CROSS: Cascaded Distillation and Dual-Constraint Grounding for Remote Sensing Referring Segmentation

Tingzhang Luo[*1], Ruizhong Liu[*2], Yichao Liu[3], Cheng Fan[1], Yu Liu[4], and Jianyuan Guo[†1]

[1] City University of Hong Kong
[2] The Hong Kong University of Science and Technology (Guangzhou)
[3] Nankai University
[4] Peking University

Abstract. Referring Remote Sensing Image Segmentation (RRSIS) has achieved significant progress through the integration of VLMs and the Segment Anything Model (SAM). However, this progress largely relies on strong pre-trained capabilities, while leaving two fundamental limitations insufficiently addressed: *(1) Architectural Weak-Coupling*, where the unidirectional flow forces reliance on coarse VLM prompts and wastes SAM's pixel-level structural guidance, causing localization drift; and *(2) Object-Centric Semantic Bias*, where models overemphasize dominant object semantics while remaining insensitive to spatial reasoning crucial for RRSIS. Motivated by these observations, we propose CROSS, a tightly integrated paradigm for RRSIS. First, we introduce Linguistic-Guided Cascaded Distillation (LGCD) to bridge the architectural gap, which distills SAM's geometric affinities as soft regularizers into VLM intermediate layers, injecting dense structural priors to refine localization. Second, Perspective-Spatial Contrastive Learning (PSCL) imposes cross-anchored constraints by mining mask-filtered deceptive distractors and spatial-linguistic counterfactuals as hard negatives, explicitly shattering semantic shortcuts to enforce genuine logical consistency. Extensive experiments on RRSIS benchmarks demonstrate that CROSS achieves state-of-the-art performance and maintains precise localization even under severe spatial description perturbations, standing as a robust new paradigm for RRSIS. `https://clarence-cv.github.io/CROSS/`.

Keywords: Remote Sensing · Referring Segmentation · Contrastive Learning

1 Introduction

Recent advances in visual understanding [5, 21, 24, 30] have increasingly emphasized models' ability to handle complex semantics, diverse visual patterns [41, 42], and open-ended real-world scenarios [22, 40, 66, 68]. Referring Remote Sensing

[*] Equal contribution. [†] Corresponding author.

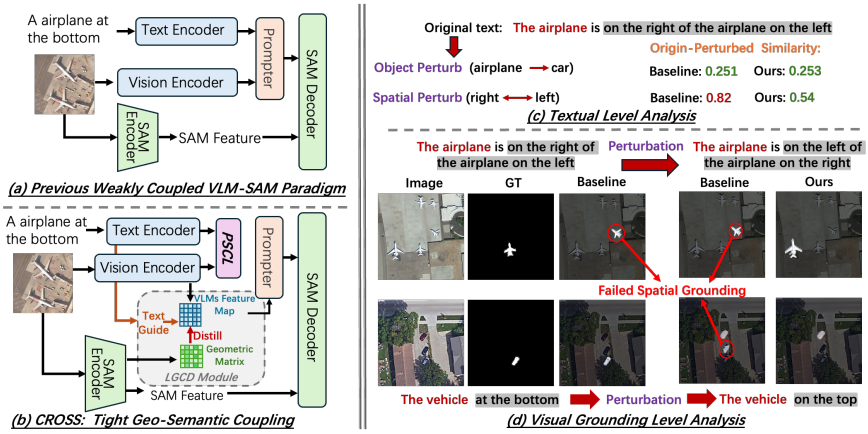


Fig. 1: Motivation of the CROSS framework. **(Left) Architectural Paradigm Comparison:** (a) Previous weakly-coupled pipelines treat VLM and SAM as isolated, fragmented modules, where SAM merely serves as a passive executor of explicit prompts. *(b)* Our deeply-coupled CROSS performs text-guided distillation of SAM-derived spatial affinity matrices into the VLM to enforce structural constraints. **(Right) Spatial Logic Probing:** *(c)* Textual Level Analysis: Baseline methods show sensitivity to object categories but remain invariant to spatial perturbations, indicating a lack of directional awareness. *(d)* Visual Grounding Analysis: Baseline exhibits logical collapse under complex spatial cues, whereas CROSS maintains logical soundness by precisely anchoring masks to the complete linguistic context.

Image Segmentation (RRSIS) [23, 37, 54, 71] aims to segment specific targets in Earth observation imagery, strictly adhering to linguistic referring expressions. Unlike natural scene Referring Image Segmentation (RIS) [11, 31, 62], RRSIS confronts unique complexities: immense visual scale, extreme homogeneity, and dense, multi-scale targets. To tackle these challenges, leveraging foundation models has become a promising paradigm. Within this evolving landscape, research efforts have bifurcated into two primary directions. While Large Multimodal Models (LMMs) [18, 26, 33, 47, 52] excel in holistic understanding and complex reasoning, the specific task of RRSIS demands precise pixel-level localization rather than generative text capabilities. Consequently, the discriminative CLIP-based paradigm (*e.g.*, SigLIP [55, 73] coupled with SAM [16, 50]) has emerged as the compelling choice, prioritizing dense feature alignment over semantic reasoning. Adhering to this latter paradigm, pioneering works such as RSRefSeg [3] and RSRefSeg 2 [2] have sought to transpose this efficient VLM-SAM[4] pipeline to remote sensing scenarios.

However, rather than achieving a genuine multi-modal synergy tailored for the dense and complex nature of Remote Sensing (RS) scenarios, current adap-

[4] The term *VLM* in our work refers to CLIP-style models. Generative MLLM such as LLaVA [33] or Qwen-VL [1] are not considered here, as preliminary evidence suggests they remain less competitive than CLIP-based approaches for precise RRSIS tasks.

tations largely coast on inherent pre-trained capabilities of the standalone foundation models. By deeply dissecting this superficial integration, we identify two fundamental bottlenecks that critically paralyze existing VLM-SAM pipelines (as illustrated in Fig. 1).

• **Bottleneck I: Architectural Weak-Coupling.** We characterize existing pipelines as *"weakly coupled"* due to their strictly unidirectional information flow (Fig. 1 a). In such designs, the VLM produces coarse semantic prompts that are subsequently consumed by SAM, while SAM's powerful image encoder remains entirely unused during prompt generation. This design overlooks a critical opportunity: SAM encodes rich pixel-level structural priors that could otherwise guide semantic localization. As a result, VLM-derived dense prompts often exhibit spatially diffuse responses in complex remote sensing scenes, reflecting the limited localization capability of CLIP-style vision encoders when operating without structural constraints. This observation raises a key question: ***Can SAM's structural priors be injected back into the VLM to explicitly constrain semantic localization?***

• **Bottleneck II: Object-Centric Semantic Bias.** Beyond architectural isolation, existing pipelines also inherit a strong bias from VLM pre-training. CLIP-style models are primarily trained for global image–text alignment, which encourages strong associations with dominant object semantics while providing limited supervision for spatial relations. As illustrated in Fig. 1, our perturbation analysis reveals a critical failure mode. When spatial attributes in the referring expression are altered, the baseline model still produces a high similarity response (Fig. 1c), yet the predicted localization becomes entirely incorrect (Fig. 1d). This phenomenon indicates that the model largely relies on object-level semantic cues while remaining insensitive to spatial descriptions that are crucial for RRSIS. Consequently, a second key question arises: ***How can we explicitly guide the network to overcome this object-centric bias and internalize robust, text-driven spatial reasoning?***

To directly address these bottlenecks, we propose CROSS: **C**ascaded Distillation and Dual-Constraint Grounding for **R**em**O**te **S**ensing Referring **S**egmentation (overall architecture illustrated in Fig. 2). Unlike conventional weakly coupled pipelines, our approach establishes a tightly integrated paradigm utilizing text-guided SAM distillation (Fig. 1 b). Specifically, to operationalize this paradigm and overcome the aforementioned architectural isolation, we propose **L**inguistic-**G**uided **C**ascaded **D**istillation (**LGCD**). Guided by an adaptive text-driven soft mask, LGCD abstracts SAM's structural priors into geometric affinity matrices and cascadedly distills them into SigLIP's shallow, deep, and final layers. Crucially, transferring relative affinities acts as a soft regularizer. By strategically leveraging different layers, this mechanism explicitly augments spatial configurations without corrupting SigLIP's inherent hierarchical semantic space. This structural enhancement is also visually corroborated by the highly concentrated spatial activations demonstrated in Fig. 5.

Furthermore, to mitigate the object-centric bias inherited from VLM pre-training and strengthen spatial reasoning, we introduce the **P**erspective-**S**patial

Contrastive **L**earning (**PSCL**) scheme. Formulated as a complementary cross-anchored contrastive space, PSCL imposes dual-level constraints. At the visual perspective level, we utilize the target mask to physically blind the true instance, explicitly mining deceptive background distractors that exhibit high semantic affinity to the text yet violate the intended spatial constraints. Concurrently, at the linguistic spatial level, we synthesize counterfactual text perturbations (*e.g.*, swapping "left" with "right") to strictly penalize spatial inconsistencies. By jointly addressing visual ambiguity and relational reasoning, PSCL encourages robust spatial-semantic grounding, ensuring accurate target localization even under severe linguistic perturbations (Fig. 1 right).

Our contributions can be summarized as follows: **(i) Conceptually**, we identify two fundamental bottlenecks in existing RRSIS paradigms: *Architectural Weak-Coupling*, which leads to localization drift due to the unidirectional flow, and *Object-Centric Semantic Bias*, where pre-trained VLMs predominantly rely on dominant object semantics as a shortcut, remaining insensitive to the spatial descriptions crucial for dense localization. **(ii) Methodologically**, CROSS introduces two core mechanisms: LGCD, which distills SAM's geometric affinities as soft structural regularizers to bridge the VLM's geometric-semantic gap; and PSCL, a heterogeneous contrastive construct that rectifies perspective and spatial-linguistic inconsistencies via dual-level hard negative mining. **(iii) Experimentally**, CROSS delivers superior performance and exceptional spatial referring robustness on RRSIS benchmarks.

2 Related Work

Referring Remote Sensing Image Segmentation (RRSIS). As an evolving paradigm in Earth observation [8, 35, 44, 57, 58, 79], RRSIS enables the extraction of specific spatial regions guided by textual prompts [56]. Unlike the remote sensing visual grounding (RSVG) [19, 53, 74] task that focuses on region understanding, RRSIS emphasizes fine-grained pixel-level analysis. Nevertheless, research in this field is still in its early stages and exploration is limited. Yuan et al. [71] first introduced this task, constructed the RefsegRS dataset and adopted the LAVT [70] framework to solve it. Following this, the first large-scale benchmark, RRSIS-D, was built upon the DIOR-RSVG dataset [74] by Liu et al. [37]. To handle arbitrary target orientations and dramatic scale changes, they developed RMSIN, a framework centered on rotational convolutions. In addition, Lei et al. propose FIANet [20], which focuses more on adaptive understanding of objects at different scales and fine-grained vision-language interaction. Recently, Chen et al. introduced the RSRefSeg [3] and RSRefSeg2 [2] methods by leveraging SigLIP to generate prompts for SAM.

Segment Anything Model. SAM [16] and its successor SAM 2 [50] have established a new paradigm for promptable segmentation, demonstrating robust generalization across diverse domains including remote sensing [7, 46, 67], video tracking, and medical imaging [4,59,72]. While various optimized versions [64,80] have enhanced its computational performance, a fundamental limitation remains:

SAM lacks inherent linguistic understanding. To bridge this gap, MLLM-based frameworks—such as LISA [18], u-LLaVA [65], and EVF-SAM [77]—leverage multimodal large language models to generate text-driven embeddings for SAM. However, these models are primarily optimized for natural images and often struggle with the extreme visual homogeneity and intricate spatial layouts characteristic of remote sensing scenes [76]. This underscores the need for specialized structural-semantic alignment tailored for RRSIS.

Vision-Language Models (VLMs). VLMs aim to establish a unified embedding space for cross-modal alignment, supporting a wide range of tasks such as image-text retrieval [15, 49, 73], reasoning segmentation [18, 39, 61, 69], remote sensing analysis [25, 26, 28, 32, 63, 75, 78], and beyond [14]. While previous cascaded works [29] rely on internal semantic refinement, CROSS leverages SAM to distill external geometric priors, effectively bridging the structural-semantic gap.

3 Preliminary

Our referring remote sensing image segmentation framework is built upon a prompt-driven multimodal architecture. The base pipeline consists of three fundamental components: 1) a vision-language foundation model (**SigLIP 2** [55]) for multimodal feature extraction, 2) a **Cross-Modal Prompter** that translates semantic features into prompts, and 3) the Segment Anything Model (**SAM 2**) as the mask decoder.

3.1 Problem Definition

Given a remote sensing image $I \in \mathbb{R}^{H_0 \times W_0 \times 3}$ and a natural language referring expression T, the remote sensing referring segmentation task aims to generate a binary segmentation mask $M \in \{0,1\}^{H_0 \times W_0}$, where $M_{ij} = 1$ indicates that pixel (i, j) belongs to the target region described by text T.

3.2 Vision-Language Model SigLIP 2

We employ SigLIP 2 [55] to extract dual-modal embeddings. For an image I and text T, the encoders $\mathcal{E}_v$ and $\mathcal{E}_t$ generate aligned features:

$$F_v = \mathcal{E}_v(I) \in \mathbb{R}^{H_v \times W_v \times C}, \quad F_t = \mathcal{E}_t(T) \in \mathbb{R}^{L \times C}, \tag{1}$$

here F_v represents the visual features and F_t denotes the textual embeddings of length L. Here, C signifies the shared embedding dimension where visual and linguistic semantics are unified through contrastive pre-training.

3.3 Cross-Modal Prompter

The prompter receives the multimodal features F_v and F_t as input. It generates a dense semantic prompt $P \in \mathbb{R}^{H_v \times W_v}$ by computing the cross-modal similarity

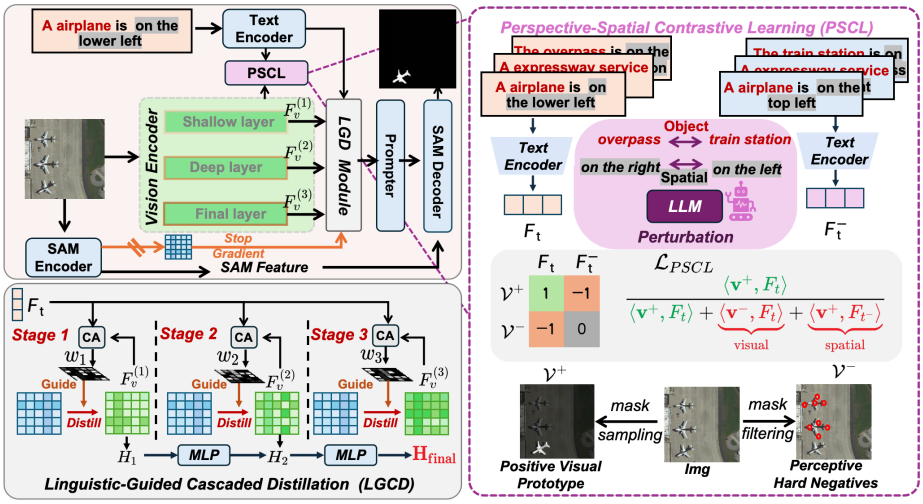


Fig. 2: Overview of the proposed CROSS. Our architecture integrates Linguistic-Guided Cascaded Distillation (LGCD) to inject SAM's structural priors into hierarchical VLM layers, and Perspective-Spatial Contrastive Learning (PSCL) which constructs visual and spatial negative samples to enhance spatial-semantic sensitivity.

between the visual feature map and the textual embeddings. This prompt P serves as a spatial prior and is passed to the SAM 2 mask decoder to guide target localization and final segmentation.

3.4 Segment Anything Model (SAM 2)

SAM 2 [50] functions as the prompt-guided mask decoder. It receives the image features F_{img} from the SAM encoder and the dense semantic prompt P. The mask decoder $\mathcal{D}_{mask}$ then integrates these inputs to generate the final binary segmentation mask M:

$$M = \mathcal{D}_{mask}(F_{img}, P) \in \{0,1\}^{H\times W}, \tag{2}$$

4 Methodology

In this section, we present CROSS (Fig. 2), which primarily introduces two core mechanisms: Linguistic-Guided Cascaded Distillation (LGCD) and Perspective-Spatial Contrastive Learning (PSCL).

4.1 Linguistic-Guided Cascaded Distillation (LGCD)

To enhance spatial precision, we propose LGCD, which injects class-agnostic geometric priors from SAM 2 encoder into the SigLIP-2 encoder via a text-aware filtering mechanism.

Cascaded Representation Extraction (CRE). Instead of relying on a singular terminal output, we harvest a cascaded fashion of visual representations $\{F^{(i)}\}_{i=1}^{3}$ from the shallow, deep, and final blocks of the SigLIP 2 encoder. Each feature map is unrolled into a token sequence $F_v^{(i)} \in \mathbb{R}^{N\times C}$, where $N = H_v \times W_v$ corresponds to the spatial resolution of the vision encoder's patch grid. At each stage i, the visual state is updated via a residual injection: $X_i = H_{i-1} + F_v^{(i)}$ ($H_0 = \mathbf{0}$), where the shared spatial dimension N enables direct patch-level summation. To integrate linguistic context as a spatial safeguard, we apply a bidirectional Cross-Attention (CA) mechanism at each stage:

$$H_i = \mathrm{MLP}(\mathrm{LN}(X_i + \gamma_i \cdot \mathrm{CA}(Q{=}X_i, K{=}F_t, V{=}F_t))) \tag{3}$$

where $\mathrm{LN}(\cdot)$ and $\mathrm{MLP}(\cdot)$ denote Layer Normalization and a two-layer feed-forward network, and γ_i is a zero-initialized learnable scalar to ensure training stability. This process yields the refined feature $H_i \in \mathbb{R}^{N\times C}$. Simultaneously, we obtain a cross-attention matrix $A_i \in \mathbb{R}^{N\times L}$ derived from the scaled dot-product between visual queries and text keys, which represents the attention weights of N spatial patches over L text tokens.

Text-Guided Relational Distillation (TGD). At each stage i, we distill the topological priors from the SAM 2 terminal feature into H_i. Given the severe background clutter in remote sensing scenes, a naive dense distillation of SAM's class-agnostic features would introduce massive structural noise. Therefore, we repurpose the cross-attention matrix A_i as a semantic mask to conditionally route only the linguistically relevant topology. For each spatial location p, the mask $M_{\text{text}}^i \in \mathbb{R}^N$ is aggregated across all L tokens:

$$M_{\text{text}}^i(p) = \frac{1}{L}\sum_{j=1}^{L} A_i^{(p,j)}, \tag{4}$$

Direct pixel-wise alignment between the heterogeneous latent spaces of SAM and the VLM inevitably causes feature distortion. Instead, we transfer the structural priors by aligning their relative spatial topologies. Specifically, we compute the Gram matrix $\mathcal{G}(\cdot) \in \mathbb{R}^{N\times N}$ to capture the pairwise feature affinities:

$$\mathcal{G}(F)_{p,q} = \frac{\langle f_p, f_q\rangle}{\|f_p\|_2\|f_q\|_2}, \tag{5}$$

where f_p, f_q represent the feature vectors at spatial positions p and q. The text-guided relational distillation loss is explicitly formulated to conditionally enforce this structural alignment:

$$\mathcal{L}_{\text{distill}}^i = \frac{1}{|\Omega|}\sum_{(p,q)\in\Omega} w_{p,q}^i \cdot \left\|\mathcal{G}(H_i)_{p,q} - \mathcal{G}(S)_{p,q}\right\|_2^2, \tag{6}$$

where S denotes the feature map extracted from the SAM 2 encoder, $\Omega = \{(p,q) \mid p,q \in [1,N]\}$ represents all spatial location pairs, and the text-guided soft weight is defined as $w_{p,q}^i = \alpha + (1-\alpha)\cdot(M_{\text{text}}^i(p) + M_{\text{text}}^i(q))/2$. The margin

$\alpha = 0.1$ is used to maintain basic spatial consistency to prevent complete collapse of the background topology. Unlike mask-level distillation, LGCD does not treat SAM 2 as a segmentation oracle or copy its masks as hard pseudo-labels; instead, it transfers text-filtered patch-to-patch affinities as soft relational regularization. The total distillation loss is dynamically aggregated across all cascade stages to provide continuous depth supervision:

$$\mathcal{L}_{\text{distill}} = \frac{1}{K} \sum_{k=1}^{K} \mathcal{L}_{\text{distill}}^{(k)}, \tag{7}$$

where K denotes the total number of cascaded stages.

4.2 Perspective-Spatial Contrastive Learning (PSCL)

As discussed earlier, pre-trained VLMs inherently suffer from object-centric semantic bias. They tend to take a lazy shortcut, focusing solely on the **primary subject** (*e.g.*, "vehicle") while ignoring critical **spatial modifiers** (*e.g.*, "left"). To explicitly shatter this spurious correlation, we propose the PSCL paradigm, which establishes a complementary, cross-anchored contrastive space. Specifically, we dismantle this shortcut from two perspectives:

Perceptive Hard Negatives. We utilize the GT mask to filter out the target region, strictly isolating the background Ω_{bg}. Within Ω_{bg}, pixels exhibiting the highest similarities to the text F_t are considered deceptive distractors, as they likely match the primary subject but violate spatial constraints. Therefore, we construct the hard negative set $\mathcal{V}^- = \{\mathbf{v}^- \in \Omega_{bg} \mid \text{Top-}K\langle \mathbf{v}^-, F_t \rangle\}$ by dynamically mining these top-K embeddings.

Spatial Counterfactual Negatives. Beyond visual discrimination, we aim to explicitly endow the model with topological reasoning capabilities. To achieve this without intractable geometric modeling, we synthesize spatial counterfactual texts t^-. For absolute positional descriptions, we invert the spatial indicators (*e.g.*, "top" $\rightarrow$ "bottom"). Crucially, for intricate relative relationships, we systematically swap the subject and the object (*e.g.*, "A golf field is on the left of the green airport" $\rightarrow$ "The green airport is on the left of a golf field") to generate logical contradictions. By leveraging a lightweight LLM (*e.g.*, Qwen2.5-7B-Instruct in our experiments) for offline preprocessing, we generate these perturbations to force the model to decode syntactic structures, thereby mitigating the reliance on superficial keyword correlations.

Dual-Constraint Objective. We integrate these perspective and spatial constraints into a unified asymmetric InfoNCE objective, as illustrated in Fig. 2. Given the positive visual prototype $\mathbf{v}^+$ (aggregated from the target mask prediction) and the original text t, the PSCL loss is formulated as:

$$\mathcal{L}_{PSCL} = -\log \frac{\exp(\langle \mathbf{v}^+, F_t \rangle / \tau)}{\exp(\langle \mathbf{v}^+, F_t \rangle / \tau) + \sum_{\mathbf{v}^- \in \mathcal{V}^-} \exp(\langle \mathbf{v}^-, F_t \rangle / \tau) + \eta \exp(\langle \mathbf{v}^+, F_{t^-} \rangle / \tau)}, \tag{8}$$

where τ is the temperature hyperparameter, $\langle \cdot, \cdot \rangle$ denotes cosine similarity, and η is a scaling coefficient balancing the gradient contribution of the spatial logic

constraint. Following common practice in previous contrastive learning [49, 60], we set $\tau = 0.07$. The scaling factor is set to $\eta = 2$ by default.

4.3 Overall Training Objective

After the cascaded refinement in the LGCD module, the terminal visual feature H_{final} (i.e., the output of the last cascade stage) and the linguistic embedding F_t are used to generate prompts. Subsequently, these prompts, alongside the image features extracted from the SAM 2 encoder, are fed into the SAM mask decoder to predict the segmentation mask $\hat{M} \in \mathbb{R}^{H \times W}$. Given the corresponding ground truth mask $Y \in \{0, 1\}^{H \times W}$, the overall training objective of CROSS is:

$$\mathcal{L}_{total} = \lambda_{ce}\mathcal{L}_{\text{ce}}(\hat{M}, Y) + \lambda_{dice}\mathcal{L}_{\text{dice}}(\hat{M}, Y) + \lambda_1\mathcal{L}_{\text{distill}} + \lambda_2\mathcal{L}_{PSCL}, \quad (9)$$

where $\mathcal{L}_{\text{ce}}$ and $\mathcal{L}_{\text{dice}}$ denote the cross-entropy loss and DICE loss [45], respectively. $\mathcal{L}_{\text{distill}}$ and $\mathcal{L}_{PSCL}$ represent the cascaded distillation loss and the perspective-spatial contrastive learning loss, respectively.

5 Experiments

5.1 Experimental Settings

Datasets. We conduct comprehensive experiments on two standard RRSIS benchmarks: As the pioneering dataset in this domain, RefSegRS consists of 512×512 resolution images, divided into 2,172, 413, and 1,817 samples for training, validation, and testing, respectively. To further assess scalability, we utilize RRSIS-D, a large-scale benchmark containing 12,181 training, 1,740 validation, and 3,481 test samples, with a higher resolution of 800×800.
Evaluation Metrics Consistent with prior work in referring image segmentation [37, 71], we evaluate our method using three standard metrics: cumulative Intersection over Union (cIoU), generalized Intersection over Union (gIoU), and Precision at specific IoU thresholds (Pr@X, where $X \in \{0.5, 0.6, \ldots, 0.9\}$).

5.2 Implementation Details

CROSS integrates the robust cross-modal alignment of SigLIP 2 [55] with the high-fidelity segmentation capabilities of SAM 2 [50] into a unified framework. The SAM 2 variant employed is 'sam2.1-hiera-large'[5], and the SigLIP 2 variant utilized is 'siglip2-so400m-patch16-512'[6]. Following the input requirements of SAM 2 and SigLIP 2, training images were resized to 512×512 and 1024×1024, respectively, without any data augmentation. Consistent with the PEFT settings in RSRefSeg 2 [2], we apply LoRA ($r = 16$) to the encoders of both SigLIP 2 and SAM 2. The trainable parameters consist of the newly introduced low-rank

[5] https://huggingface.co/facebook/sam2.1-hiera-large
[6] https://huggingface.co/google/siglip2-so400m-patch16-512

Table 1: Performance comparison across various evaluation metrics on the RefSegRS test dataset.

Method	Publication	Pr@0.5	Pr@0.6	Pr@0.7	Pr@0.8	Pr@0.9	cIoU	gIoU
BRINet [10]	CVPR'20	20.72	14.26	9.87	2.98	1.14	58.22	31.51
LSCM [13]	ECCV'20	31.54	20.41	9.51	5.29	0.84	61.27	35.54
CMPC [12]	CVPR'20	32.36	14.14	6.55	1.76	0.22	55.39	40.63
CMPC+ [36]	TPAMI'21	49.19	28.31	15.31	8.12	2.55	66.53	43.65
CRIS [62]	CVPR'22	35.77	24.11	14.36	6.38	1.21	65.87	43.26
LAVT [70]	CVPR'22	51.84	30.27	17.34	9.52	2.09	71.86	47.40
CARIS [38]	ACM MM'23	45.40	27.19	15.08	8.87	1.98	69.74	42.66
RIS-DMMI [9]	CVPR'23	63.89	44.30	19.81	6.49	1.00	68.58	52.15
CrossVLT [6]	TMM'23	71.16	58.28	34.51	16.35	5.06	77.44	58.84
LGCE [71]	TGRS'24	73.75	61.14	39.46	16.02	5.45	76.81	59.96
DANet [48]	ACM MM'24	76.61	64.59	42.72	18.29	8.04	79.53	62.14
RMSIN [37]	CVPR'24	79.20	65.99	42.98	16.51	3.25	75.72	62.58
FIANet [20]	TGRS'24	84.09	77.05	61.86	33.41	7.10	78.32	68.67
SBANet [27]	ISPRS'25	77.02	-	44.15	-	8.97	79.86	62.73
SegEarth-R1 [26]	Arxiv'25	86.30	79.53	69.57	48.87	10.73	79.00	72.45
RS2-SAM 2 [51]	AAAI'26	84.31	79.42	70.89	55.70	21.19	80.87	73.90
RSRefSeg-2 [2]	TGRS'26	88.22	82.99	73.97	60.92	34.40	81.24	77.39
Ours	-	**88.61**	**83.98**	**76.39**	**64.61**	**41.32**	**83.25**	**79.51**

modules, the LGCD module, and the SAM decoder, while all other foundation model backbone weights remain frozen. This configuration results in only **7.2%** of the total parameters being updated. For our cascaded distillation, the soft mask weight is set to $\alpha = 0.1$. During the contrastive process, we utilize $K = 8$ visual negative samples with a spatial penalty weight of $\eta = 2$. The balancing coefficients for the objective function are assigned as $\lambda_{ce} = \lambda_{dice} = 5.0$ following [2], while the weights for our auxiliary components are set to $\lambda_1 = 0.5$ and $\lambda_2 = 0.2$ via empirical hyperparameter tuning. Training is performed using the AdamW optimizer with a peak learning rate of 1×10^{-4} and a batch size of 8 over 300 epochs. We utilize BF16 precision and the DeepSpeed ZeRO-2 framework on 8 NVIDIA RTX PRO 6000 GPUs.

5.3 Comparison with State-of-the-Art Methods

RefSegRS Dataset. As detailed in Tab. 1, our method establishes a new state-of-the-art across all evaluation metrics, achieving a **cIoU of 83.25%** and a **gIoU of 79.51%**. Notably, our framework consistently outperforms the strongest competitor, RSRefSeg 2, across the entire precision spectrum. While maintaining a steady lead at looser overlap requirements (+**0.39%** at Pr@0.5 and +**0.99%** at Pr@0.6), the performance gap widens substantially under more rigorous criteria. Specifically, our method surpasses RSRefSeg 2 by +**3.69% at Pr@0.8** and a remarkable +**6.92% at Pr@0.9**. This performance trajectory demonstrates that our cascaded distillation and contrastive learning effectively mitigate representational drift, enabling the generation of high-fidelity masks that strictly adhere to target boundaries even under the most stringent overlap constraints.

Table 2: Performance comparison across various evaluation metrics on the RRSIS-D test dataset.

Method	Publication	Pr@0.5	Pr@0.6	Pr@0.7	Pr@0.8	Pr@0.9	cIoU	gIoU
BRINet [10]	CVPR'20	56.90	48.77	39.12	27.03	8.73	69.88	49.65
CMPC+ [36]	TPAMI'21	57.65	47.51	36.97	24.33	7.78	68.64	50.24
LAVT [70]	CVPR'22	69.52	63.63	53.29	41.60	24.94	77.19	61.04
RIS-DMMI [9]	CVPR'23	68.74	60.96	50.33	38.38	21.63	76.20	60.12
CrossVLT [6]	TMM'23	70.38	63.83	52.86	42.11	25.02	76.32	61.00
LGCE [71]	TGRS'24	67.65	61.53	51.45	39.62	23.33	76.34	59.37
EVF-SAM [77]	Arxiv'24	72.16	66.50	56.59	43.92	25.48	76.77	62.75
FIANet [20]	TGRS'24	74.46	66.96	56.31	42.83	24.13	76.91	64.01
RMSIN [37]	CVPR'24	74.26	67.25	55.93	42.55	24.53	77.79	64.20
CADFormer [34]	JSTARS'25	74.20	67.62	55.59	42.37	23.59	77.26	63.77
LSCF [43]	TGRS'25	74.30	67.69	56.32	43.08	25.67	77.42	64.25
RSRefSeg-l [3]	IGARSS'25	74.49	68.33	58.73	48.50	30.80	77.24	64.67
SegEarth-R1 [26]	Arxiv'25	76.96	-	-	-	-	78.01	66.40
RS2-SAM 2 [51]	AAAI'26	77.56	72.34	61.76	47.92	29.73	78.99	66.72
RSRefSeg-2 [3]	TGRS'26	80.23	**75.78**	**65.41**	50.65	31.05	79.45	**69.17**
Ours	-	**81.24**	74.56	64.80	**51.74**	**32.82**	**79.89**	68.92

Table 3: Ablation Studies on RefSegRS and RRSIS-D. (a) Analysis of the proposed **LGCD** (CRE, TGD) and **PSCL** ($\mathcal{L}_{\text{PSCL}}$) modules. (b) Impact of different negative sample types in PSCL. All results in cIoU (%).

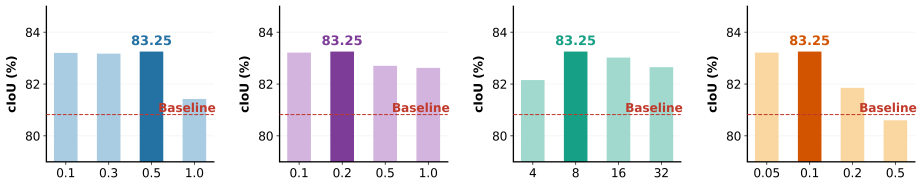
LGCD Module		PSCL	Benchmarks	
CRE	TGD	$\mathcal{L}_{\text{PSCL}}$	RefSegRS	RRSIS-D
✗	✗	✗	80.82	77.80
✓	✗	✗	82.55	78.63
✓	✓	✗	83.15	79.02
✗	✗	✓	82.11	78.12
✓	✓	✓	**83.25**	**79.89**

(a) Main Components

PSCL Negatives		RefSegRS	RRSIS-D
Perspective	Spatial		
✗	✗	83.15	79.02
✓	✗	83.21	79.34
✗	✓	83.20	78.70
✓	✓	**83.25**	**79.89**

(b) Formulation of $\mathcal{L}_{\text{PSCL}}$

RRSIS-D Dataset. Tab. 2 demonstrates that CROSS achieves state-of-the-art performance on RRSIS-D, leading in cIoU (**79.89%**) and Pr@0.5 (**81.24%**). Notably, our framework maintains substantial gains under the most stringent evaluation criteria, outperforming RSRefSeg 2 by **+1.77%** at the Pr@0.9 threshold. This performance delta at high precision intervals underscores the efficacy of our Filter-Refine-Verify paradigm; while weakly-coupled baselines suffer from logical drifting, CROSS ensures logical soundness by anchoring masks to the complete linguistic context via LGCD. The marginal gIoU deficit is likely due to the extreme scale diversity of dense small objects in RRSIS-D, where our model prioritizes global logical grounding and high-precision localization over heuristic boundary alignment.



(a) Weight λ_1 **(b)** Weight λ_2 **(c)** Candidates K **(d)** Weight α

Fig. 3: Sensitivity analysis on RefSegRS with respect to cIoU. We evaluate the impact of **(a)** λ_1, **(b)** λ_2, **(c)** top-k candidates K, and **(d)** soft weight α. The results demonstrate that CROSS maintains stable and superior performance across a broad range of hyperparameter settings.

5.4 Ablation Studies

Effectiveness of Main Components. We conduct a series of component-wise ablations on RefSegRS and RRSIS-D to verify the contribution of each module, as reported in Tab. 3 (a). The baseline (SigLIP 2 + SAM 2) achieves **80.82%** cIoU on RefSegRS. Individually, we first conduct experiments by integrating Cascaded Representation Extraction (**CRE**), which yields a significant gain (+**1.73%**), validating that transferring SAM's geometric priors is essential for dense prediction. Meanwhile, Text-Guided Relational Distillation (**TGD**) leverages textual cues to mitigate background clutter within intermediate layers, effectively filtering out irrelevant noise across different stages. Jointly, these modules reach **83.15%**, demonstrating that bridging the semantic-geometric gap is vital for robust performance. Finally, the inclusion of $\mathcal{L}_{\text{PSCL}}$ yields the peak result **83.25%** on RefSegRS and **79.89%** on RRSIS-D, as it provides the necessary discriminative power to suppress illusory binding against visually similar distractors.

Investigation of Negative Samples in Contrastive Object. We investigate the negative formulations in Tab. 3 (b). Relying exclusively on spatial contrast leads to a performance decrease (**-0.05%** on RefSegRS), as spatial cues without visual grounding cause ambiguity in cluttered remote sensing scenes. Conversely, incorporating visual negatives improves the result to **83.21%**. Ultimately, combining both yields the peak **83.25%** on RefSegRS and **79.89%** on RRSIS-D, ensuring discriminability in both perspective and spatial dimensions.

Hyperparameter Analysis. Firstly, we conduct a comprehensive sensitivity analysis of four key hyperparameters in Fig. 3. Crucially, across comprehensive tested ranges, CROSS consistently maintains a superior performance margin over the baseline (**80.82%**), as indicated by the red dashed lines. The optimal weights are achieved at

Layer Indices	RefSegRS	RRSIS-D
$\{18, 24, 27\}$	83.19	79.71
$\{5, 15, 27\}$	82.76	79.42
$\{25, 27, 27\}$	83.14	79.50
$\{9, 18, 27\}$	**83.25**	**79.89**

Table 5: Ablation on layer selection (cIoU %).

$\lambda_1 = 0.5$ and $\lambda_2 = 0.2$; increasing λ_1 beyond 0.5 tends to cause gradient dominance that destabilizes the joint optimization. For the hard negative set, $K = 8$ performs best by maintaining a balanced contribution between perceptive and spatial negatives. Regarding the text-guided background soft filtering, $\alpha = 0.1$ is found to be ideal, whereas a larger value (*e.g.*, 0.5) leads to excessive suppression of potential target features. These stable trends across both datasets validate our empirical configurations and ensure the model's robustness. **Secondly,** we investigate the choice of intermediate layers, adopting $\{9, 18, 27\}$ as the final configuration, as shown in Tab. 5. Notably, we do not emphasize the strict optimality of these specific indices, as our primary goal is simply to ensure the extraction of features from diverse intermediate stages. The marginal performance degradation observed only with excessively shallow layers (*e.g.*, $\{5, 15, 27\}$) confirms that the framework remains robust to layer selection within a reasonable range.

5.5 Further Analysis and Discussion

•**Parameter Overhead Analysis.** As summarized in Tab. 6, CROSS updates **106.09 M** parameters, representing a **7.29%** trainable ratio. Compared to the baseline and RSRefSeg 2, our framework introduces a parameter increase of **1.23%** and **0.57%**, respectively, relative to the total 1.4B+ backbone scale. This increase is primarily attributed to the integrated cascaded distillation module. Given the resulting performance gains across multiple evaluation metrics, we consider this additional parameter cost to be acceptable.

Table 6: Comparative analysis of parameter scales. The total and trainable parameters are evaluated across the baseline (SAM 2 + SigLIP 2), RSRefSeg 2, and our CROSS.

Method	Total Params (B)	Trainable Params (M)	Ratio (%)
Baseline (SAM 2+SigLIP 2)	1.438 B	88.41 M	6.15%
RSRefSeg 2	1.447 B	97.87 M	6.76%
CROSS	**1.455 B**	**106.09 M**	**7.29%**

•**Robust Spatial Grounding.** Fig. 4 probes the grounding logic of CROSS against SOTA method RSRefSeg 2 in complex scenarios. For the nested prompt ***"The airplane is on the right of the airplane on the left"***, RSRefSeg 2 fails to parse the full relational context, drifting toward the partial clause ***"on the right"***. In addition, in the scenario ***"A ground track field is in the gray large stadium"***, where two identical track fields are present, the baseline erroneously segments both instances. This oversight indicates a failure to process the spatial qualifier ***"is in the gray large stadium"***, effectively reducing the logic-driven task to a simple category-level search. CROSS resolves these via **PSCL**, which enforces strict logical alignment through contrastive verification.

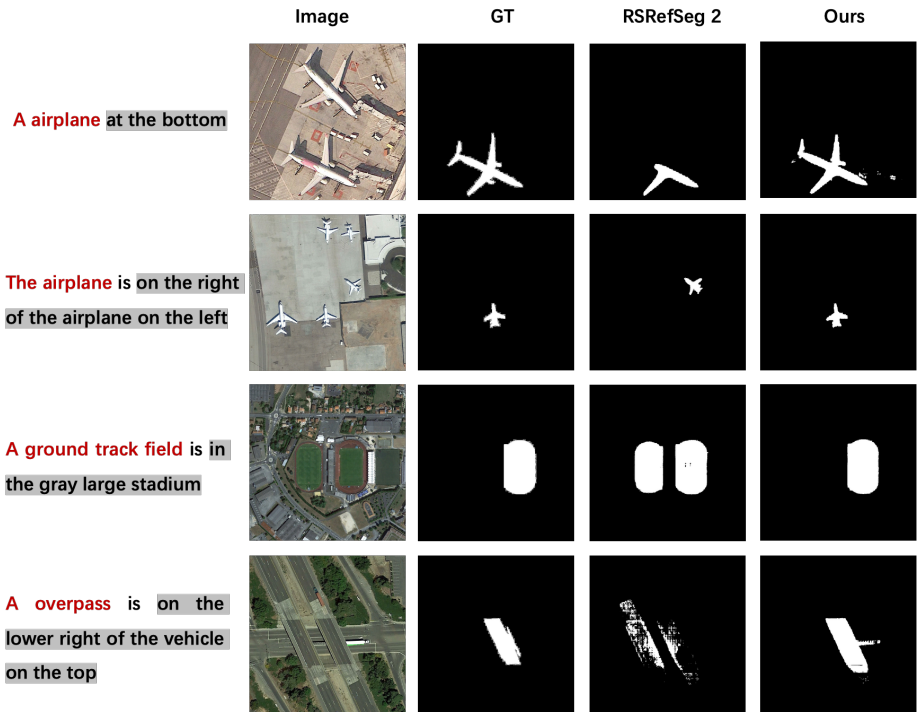


Fig. 4: Visualization result on RRSIS-D. The targets and spatial descriptions are highlighted in **red** and **gray**, respectively. Compared to RSRefSeg 2, our method achieves more accurate spatial referring and precise boundaries.

By penalizing "logical collapse," CROSS effectively filters distractors and ensures spatial constraint is actively utilized for disambiguation.

•**Effects of Structural Distillation.** As shown in Fig. 5, CROSS yields more concentrated dense prompt heatmaps compared to RSRefSeg 2. This provides direct evidence that our distillation successfully injects SAM's structural priors into the VLM, enhancing fine-grained localization precision.

•**Structural Consistency Analysis.** To quantitatively assess the internal representation dynamics, we employ Centered Kernel Alignment (CKA) [17] to visualize layer-wise feature similarities. Fig. 6 compares our CROSS with the baseline. The **red box** reveals stable alignment between the final layer and preceding deep hierarchies, while the blue box illustrates notably smoother transitions between adjacent layers. These phenomena stem directly from our cascaded distillation design. By injecting a shared SAM-derived geometric affinity matrix across multiple layers, we establish a consistent structural prior that regularizes the representational trajectory. Crucially, to strictly prevent this shared constraint from inducing detrimental layer homogenization (i.e., collapsing hierarchical feature diversity), the distillation is dynamically modulated by a *layer-specific* soft mask (Eq. 4). By guiding the shared geometric anchor through this

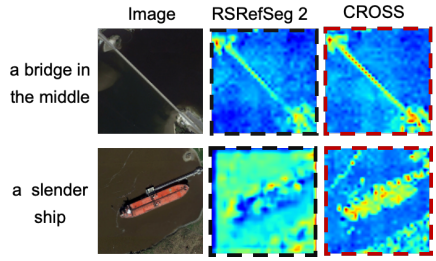


Fig. 5: Visualization of dense prompt heatmaps. Compared to the baseline RSRefSeg 2, the heatmaps of CROSS are significantly more concentrated.

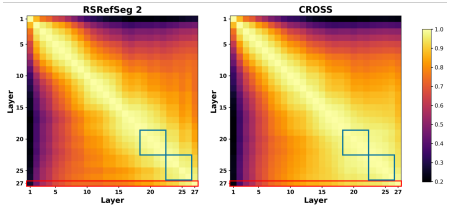


Fig. 6: Internal representation analysis via CKA [17]. Comparison of layer-wise similarity heatmaps. The **red box** highlights stable alignment, while the blue box illustrates smoother transitions.

layer-adaptive masking, our design ensures that the network internalizes robust structural awareness while strictly preserving its capacity for adaptive feature evolution.

6 Conclusion

In this paper, we presented CROSS, a novel framework for Referring Remote Sensing Image Segmentation (RRSIS). Our research identifies and addresses two fundamental bottlenecks in existing foundation-model adaptations: the semantic-geometric gap arising from weakly coupled architectures, and the illusory spatial-semantic binding inherent in pre-trained VLMs. By introducing Linguistic-Guided Cascaded Distillation (LGCD) for structural prior distillation and Perspective-Spatial Contrastive Learning (PSCL) for contrastive refinement, CROSS effectively restores representational smoothness and enhances geometric fidelity. Extensive experiments on RRSIS benchmarks demonstrate that CROSS achieves state-of-the-art performance across comprehensive evaluation metrics, particularly under stringent precision requirements, and demonstrates a superior understanding of spatial referring.

Acknowledgements

This work was supported by the Hong Kong RGC under Grant 21218026, and the City University of Hong Kong under Grants 9382010 and 7020171.

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)

2. Chen, K., Liu, C., Chen, B., Zhang, J., Zou, Z., Shi, Z.: Rsrefseg 2: Decoupling referring remote sensing image segmentation with foundation models. IEEE Transactions on Geoscience and Remote Sensing **64**, 1–20 (2026). https://doi.org/10.1109/TGRS.2025.3647535
3. Chen, K., Zhang, J., Liu, C., Zou, Z., Shi, Z.: Rsrefseg: Referring remote sensing image segmentation with foundation models. In: IGARSS 2025-2025 IEEE International Geoscience and Remote Sensing Symposium. pp. 1070–1074. IEEE (2025)
4. Cheng, Y., Li, L., Xu, Y., Li, X., Yang, Z., Wang, W., Yang, Y.: Segment and track anything. arXiv preprint arXiv:2305.06558 (2023)
5. Cho, J.H., Madotto, A., Mavroudi, E., Afouras, T., Nagarajan, T., Maaz, M., Song, Y., Ma, T., Hu, S., Jain, S., et al.: Perceptionlm: Open-access data and models for detailed visual understanding. Advances in Neural Information Processing Systems **38**, 23475–23537 (2026)
6. Cho, Y., Yu, H., Kang, S.J.: Cross-aware early fusion with stage-divided vision and language transformer encoders for referring image segmentation. IEEE Transactions on Multimedia **26**, 5823–5833 (2023)
7. Gao, J., Zhang, D., Wang, F., Ning, L., Zhao, Z., Li, X.: Combining sam with limited data for change detection in remote sensing. IEEE Transactions on Geoscience and Remote Sensing (2025)
8. Guan, R., Liu, T., Tu, W., Tang, C., Luo, W., Liu, X.: Sampling enhanced contrastive multi-view remote sensing data clustering with long-short range information mining. IEEE Transactions on Knowledge and Data Engineering (2025)
9. Hu, Y., Wang, Q., Shao, W., Xie, E., Li, Z., Han, J., Luo, P.: Beyond one-to-one: Rethinking the referring image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4067–4077 (2023)
10. Hu, Z., Feng, G., Sun, J., Zhang, L., Lu, H.: Bi-directional relationship inferring network for referring image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4424–4433 (2020)
11. Huang, J., Xu, Z., Liu, T., Liu, Y., Han, H., Yuan, K., Li, X.: Densely connected parameter-efficient tuning for referring image segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 3653–3661 (2025)
12. Huang, S., Hui, T., Liu, S., Li, G., Wei, Y., Han, J., Liu, L., Li, B.: Referring image segmentation via cross-modal progressive comprehension. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10488–10497 (2020)
13. Hui, T., Liu, S., Huang, S., Li, G., Yu, S., Zhang, F., Han, J.: Linguistic structure guided context modeling for referring image segmentation. In: European conference on computer vision. pp. 59–75. Springer (2020)
14. Ito, K.: Feature design for bridging sam and clip toward referring image segmentation. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 8368–8378. IEEE (2025)
15. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
16. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)

17. Kornblith, S., Norouzi, M., Lee, H., Hinton, G.: Similarity of neural network representations revisited. In: International conference on machine learning. pp. 3519–3529. PMlR (2019)
18. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9579–9589 (2024)
19. Lan, M., Rong, F., Jiao, H., Gao, Z., Zhang, L.: Language query-based transformer with multiscale cross-modal alignment for visual grounding on remote sensing images. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–13 (2024)
20. Lei, S., Xiao, X., Zhang, T., Li, H.C., Shi, Z., Zhu, Q.: Exploring fine-grained image-text alignment for referring remote sensing image segmentation. IEEE Transactions on Geoscience and Remote Sensing **63**, 1–11 (2025). https://doi.org/10.1109/TGRS.2024.3522293
21. Li, B., Dong, H., Zhang, D., Zhao, Z., Sun, H., Gao, J.: Exploring efficient open-vocabulary segmentation in the remote sensing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 5982–5991 (2026)
22. Li, B., Huo, T., Dong, H., Zhang, D., Zhao, Z., Gao, J., Li, X.: Towards realistic open-vocabulary remote sensing segmentation: Benchmark and baseline. arXiv preprint arXiv:2604.15652 (2026)
23. Li, B., Zhang, D., Huo, T., Zhao, Z., Gao, J., Li, X.: An open-source benchmark and baseline for multi-temporal referring segmentation. arXiv preprint arXiv:2606.00987 (2026)
24. Li, B., Zhang, D., Zhao, Z., Gao, J., Li, X.: Stitchfusion: Weaving any visual modalities to enhance multimodal semantic segmentation. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 1308–1317 (2025)
25. Li, B., Zhang, D., Zhao, Z., Gao, J., Li, X.: U3m: Unbiased multiscale modal fusion model for multimodal semantic segmentation. Pattern Recognition **168**, 111801 (2025)
26. Li, K., Xin, Z., Pang, L., Pang, C., Deng, Y., Yao, J., Xia, G., Meng, D., Wang, Z., Cao, X.: Segearth-r1: Geospatial pixel reasoning via large language model. arXiv preprint arXiv:2504.09644 (2025)
27. Li, K., Vosselman, G., Yang, M.Y.: Scale-wise bidirectional alignment network for referring remote sensing image segmentation. ISPRS Journal of Photogrammetry and Remote Sensing **226**, 350–363 (2025)
28. Li, X., Wen, C., Hu, Y., Zhou, N.: Rs-clip: Zero shot remote sensing scene classification via contrastive vision-language supervision. International Journal of Applied Earth Observation and Geoinformation **124**, 103497 (2023)
29. Li, Y., Li, Z., Zeng, Q., Hou, Q., Cheng, M.M.: Cascade-clip: Cascaded vision-language embeddings alignment for zero-shot semantic segmentation. arXiv preprint arXiv:2406.00670 (2024)
30. Lin, B., Li, Z., Cheng, X., Niu, Y., Ye, Y., He, X., Yuan, S., Yu, W., Wang, S., Ge, Y., et al.: Uniworld-v1: High-resolution semantic encoders for unified visual understanding and generation. arXiv preprint arXiv:2506.03147 (2025)
31. Liu, C., Lin, Z., Shen, X., Yang, J., Lu, X., Yuille, A.: Recurrent multimodal interaction for referring image segmentation. In: Proceedings of the IEEE international conference on computer vision. pp. 1271–1280 (2017)
32. Liu, F., Chen, D., Guan, Z., Zhou, X., Zhu, J., Ye, Q., Fu, L., Zhou, J.: Remoteclip: A vision language foundation model for remote sensing. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–16 (2024)
33. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)

34. Liu, M., Jiang, X., Zhang, X.: 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 (2025)
35. Liu, R., Luo, T., Huang, S., Wu, Y., Jiang, Z., Zhang, H.: Crossmatch: Cross-view matching for semi-supervised remote sensing image segmentation. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–15 (2024)
36. Liu, S., Hui, T., Huang, S., Wei, Y., Li, B., Li, G.: Cross-modal progressive comprehension for referring segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(9), 4761–4775 (2021)
37. Liu, S., Ma, Y., Zhang, X., Wang, H., Ji, J., Sun, X., Ji, R.: Rotated multi-scale interaction network for referring remote sensing image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26658–26668 (2024)
38. Liu, S.A., Zhang, Y., Qiu, Z., Xie, H., Zhang, Y., Yao, T.: Caris: Context-aware referring image segmentation. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 779–788 (2023)
39. Liu, Y., Peng, B., Zhong, Z., Yue, Z., Lu, F., Yu, B., Jia, J.: Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. arXiv preprint arXiv:2503.06520 (2025)
40. Luo, T., Du, M., Shi, J., Chen, X., Zhao, B., Huang, S.: Contextuality helps representation learning for generalized category discovery. In: 2024 IEEE International Conference on Image Processing (ICIP). pp. 687–693. IEEE (2024)
41. Luo, T., Liu, Y., Liu, Y., Zhang, A., Wang, X., Zhan, Y., Tang, C., Liu, L., Chen, Z.: Dig-face: De-biased learning for generalized facial expression category discovery. arXiv preprint arXiv:2409.20098 (2024)
42. Luo, T., Liu, Y., Xu, Y., Liu, R., Wang, X., Zeng, H., Huang, S., Zhang, H.: Stroke-based perception: Discover novel oracle characters. IEEE Transactions on Multimedia (2026)
43. Ma, Q., Li, L., Lu, X., Jiao, L., Liu, F., Ma, W., Liu, X., Sun, L.: Lscf: Long-term semantic-guidance convformer for referring remote sensing image segmentation. IEEE Transactions on Geoscience and Remote Sensing (2025)
44. Ma, X., Lian, R., Wu, Z., Guan, R., Hong, T., Zhao, M., Ma, M., Nie, J., Du, Z., Song, S., et al.: A novel scene coupling semantic mask network for remote sensing image segmentation. ISPRS Journal of Photogrammetry and Remote Sensing **221**, 44–63 (2025)
45. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: Proceedings of the International Conference on 3D Vision (3DV). pp. 565–571 (2016)
46. Osco, L.P., Wu, Q., De Lemos, E.L., Gonçalves, W.N., Ramos, A.P.M., Li, J., Junior, J.M.: The segment anything model (sam) for remote sensing applications: From zero to one shot. International Journal of Applied Earth Observation and Geoinformation **124**, 103540 (2023)
47. Ou, R., Hu, Y., Zhang, F., Chen, J., Liu, Y.: Geopix: A multimodal large language model for pixel-level image understanding in remote sensing. IEEE Geoscience and Remote Sensing Magazine (2025)
48. Pan, Y., Sun, R., Wang, Y., Zhang, T., Zhang, Y.: Rethinking the implicit optimization paradigm with dual alignments for referring remote sensing image segmentation. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 2031–2040 (2024)

49. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
50. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
51. Rong, F., Lan, M., Zhang, Q., Zhang, L.: Customized sam 2 for referring remote sensing image segmentation. arXiv e-prints pp. arXiv–2503 (2025)
52. Shabbir, A., Zumri, M., Bennamoun, M., Khan, F.S., Khan, S.: Geopixel: Pixel grounding large multimodal model in remote sensing. arXiv preprint arXiv:2501.13925 (2025)
53. Sun, Y., Feng, S., Li, X., Ye, Y., Kang, J., Huang, X.: Visual grounding in remote sensing images. In: Proceedings of the 30th ACM International conference on Multimedia. pp. 404–412 (2022)
54. Sun, Y., Dong, Z., Jiang, H., Gu, Y., Liu, T.: Crobim-u: Uncertainty-driven referring remote sensing image segmentation. IEEE Transactions on Geoscience and Remote Sensing (2026)
55. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
56. Wang, C., Ji, Y., Meng, Y., Zhang, Y., Zhu, Y.: Rs-ood: A vision-language augmented framework for out-of-distribution detection in remote sensing. arXiv preprint arXiv:2509.02273 (2025)
57. Wang, C., Ji, Y., Meng, Y., Zhang, Y., Zhu, Y.: Sopseg: Prompt-based small object instance segmentation in remote sensing imagery. arXiv preprint arXiv:2509.03002 (2025)
58. Wang, C., Xi, Z., Liu, D., Feng, Y., Deng, Y., Li, K., Chen, J., Chen, J., Meng, Y.: Pcp: A prompt-based cartographic-level polygonal vector extraction framework for remote sensing images. IEEE Transactions on Geoscience and Remote Sensing (2025)
59. Wang, D., Zhang, J., Du, B., Xu, M., Liu, L., Tao, D., Zhang, L.: Samrs: Scaling-up remote sensing segmentation dataset with segment anything model. Advances in Neural Information Processing Systems **36**, 8815–8827 (2023)
60. Wang, F., Liu, H.: Understanding the behaviour of contrastive loss. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2495–2504 (2021)
61. Wang, J., Ke, L.: Llm-seg: Bridging image segmentation and large language model reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1765–1774 (2024)
62. Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M., Liu, T.: Cris: Clip-driven referring image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11686–11695 (2022)
63. Wang, Z., Prabha, R., Huang, T., Wu, J., Rajagopal, R.: Skyscript: A large and semantically diverse vision-language dataset for remote sensing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 5805–5813 (2024)
64. Xiong, Y., Varadarajan, B., Wu, L., Xiang, X., Xiao, F., Zhu, C., Dai, X., Wang, D., Sun, F., Iandola, F., et al.: Efficientsam: Leveraged masked image pretraining for efficient segment anything. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16111–16121 (2024)

65. Xu, J., Xu, L., Yang, Y., Li, X., Wang, F., Xie, Y., Huang, Y.J., Li, Y.: u-llava: Unifying multi-modal tasks via large language model. arXiv preprint arXiv:2311.05348 (2023)
66. Xu, Y., Guo, C., Shuai, Y., Ni, J.: Relational retrieval: Leveraging known-novel interactions for generalized category discovery. In: Proceedings of the 2026 International Conference on Multimedia Retrieval. pp. 410–414 (2026)
67. Yan, Z., Li, J., Li, X., Zhou, R., Zhang, W., Feng, Y., Diao, W., Fu, K., Sun, X.: Ringmo-sam: A foundation model for segment anything in multimodal remote-sensing images. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–16 (2023)
68. Yang, K., Wang, Y., Luo, T.: Assignment-driven hash learning in a hyper-semantic space for on-the-fly category discovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11303–11312 (2026)
69. Yang, S., Xia, M., Li, G., Zhou, H.Y., Yu, Y.: Bottom-up shift and reasoning for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11266–11275 (2021)
70. Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., Torr, P.H.: Lavt: Language-aware vision transformer for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18155–18165 (2022)
71. Yuan, Z., Mou, L., Hua, Y., Zhu, X.X.: Rrsis: Referring remote sensing image segmentation. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–12 (2024)
72. Yue, W., Zhang, J., Hu, K., Xia, Y., Luo, J., Wang, Z.: Surgicalsam: Efficient class promptable surgical instrument segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 6890–6898 (2024)
73. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023)
74. Zhan, Y., Xiong, Z., Yuan, Y.: Rsvg: Exploring data and models for visual grounding on remote sensing data. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–13 (2023)
75. Zhang, W., Cai, M., Zhang, T., Zhuang, Y., Mao, X.: Earthgpt: A universal multimodal large language model for multisensor image comprehension in remote sensing domain. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–20 (2024)
76. Zhang, X., Zhang, H., Wang, G., Zhang, Q., Zhang, L., Du, B.: Uniuir: Considering underwater image restoration as an all-in-one learner. IEEE Transactions on Image Processing (2025)
77. Zhang, Y., Cheng, T., Hu, R., Liu, L., Liu, H., Ran, L., Chen, X., Liu, W., Wang, X.: Evf-sam: Early vision-language fusion for text-prompted segment anything model. arXiv preprint arXiv:2406.20076 (2024)
78. Zhang, Z., Li, J., Liang, Y., Yan, J., Xiao, Y., Su, X., Yuan, Q.: Ecrformer: An efficient cloud removal transformer with semantic-decoupled learning for multimodal satellite imagery. ISPRS Journal of Photogrammetry and Remote Sensing **237**, 323–338 (2026)
79. Zhang, Z., Yan, J., Liang, Y., Feng, J., He, H., Cao, L.: Multi-scale restoration of missing data in optical time-series images with masked spatial-temporal attention network. IEEE Transactions on Geoscience and Remote Sensing (2025)
80. Zhong, Z., Tang, Z., He, T., Fang, H., Yuan, C.: Convolution meets lora: Parameter efficient finetuning for segment anything model. arXiv preprint arXiv:2401.17868 (2024)