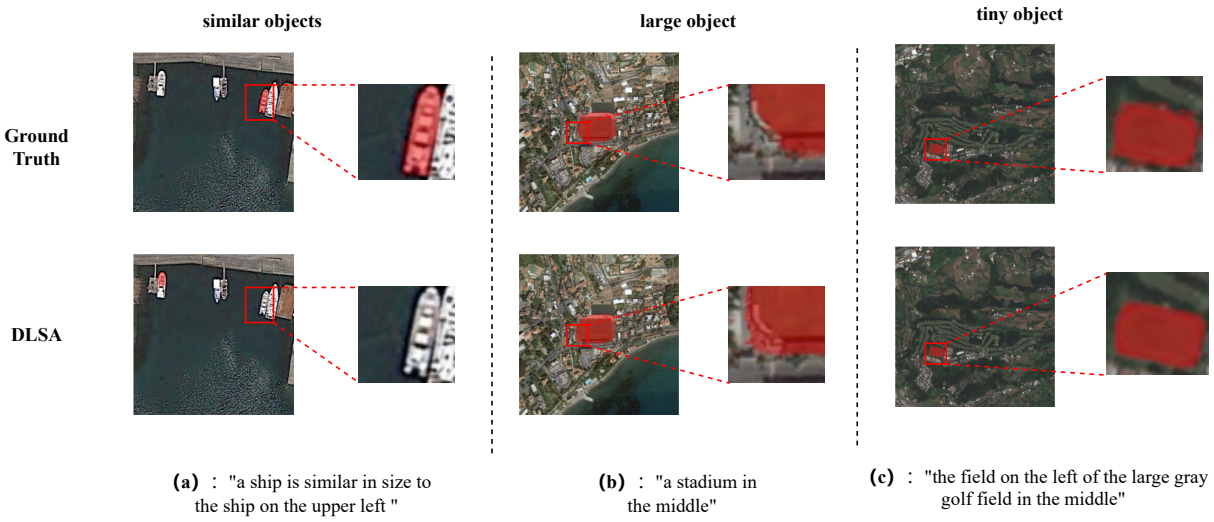


Figure 5: Representative failure cases of DLSA.