remote sensing

MDPI

Article

UniRS-Seg: Shared Sparse–Dense Query Decoding for Joint Semantic and Referring Remote Sensing Segmentation

Aya Taourirte and Hai Huan *

School of Artificial Intelligence, Nanjing University of Information Science and Technology, Nanjing 210044, China; 202551620018@nuist.edu.cn
* Correspondence: haihuan@nuist.edu.cn

Highlights

What are the main findings?

- UniRS-Seg unifies semantic and referring remote sensing segmentation in one adapted SAM ViT-B + SigLIP 2 model with a shared, query-conditioned mask decoder.
- It achieves the highest listed mIoU/gIoU on six benchmarks: 57.30% (LoveDA), 85.31% (Vaihingen), 88.24% (Potsdam), 78.18% (RefSegRS), 70.88% (RRSIS-D), and 72.91% (RISBench).

What are the implications of the main findings?

- A common query-conditioned decoder supports both task families; the joint-training comparison improves the reported mean metrics over the single-task controls.
- Joint task-balanced training with cross-modal consistency yields positive transfer between semantic and referring segmentation, supporting unified segmentation pipelines for Earth observation.

Abstract

Remote sensing semantic and referring segmentation share mask geometry but differ in whether a category vocabulary or a natural-language expression selects the output. UniRS-Seg bridges these interfaces through a shared sparse–dense query compiler and a hierarchical mask decoder, jointly trained on independently annotated datasets. Class names and referring expressions are projected from SigLIP 2 into the same decoder space; dense image–query activations supply spatial support, and sparse query tokens specify the requested region. An adapted SAM ViT-B pathway and parallel convolutional features recover local geometry, while learnable soft mask cross-attention refines both tasks and class-balanced mask-guided self-attention supports semantic parsing. UniRS-Seg achieves 57.30%, 85.31%, and 88.24% mIoU on LoveDA, Vaihingen, and Potsdam, and 78.18% and 70.88% gIoU on RefSegRS and RRSIS-D. These scores exceed the strongest listed baseline means by 0.68, 0.99, 1.18, 0.79, and 1.71 percentage points, respectively. Against separately trained semantic-only and referring-only versions of the shared architecture, joint training improves the four evaluated datasets by 0.45, 0.89, 1.40, and 0.97 points. An additional supervised RISBench evaluation achieves 72.91% gIoU and 74.85% cIoU, with the highest gIoU among the listed methods. On RefSegRS, adding learnable mask cross-attention to combined sparse and dense prompts increases gIoU by 0.86 points. The results support supervised dual-task transfer through a common query-conditioned decoder; the comparison uses published baseline results and does not establish unseen-domain or zero-shot generalization.

Check for updates

Academic Editor: Tania Stathaki

Received: 4 September 2026
Revised: 8 October 2026
Accepted: 20 October 2026
Published: 24 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Keywords: remote sensing; semantic segmentation; referring image segmentation; Segment Anything Model; SigLIP 2; vision-language model; parameter-efficient fine-tuning; unified segmentation

1. Introduction

Pixel-level remote sensing interpretation must recover both the category and spatial extent of geospatial regions. Semantic segmentation labels every pixel using a predefined land-cover vocabulary, whereas referring segmentation extracts the region selected by a natural-language expression. These interfaces are useful for complementary operations: mapping all buildings in a scene and isolating the building described relative to a road, for example. Both depend on geometry that is difficult in overhead imagery, including densely repeated small objects, large scale variation, irregular vegetation, low-contrast boundaries, and shadows [1]. Training two independent pipelines duplicates the adaptation and boundary-recovery machinery even though the outputs are masks of related scene content.

Recent semantic methods address global context and local detail with Transformer and state-space designs, including UNetFormer, RS3Mamba, UNetMamba, and PyramidMamba [2–5]. Attention-enhanced segmentation also uses cross-region and channel interactions to preserve small objects and boundaries [6]. In the semisupervised setting, APRW combines adaptive patch augmentation (APA) with class ranking weights to address perturbation diversity and imbalance [7]. These advances concern visual representation or supervision, but do not directly connect exhaustive land-cover prediction to expression-conditioned selection. Conversely, RefSegRS and RRSIS-D establish language-guided overhead segmentation benchmarks [8,9]; RMSIN emphasizes rotated multi-scale interaction, and FIANet improves fine-grained image–text alignment [9,10]. Their referring masks remain distinct from a full-scene semantic labeling interface.

Foundation models offer complementary priors for connecting the tasks. SAM supplies prompt-conditioned mask geometry, while adapters help bridge the natural-image/overhead-image gap [11–13]. RemoteCLIP and SigLIP 2 strengthen image–text representations, and SkySense supports broad Earth-observation interpretation [14–16]. RSRefSeg, RS2-SAM2, and RSRefSeg 2 combine vision–language localization with SAM-family segmentation [17–19]. Unification itself is also an active direction: UniGeoSeg extends instruction-driven geospatial segmentation across referring, interactive, and reasoning tasks [20]. Thus, neither combining foundation models nor pursuing several segmentation interfaces is by itself a sufficient novelty claim.

The question addressed here is more specific: can a class-vocabulary query bank and a free-form expression use the same sparse–dense interface and mask decoder, and does joint supervision improve both task families? A fixed semantic classifier is tied to one taxonomy, whereas an expression decoder must discriminate instances using attributes and relations. Sharing their weights can transfer category and contour evidence, but can also introduce conflicts between exhaustive class coverage and selective instance extraction. Independently annotated datasets further prevent assuming that every image has both forms of supervision. The architecture and training scheme must therefore connect compatible signals without inventing referring annotations from semantic labels.

We propose UniRS-Seg to address this gap. SigLIP 2 encodes dataset class names or a referring expression in a common query space. Query-to-patch similarities generate a dense spatial prompt, and activation-weighted visual pooling forms a sparse mask token. Both condition an adapted SAM ViT-B pathway with convolutional detail recovery and one hierarchical decoder. Learnable mask cross-attention uses the current soft prediction as

a spatial prior; semantic-mode class-balanced attention reduces majority-class dominance. Semantic masks support conservative class-union expressions during training, while genuine referring annotations supervise expression-specific selection. Task-balanced sampling keeps either annotation family from dominating. Class queries are cached for image-only semantic inference, and referring inference accepts an expression without manual point or box prompts. The scope is supervised segmentation over the evaluated taxonomies, rather than zero-shot recognition of unseen categories.

The contributions are:

- A shared sparse–dense query interface and dynamic mask decoder that support both full-scene semantic parsing and expression-conditioned remote sensing segmentation.
- Complementary geometry and language pathways, with soft mask attention, semantic class balancing, and hierarchical reconstruction for overhead imagery.
- Annotation-compatible joint training with task-balanced sampling, cross-modal alignment, and class-union consistency, evaluated against separate and single-task controls.
- Evaluation on three semantic and three referring benchmarks, including an additional supervised RISBench experiment, with component, prompt, confidence, augmentation, training-regime, and adaptation ablations, corruption analysis, and measured computational profiling.

2. Related Work

2.1. Remote Sensing Semantic Segmentation

Fully convolutional networks established end-to-end dense prediction by replacing fully connected layers with spatial classifiers [21]. U-Net popularized symmetric encoder–decoder processing and skip connections for recovering high-resolution structure [22]. Atrous convolution and pyramid pooling subsequently enlarged receptive fields without proportionally reducing output resolution [23]. These ideas are particularly relevant in remote sensing, where a single tile may contain compact cars, elongated roads, individual roofs, and broad vegetation regions. Transformer models introduced long-range token interaction into segmentation. Swin Transformer provided hierarchical features through shifted-window attention [24]. BANet combined convolutional and Transformer representations for very-fine-resolution urban imagery, DC-Swin introduced a Transformer scheme for fine-resolution segmentation, and UNetFormer combined an efficient encoder with a global–local Transformer decoder [2,25,26]. More recent state-space designs pursue global modeling with approximately linear sequence complexity. RS3Mamba, UNetMamba, UMFormer, PyramidMamba, and PPMamba explore different scan, fusion, and decoder strategies [3–5,27,28]. Attention-based alternatives address spatial and channel context explicitly. Tong et al. combine CrossFormer with attention-enhanced multi-scale processing for urban segmentation [6]. APRW uses adaptive patch augmentation and class ranking weights in semisupervised segmentation [7]; its unlabeled-data and pseudo-label setting differs from the supervised dual-task protocol used here. WASP-Mamba combines multi-scale state-space aggregation with frequency-aware reconstruction and reports results on all three semantic datasets used here [29].

Explicit category modeling provides complementary context. D2LS dynamically refines class embeddings through iterative image–dictionary cross-attention, addressing appearance ambiguity with input-conditioned semantic representations [30]. SA-DDL extends this direction with a spatially adaptive dictionary field, a Local Context Modulator, and a Region-Aware Interaction Decoder [31]. These methods focus on semantic segmentation and local category variation. In contrast, CMAttn uses soft prediction-weighted class prototypes within the semantic branch of a shared semantic/referring decoder; expression-guided inference retains LMCAM while bypassing semantic class bal-

ancing. This distinction separates the present task-unification study from dictionary-based semantic representation learning.

Despite these advances, almost all semantic systems retain a dataset-specific output classifier. Changing from LoveDA's seven labels to the ISPRS taxonomy requires replacing and retraining the classifier. By treating label names as text queries, UniRS-Seg makes the output vocabulary explicit and uses the same dynamic mask generator across taxonomies. This does not automatically make the system open-vocabulary: class performance is still learned from the training taxonomies, and evaluation on unseen categories would require a separate protocol. The benefit claimed here is architectural unification and vocabulary-conditioned decoding, not unmeasured zero-shot recognition.

2.2. Referring Image and Referring Remote Sensing Segmentation

Early referring segmentation models repeatedly fused recurrent language representations with convolutional image features. RRN refined predictions recurrently [32]; CMSA used cross-modal self-attention [33]; BRINet reasoned bidirectionally over relationships [34]; and LSCM modeled linguistic structure [35]. CMPC and its journal extension progressively decomposed and aligned visual and language cues [36,37]. Transformer-based methods strengthened cross-modal context: LAVT injected language into a hierarchical vision Transformer, CRIS exploited CLIP-aligned representations, RIS-DMMI reconsidered one-to-one matching, CARIS modeled context, and CrossVLT divided cross-aware fusion across encoder stages [38–42]. RRSIS adds domain-specific difficulties. RefSegRS contains 4420triplets from 285 aerial scenes and includes categories, attributes, and spatial relations [8,43]. Its language-guided cross-scale enhancement baseline showed the need to combine shallow detail and deep semantics. RRSIS-D, derived from remote sensing visual grounding data, contains 17,402 triplets across 20 categories and substantial scale and rotation variation [9,44]. RMSIN explicitly introduced rotated multi-scale interaction. FIANet decomposes expressions into object and position semantics for finer alignment [10]. More recent methods explore CLIP-driven dynamic feature selection, two-way guidance, and grounding-aware stagewise reasoning [45–47]. The common lesson is that language grounding and accurate mask delineation are separable but interdependent. Weak grounding selects the wrong instance, while weak decoding produces incomplete or blurred boundaries even when the instance is localized. UniRS-Seg adopts coarse-to-fine prompting for this reason, but it differs in scope: class-name prompts and referring expressions enter the same prompt interface and mask decoder.

2.3. Segmentation and Vision-Language Foundation Models

SAM consists of a large image encoder, prompt encoder, and lightweight two-way mask decoder [11]. Its promptability and strong boundary prior are attractive for remote sensing, but direct transfer is affected by the gap between ground-view natural images and overhead scenes [13]. SAM-Adapter demonstrated that lightweight trainable modules can inject task-specific information without fully fine-tuning the foundation backbone [12]. Parameter-efficient tuning is especially relevant when combining multiple foundation models. Bottleneck adapters modify activations with small residual modules [48]; LoRA instead parameterizes low-rank weight updates [49]. We use bottleneck adapters to adapt spatial activations while retaining the pretrained SAM weights; Section 5.4 compares this choice with SAM LoRA.

CLIP learned transferable image–text alignment through contrastive supervision [50]. RemoteCLIP and SkySense adapt large-scale multimodal representation learning to Earth observation [14,16]. SigLIP 2 extends sigmoid-loss image–text pretraining with captioning, self-supervised, masked-prediction, and data-curation components; its reported improve-

ments include localization and dense features, making it a suitable interface between text and pixel prediction [15]. We use both its text representations and dense visual tokens. The SigLIP2 image tower supplies semantically aligned localization cues, whereas SAM supplies segmentation-oriented geometry. A learned fusion layer aligns their dimensions and resolutions; the two encoders are not assumed to be natively interchangeable.

Foundation-model RRSIS is a rapidly moving area. EVF-SAM performs early vision-language fusion for text-prompted SAM [51]. RSRefSeg employs CLIP activations as SAM prompts [17]. RS2-SAM2 jointly encodes visual and textual inputs, adapts hierarchical features, and generates a pseudo-mask dense prompt [18]. RSRefSeg 2 uses SigLIP 2 and SAM2 to decouple localization from fine segmentation [19]. These studies motivate language-guided foundation-model prompting and establish the relevant prior art. The architectural distinction evaluated by UniRS-Seg is that the same query-conditioned interface can support both full semantic parsing and referring segmentation while reusing an adapted SAM1 detail-recovery path.

2.4. Universal Segmentation and the Remaining Gap

Mask-classification decoders, exemplified by Mask2Former, represent segmentation as a set of learned queries and masks [52]. This view unifies semantic, instance, and panoptic segmentation more naturally than a fixed pixel classifier. Referring segmentation also produces a query-specific mask, but its query originates in language rather than a learned object slot. UniRS-Seg connects these views by treating a semantic vocabulary as multiple text queries and a referring expression as one text query. The decoder always predicts masks conditioned on query embeddings; task-specific post-processing determines whether the outputs are normalized across classes or thresholded as one binary mask. UniGeoSeg addresses a broader instruction-driven setting with referring, interactive, and reasoning segmentation [20]. CrossEarth uses multi-task training for remote sensing domain generalization [53]. UniRS-Seg instead evaluates supervised semantic/referring transfer through a shared foundation-model prompt interface. A shared model can benefit from complementary supervision or suffer negative transfer between class coverage and instance discrimination. The reported controls therefore compare single-task training, two task-specific models, task-balanced shared training, and shared training with consistency. They assess the present supervised setting rather than replacing a cross-domain evaluation.

3. Materials and Methods

3.1. Problem Formulation

Let $I \in \mathbb{R}^{H\times W\times 3}$ denote an optical remote sensing image. In semantic mode, a dataset supplies a vocabulary $\mathcal{C} = \{c_1, \ldots, c_C\}$ and a dense label map $Y^{\text{sem}} \in \{1, \ldots, C\}^{H\times W}$. The desired output is a set of C logit maps whose pixel-wise softmax gives the semantic prediction. In referring mode, the input additionally contains an expression $e = (w_1, \ldots, w_L)$ and the target is a binary mask $Y^{\text{ref}} \in \{0, 1\}^{H\times W}$. Rather than defining two output heads, we express both inputs as a query set Q. Semantic mode contains one query per class name; referring mode contains one query representing the expression, accompanied by its token-level context.

This abstraction separates three concerns. The visual pathway answers what geometric and semantic evidence exists in the image. The query pathway states which classes or described instance should be extracted. The shared decoder maps their interaction to masks. Dataset taxonomies can therefore differ in size without changing the decoder weights. All classes, including LoveDA background and ISPRS clutter/background, remain explicit during training; evaluation may omit a class only when the established benchmark protocol does so.

3.2. Architecture Overview

Figure 1 summarizes UniRS-Seg. The image is processed by two complementary foundation pathways. The adapted SAM1 branch supplies segmentation-oriented tokens and multi-level spatial features. A lightweight CNN in parallel retains high-frequency texture and feeds cross-branch attention. The SigLIP 2 image tower supplies dense vision-language features aligned with the text tower. The text input is either a prompt ensemble built from semantic class names or a free-form referring expression. A query compiler projects both forms into a common decoder space. The two foundation pathways are detailed in Sections 3–3.6.3 below.

The shared prompt interface performs query-to-image matching on the SigLIP 2 grid. It returns a coarse dense activation prompt and a sparse query token for every requested mask. These prompts condition a generalized SAM mask decoder. Learnable mask cross-attention (LMCAM) uses the current coarse probability as a differentiable spatial prior rather than a hard gate. In semantic mode, class-balanced mask-guided self-attention (CMAttn) further refines image embeddings with soft class prototypes while compensating for the long-tailed pixel distribution. Finally, the hierarchical pixel decoder combines decoder states, intermediate SAM features, and CNN features to recover thin objects and boundaries. Semantic mode runs the query set in parallel and applies a pixel-wise softmax; referring mode applies a sigmoid to one mask.

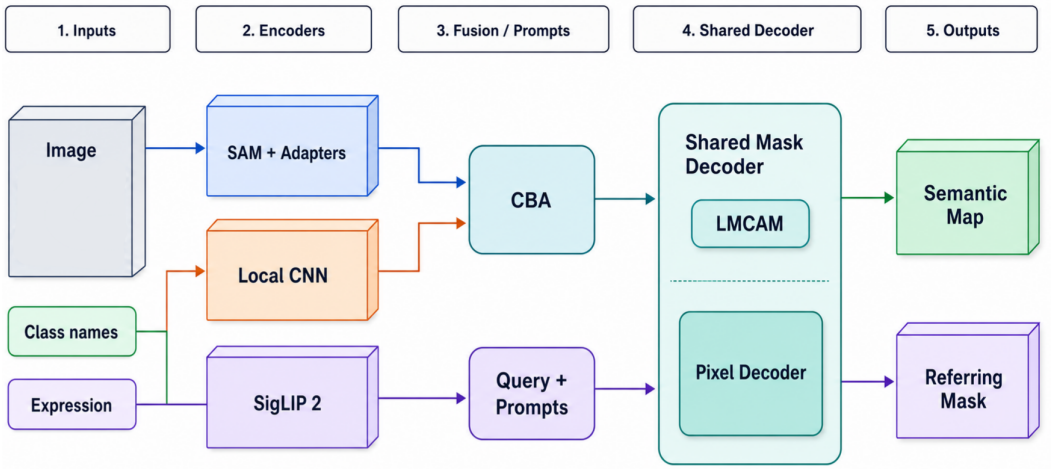


Figure 1. Overall architecture of UniRS-Seg. The gray block denotes the image; blue denotes the adapted SAM encoder; orange denotes the local CNN; purple denotes SigLIP 2, query/prompt construction, and the referring-mask output; green denotes class-name inputs and the semantic-map output. Teal blocks denote cross-branch attention (CBA) and the shared mask/pixel decoder, including LMCAM. Semantic class names and referring expressions provide the two task interfaces to the shared query-conditioned decoder. Figure 3 details the image–text pathway and sparse–dense prompt construction.

3.3. SAM Domain Adaptation and Local-Detail Pathway

3.3.1. Bottleneck Adapters

The SAM ViT-B image encoder contains 12 Transformer blocks. Its pretrained weights are frozen by default, and a residual bottleneck adapter is inserted after the attention and feed-forward sublayers of each block. For a token tensor $F^{(k)}$ at block k, the adapter update is

$$\widetilde{F}^{(k)} = F^{(k)} + s_k \,\mathrm{GELU}\Big(\mathrm{LN}(F^{(k)})W^{(k)}_{\mathrm{down}}\Big)W^{(k)}_{\mathrm{up}}, \tag{1}$$

Tokensare rows of $F^{(k)} \in \mathbb{R}^{N_s \times d_{\mathrm{sam}}}$; $W_{\mathrm{down}} \in \mathbb{R}^{d_{\mathrm{sam}} \times d_b}$ reduces channels and $W_{\mathrm{up}} \in \mathbb{R}^{d_b \times d_{\mathrm{sam}}}$ restores them. The scalar s_k is a learnable residual scale initialized near zero.

Unlike a prompt vector shared across spatial locations, this adapter transforms each token and can learn remote-sensing-specific appearance while preserving the pretrained backbone at initialization. The bottleneck ratio is 1:4. The adaptation ablation compares this configuration with frozen SAM and SAM LoRA; it does not include a bottleneck-ratio sweep or full fine-tuning.

Features are retained from blocks 3, 6, 9, and 12. Layer normalization and 1×1 projections map these tensors to a common pyramid width before hierarchical decoding. Retaining intermediate features is necessary because the final SAM grid alone is too coarse for cars, narrow roads, roof outlines, and thin vegetation boundaries.

3.3.2. Parallel CNN and Cross-Branch Attention

The parallel convolutional branch contains a stem followed by four stages at strides 2, 4, 8, and 16. Each stage uses depthwise separable 3×3 convolutions, normalization, GELU, and residual connections. The branch is intentionally lightweight: its role is to preserve local contrast, corners, texture, and boundary frequency rather than duplicate the global encoder. The stride-16 CNN feature F_c is aligned with the final SAM token grid F_s through learned projections.

Cross-branch attention uses SAM tokens as queries and CNN tokens as keys and values:

$$F_{\mathrm{sc}} = F_s + \mathrm{softmax}\left(\frac{(F_s W_q)(F_c W_k)^{\mathsf{T}}}{\sqrt{d}} + B_{\mathrm{rel}}\right)(F_c W_v), \tag{2}$$

where B_{rel} is a relative-position bias and d is the attention width. A feed-forward residual block follows the attention. This direction of attention allows each globally contextual SAM token to retrieve the local evidence most relevant to its location. The lower-resolution CNN stages bypass this module and enter the hierarchical decoder as skip features. Figure 2 illustrates the complete adaptation and detail-recovery pathway.

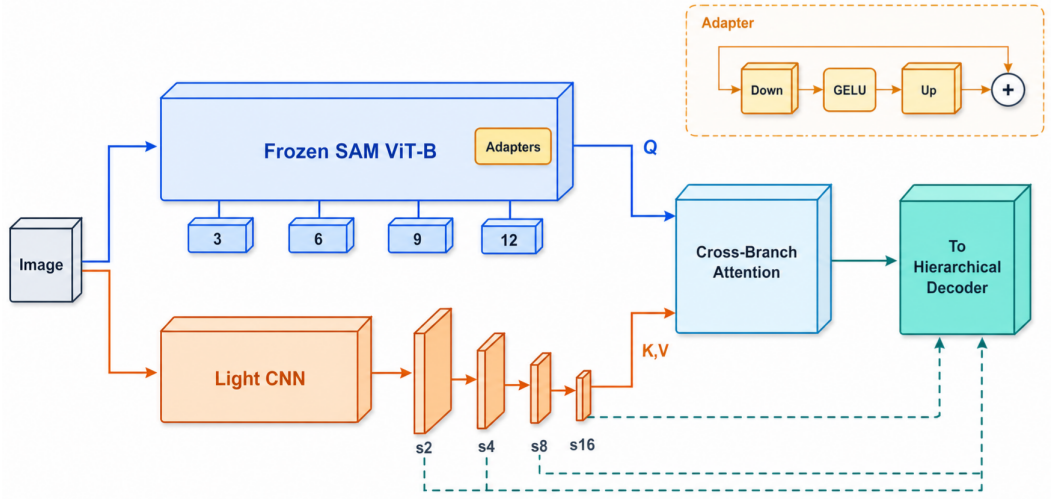


Figure 2. SAM domain adaptation and local-detail pathway. Blue blocks denote the frozen SAM ViT-B backbone and features retained from blocks 3, 6, 9, and 12; yellow/orange blocks denote trainable bottleneck adapters, their down–GELU–up residual update, and the lightweight CNN with stride-2/4/8/16 features. The light-blue block denotes cross-branch attention, and the teal block denotes the interface to the hierarchical decoder. SAM tokens supply queries (Q), while aligned CNN features supply keys and values (K, V). Dashed teal arrows carry local skip features to the decoder; gray denotes the input image.

3.4. SigLIP 2 Query Pathway

3.4.1. Dense Visual and Text Representations

The fixed-resolution `siglip2-so400m-patch16-512` checkpoint uses separate ViT image and text towers with a shared 1152-dimensional projection space. For a 512×512 image and a patch size of 16, the vision tower returns $N_v = 32 \times 32 = 1024$ unpooled patch tokens. The text tower accepts at most 64 tokens and returns both token-level states and a pooled representation [15]. We denote these outputs by

$$X^v = E_v(I_{512}) \in \mathbb{R}^{N_v \times d_s}, \qquad X^t(s) = E_t(s) \in \mathbb{R}^{L_s \times d_s}, \qquad g^t(s) = \mathrm{Pool}\big(X^t(s)\big), \tag{3}$$

Here, $d_s = 1152$, $L_s \leq 64$, and Pool denotes the checkpoint's pooled text output used by the query pathway. The pooled image embedding is useful for image–text alignment, but UniRS-Seg retains the unpooled patch states X^v because the dense spatial grid is required for localization. Padding patches introduced by aspect-ratio-preserving preprocessing are excluded with the recorded geometric valid-region mask.

SigLIP 2 retains SigLIP's pairwise sigmoid objective and augments its pretraining recipe with captioning/localization supervision, self-distillation, and masked-patch prediction. For a batch of B matched image–text pairs, the pairwise component follows the original SigLIP batch normalization [54]:

$$\mathcal{L}_{\mathrm{sig}} = \frac{1}{B} \sum_{i=1}^{B} \sum_{j=1}^{B} \log\Big[1 + \exp\Big(-y_{ij}\Big(\exp(t)\big\langle \hat{g}_i^v, \hat{g}_j^t \big\rangle + b\Big)\Big)\Big], \quad y_{ij} = \begin{cases} +1, & i = j, \\ -1, & i \neq j, \end{cases} \tag{4}$$

where hats indicate ℓ_2 normalization, t is a learned log-scale, and b is a learned bias. The auxiliary decoder used during SigLIP 2 pretraining is not part of the released encoder. UniRS-Seg therefore uses the released global embeddings and dense hidden states, while its own query compiler and mask decoder learn the remote-sensing task.

3.4.2. Semantic Class Queries

For a semantic dataset, every category name c is inserted into a fixed template set $\mathcal{T}$. Examples include "an aerial image of {class}", "a remote sensing view of {class}", and "the land-cover class {class}." If $s_{c,m}$ is the mth prompted class string, its query is

$$\bar{q}_c = \frac{1}{|\mathcal{T}|} \sum_{m=1}^{|\mathcal{T}|} \frac{W_t g^t(s_{c,m})}{\|W_t g^t(s_{c,m})\|_2}, \qquad q_c = \mathrm{LN}\Big(\frac{\bar{q}_c}{\|\bar{q}_c\|_2} + e_{\mathcal{D}}\Big), \tag{5}$$

where $W_t \in \mathbb{R}^{d_q \times d_s}$ is shared across datasets and $e_{\mathcal{D}}$ is a learned dataset embedding that captures taxonomy-specific label usage, such as the meaning of "background." Stacking the C vectors gives $Q^{\mathrm{sem}} = [q_1; \ldots; q_C] \in \mathbb{R}^{C \times d_q}$. These queries are cached after training; consequently, semantic inference requires only the image even though the classifier weights originate from text.

The prompt ensemble makes the classifier vocabulary explicit while avoiding dependence on a single sentence form. It is not used to infer synonyms at evaluation time: benchmark class names and their documented meanings are fixed before training.

3.4.3. Referring Expression Queries

For a referring expression e, the pooled text vector initializes a single query, while token states preserve attributes and relations. A two-layer compiler first attends to the expression tokens and then to the dense image tokens:

$$\begin{aligned} T_e &= \mathrm{SA}(X^t(e)), \qquad q_e^{(0)} = W_g g^t(e), \\ q_e^{(1)} &= q_e^{(0)} + \mathrm{MHA}\Big(q_e^{(0)}, T_e W_k^t, T_e W_v^t\Big), \\ q_e &= \mathrm{LN}\Big(q_e^{(1)} + \mathrm{MHA}\Big(q_e^{(1)}, X^v W_k^v, X^v W_v^v\Big) + \mathrm{MLP}(q_e^{(1)})\Big). \end{aligned} \tag{6}$$

Here, SA is token self-attention and $\mathrm{MHA}(Q, K, V)$ is multi-head cross-attention. The first interaction preserves word-level modifiers such as color, size, absolute direction, and relative position; the second makes the final query image-conditioned so that identical language can select different spatial evidence in different scenes. The referring query set is $Q^{\mathrm{ref}} = [q_e] \in \mathbb{R}^{1 \times d_q}$.

3.4.4. Parameter-Efficient VLM Adaptation

The first optimization stage freezes SigLIP 2 and trains only the query compiler and projection layers. The second stage enables low-rank adaptation (LoRA) in the last four image-tower blocks and the last two text-tower blocks. For a frozen projection $W_0 \in \mathbb{R}^{d_o \times d_i}$, the adapted mapping is

$$h' = W_0 h + \frac{\alpha}{r} B A h, \qquad A \in \mathbb{R}^{r \times d_i}, \quad B \in \mathbb{R}^{d_o \times r}, \tag{7}$$

with ranks $r_v = 16$ for vision and $r_t = 8$ for text. Only A and B are optimized. Staging the adaptation prevents randomly initialized prompt and decoder layers from immediately perturbing the aligned embedding space. All reported results use the So400m/16 checkpoint at 512×512; smaller checkpoints are treated as distinct model variants rather than silent replacements.

3.5. Shared Sparse–Dense Prompt Interface

Let $V \in \mathbb{R}^{h_v \times w_v \times d_q}$ be the projected SigLIP 2 image grid and let $q_j \in \mathbb{R}^{d_q}$ be the jth semantic or referring query. Their normalized similarity yields a coarse activation map

$$A_j(u, v) = \exp(-t_a) \left\langle \frac{V(u, v)}{\|V(u, v)\|_2}, \frac{W_q q_j}{\|W_q q_j\|_2} \right\rangle, \qquad D_j = \mathrm{Up}_{H/4, W/4}(A_j), \tag{8}$$

where t_a is a learned log-temperature and D_j is the dense prompt. Only logits used in the spatial softmax are masked to $-\infty$ at padded locations; dense-prompt interpolation uses finite activation values and the recorded valid-region geometry. Activation-weighted visual evidence and the sparse prompt token are

$$\alpha_j(u, v) = \frac{\exp A_j(u, v)}{\sum_{u', v'} \exp A_j(u', v')}, \qquad r_j = \sum_{u, v} \alpha_j(u, v) V(u, v), \qquad S_j = W_s[q_j \,\|\, r_j] + p_j, \tag{9}$$

where $\|$ denotes concatenation and p_j is the positional encoding in the SAM decoder space. We adopt cosine similarity because it decouples the match score from the magnitude of the SigLIP 2 embeddings, so a query's alignment with a spatial location depends only on directional agreement rather than on feature norm, which can otherwise vary systematically across classes or expressions.

This interface preserves the useful coarse-localization/fine-segmentation decomposition found in foundation-model RRSIS, but it operates identically for C semantic queries or one referring query. In semantic mode the entire class bank is processed in parallel. In referring mode only the expression query is active. The interface does not derive hard points or boxes, avoiding non-differentiable prompt selection and threshold-sensitive failures on small objects.

Figure 3 summarizes the query construction and prompt transfer from SigLIP 2 to the shared decoder.

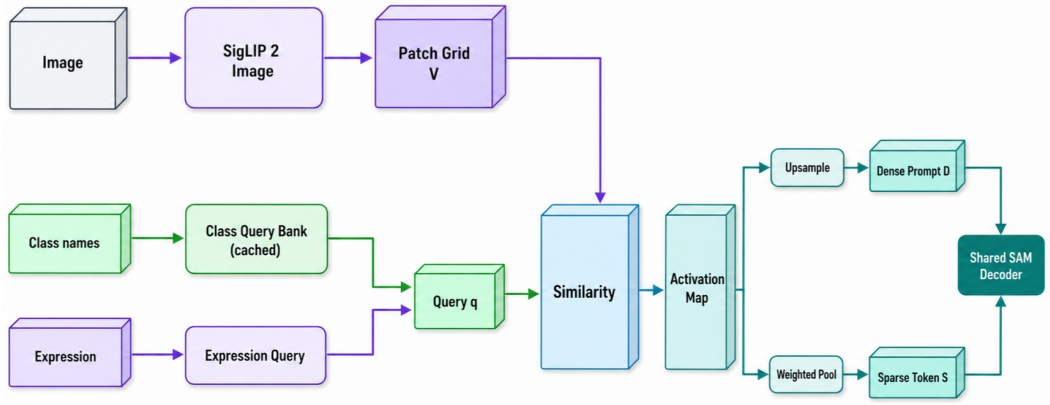


Figure 3. SigLIP 2 query and sparse–dense prompt pathway. Purple blocks denote image encoding, the patch grid V, and the expression-query branch; green blocks denote class-name inputs, the cached semantic query bank, and the common query q. The blue block computes query–patch similarity. Teal blocks denote the activation map, upsampling to the dense prompt D, activation-weighted pooling to the sparse token S, and the shared SAM decoder. Gray denotes the input image. Semantic and referring queries use the same prompt-construction operations.

3.6. Query-Conditioned Mask Decoding

3.6.1. Dynamic SAM Mask Tokens

The original SAM mask decoder uses a fixed set of output tokens and hypernetworks. We retain its two-way Transformer initialization but make mask tokens query-conditioned. A shared learned seed is added to a projected query token, yielding one mask token for each active query. Batching these tokens permits a variable number of masks without changing decoder weights. The corresponding hypernetwork maps each decoded query token to channel weights applied to the upsampled image feature. This produces coarse mask logits at one-quarter input resolution.

The decoder also receives the dense prompt D_j through the SAM dense-prompt projection. Consequently, the sparse token carries category or expression semantics, while the dense prompt supplies coarse spatial support. An ablation separates sparse-only, dense-only, and joint prompting.

3.6.2. Learnable Soft Mask Cross-Attention

Hard binary attention masks can suppress a target permanently when an early prediction is wrong. We instead turn the current probability map $M_j \in [0,1]^{h\times w}$ into a learnable additive bias:

$$\mathrm{LMCAM}(Q_j, K, V) = \mathrm{softmax}\left(\frac{Q_j K^{\mathsf{T}}}{\sqrt{d}} + \beta_j \log(\epsilon + \mathrm{vec}(M_j))\right) V, \tag{10}$$

where $\beta_j = \mathrm{softplus}(b_j)$ controls confidence in the mask and ϵ prevents numerical singularity. At initialization β_j is small, so cross-attention can recover from poor localization. As training proceeds, the model can use reliable masks to suppress irrelevant regions. This

formulation is differentiable everywhere and applies the mask before value aggregation, unlike multiplying an already aggregated attention output.

Figure 4 makes the operation order explicit: the mask is transformed into an additive log-prior before softmax, rather than multiplied onto an already normalized saliency map. For query j and spatial token n, let $z_{jn} = Q_j K_n^{\mathsf{T}} / \sqrt{d}$. Equation (10) gives attention proportional to $\exp(z_{jn})(\epsilon + M_{jn})^{\beta_j}$. For two locations a and b, the attention ratio is

$$\frac{a_{ja}}{a_{jb}} = \exp(z_{ja} - z_{jb}) \left(\frac{\epsilon + M_{ja}}{\epsilon + M_{jb}} \right)^{\beta_j}. \tag{11}$$

This separates feature compatibility from mask confidence. At $\beta_j = 0$ the limiting operation is ordinary cross-attention. For $0 < \beta_j < 1$, mask contrast is softened; at $\beta_j = 1$, the prior is proportional to the current probability; and $\beta_j > 1$ sharpens preference for already confident locations. Softplus constrains the learned confidence to positive values, with initialization near zero reducing early error reinforcement. A larger ϵ weakens the distinction between low-confidence locations, while an excessively small value increases the gradient factor $\beta_j / (\epsilon + M_{jn})$ near zero.

The query–key term can confuse repeated roofs or vehicles when appearance is similar. The mask prior can reinforce an incorrect instance or fragment a small target if the initial activation is poor. The learnable confidence adjusts reliance on that prior but does not guarantee recovery from a confidently wrong localization. Value features supply the content used for reconstruction; attention reweighting cannot restore an object absent from those features. The CNN pathway and hierarchical decoder preserve local evidence, while soft masking leaves a recovery path through nonzero attention. These components address different error sources rather than ensuring invariance to occlusion, resolution loss, or unseen scenes.

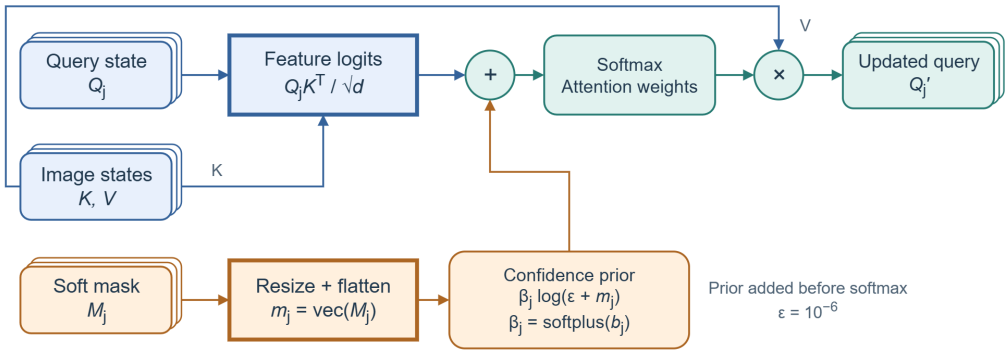


Figure 4. Learnable soft mask cross-attention corresponding to Equation (10). Blue token stacks and the score block denote projected query/image states and feature logits; orange blocks align the predicted soft mask and scale its log-prior by learned confidence; teal circles denote score addition and value aggregation, and the teal rounded block normalizes attention. The mask prior is added to query–key logits before softmax. Gradients pass through both the soft mask and confidence.

3.6.3. Class-Balanced Mask-Guided Self-Attention

Remote sensing taxonomies are strongly imbalanced: extensive background, building, or vegetation regions can dominate the features used to refine much smaller classes such as cars and narrow roads. CMAttn is therefore enabled in semantic mode after LMCAM. It uses the decoder's current soft predictions rather than ground-truth masks, avoiding a train–test mismatch. Let $E \in \mathbb{R}^{N \times d}$ denote image embeddings and $P \in [0,1]^{C \times N}$ the

semantic probabilities at the same spatial resolution. For each class, the mask-weighted prototype and dispersion are

$$\mu_c = \frac{\sum_n P_{cn} E_n}{\sum_n P_{cn} + \epsilon}, \qquad \sigma_c^2 = \frac{\sum_n P_{cn} (E_n - \mu_c)^2}{\sum_n P_{cn} + \epsilon}, \qquad \widetilde{\mu}_c = \mu_c + \gamma\, \widehat{w}_c\, \xi_c \odot \sqrt{\sigma_c^2 + \epsilon}, \tag{12}$$

Here γ controls perturbation strength, $\xi_c \sim \mathcal{N}(0, I)$ during training, and $\widehat{w}_c$ is the normalized effective-number weight computed from training-pixel frequency. Squaring, square-root evaluation, and $\odot$ act element-wise on channels. The rare-class prototypes consequently receive stronger, variance-aware augmentation. The augmented prototypes are concatenated with the image tokens, processed by self-attention and a feed-forward block, and returned through a gated residual connection. At inference, Gaussian sampling is disabled ($\xi_c = 0$), all class queries are retained, and no ground-truth label is required.

In semantic decoding, the same effective-number weights scale cross-entropy and parameterize a learned class-frequency embedding that refines the initial class query q_c from Equation (5). Query dropout affects frequent classes only in auxiliary decoder layers; the final layer always predicts the complete vocabulary. Referring mode has one expression query rather than a semantic class distribution, so it bypasses stochastic class balancing while retaining LMCAM and the shared decoder. Figure 5 illustrates the semantic-mode CMAttn path from mask-guided prototypes to enhanced image embeddings.

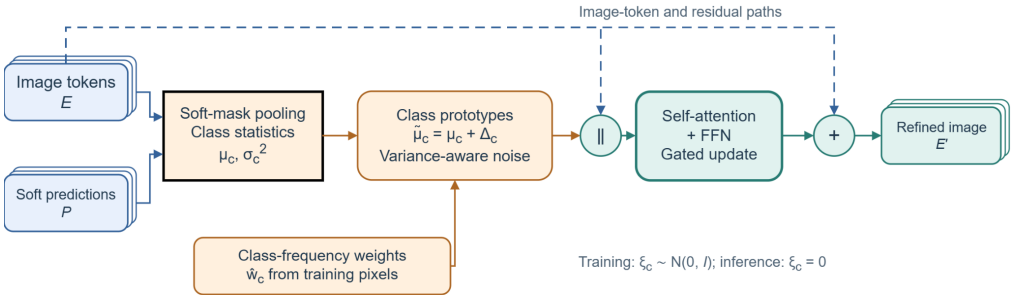


Figure 5. Semantic-mode class-balanced mask-guided self-attention. Blue token stacks denote image embeddings and soft semantic probabilities; orange blocks denote soft-mask pooling, class-frequency weights, and variance-aware prototype perturbation. Teal denotes token concatenation (‖), self-attention/FFN with a gated update, and residual addition (+); the dashed blue path supplies image tokens and the residual. Gaussian sampling is disabled at inference. Soft predictions, rather than ground-truth masks, form the prototypes. Referring inference bypasses this semantic class-balancing path.

3.6.4. Hierarchical Pixel Decoder

The hierarchical pixel decoder reconstructs full-resolution logits in three stages. Starting from the stride-16 decoder feature, it upsamples to strides 8, 4, and 1. Before each concatenation, retained SAM and CNN features are projected and resized to the current decoder grid; they are then combined with the upsampled query-conditioned state. A 3×3 depthwise separable fusion block and a residual 1×1 projection refine the result. Auxiliary masks at strides 8 and 4 provide deep supervision; only the full-resolution output is used for reporting.

For semantic mode, the C full-resolution mask logits are stacked and normalized across the query dimension. For referring mode, the single logit map is passed through a sigmoid. In compact notation,

$$P^{\text{sem}}(u, v, c) = \text{softmax}_c\{Z_c(u, v)\}, \qquad P^{\text{ref}}(u, v) = \sigma\{Z_e(u, v)\}. \tag{13}$$

The masks are therefore generated by the same decoder parameters, while output normalization expresses the different task semantics.

3.7. Joint Training Objective

Training alternates task-balanced mini-batches rather than forcing semantic and referring samples into one batch structure. For semantic data, the main losses are class-balanced cross-entropy and multi-class Dice. For referring data, binary cross-entropy and binary Dice supervise the expression mask. A boundary loss compares morphological gradients of prediction and target at a fixed three-pixel radius after resizing to a common evaluation scale. Auxiliary decoder stages receive the same task-appropriate mask losses at lower weight.

Two cross-modal objectives regularize the shared interface. First, masked average visual features are aligned with their class or expression query using an in-batch contrastive loss. Second, each semantic mask produces a conservative template expression of the form "all {class}"; the referring-mode output for that template is constrained to agree with the semantic class probability. This consistency loss is applied only to semantic samples and only to class-union language, so it does not pretend that semantic annotations contain instance-level relations.

The total objective is

$$\mathcal{L} = \lambda_{\text{sem}}\mathcal{L}_{\text{sem}} + \lambda_{\text{ref}}\mathcal{L}_{\text{ref}} + \lambda_{\text{aux}}\mathcal{L}_{\text{aux}} + \lambda_{\text{bnd}}\mathcal{L}_{\text{bnd}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{cons}}\mathcal{L}_{\text{cons}}. \quad (14)$$

Task terms absent from the current sample are set to zero. We use $\lambda_{\text{sem}} = \lambda_{\text{ref}} = 1$, $\lambda_{\text{aux}} = 0.4$, $\lambda_{\text{bnd}} = 0.2$, $\lambda_{\text{align}} = 0.1$, and $\lambda_{\text{cons}} = 0.1$ in all primary experiments.

3.8. Training and Inference Modes

Training has three stages. Stage 1 freezes both foundation encoders and trains adapters, projections, the CNN branch, prompt interface, and decoder for stable initialization. Stage 2 continues training the SAM adapters and enables the selected SigLIP 2 LoRA blocks. Stage 3 jointly fine-tunes on task-balanced batches with a reduced learning rate and cross-task consistency. Dataset sampling is balanced first by task and then by dataset, preventing the larger RRSIS-D set from dominating RefSegRS and preventing tiled semantic datasets from overwhelming referring data.

At semantic inference, the chosen dataset vocabulary is encoded once and cached. All class queries are decoded in parallel, and each pixel receives the class with maximum probability. At referring inference, the expression is encoded online and one mask is thresholded at 0.5; no point, box, or human correction is used. For large tiles, overlapping sliding-window inference with logit averaging is applied before argmax or thresholding.

4. Experimental Setup

4.1. Datasets

4.1.1. LoveDA

LoveDA contains 5987 high-spatial-resolution images of size 1024×1024 pixels from Nanjing, Changzhou, and Wuhan, China [55]. The 2522/1669/1796 image train/validation/test split covers urban and rural domains. Its seven labels are background, building, road, water, barren land, forest, and agriculture. The combination of scene-domain shift, class imbalance, small structures, and visually diverse background makes it suitable for testing both the adapted geometry pathway and vocabulary-conditioned mask queries. We use the official split, train on the training set, select hyperparameters on validation only, and

obtain test labels through the official evaluation route. No test image is used for prompt selection or early stopping.

4.1.2. ISPRS Vaihingen

The Vaihingen benchmark contains 33 true orthophoto tiles with an average size of approximately 2494 × 2064 pixels and a ground sampling distance of 9 cm [56]. The imagery provides near-infrared, red, and green channels, DSM data, and labels for impervious surfaces, buildings, low vegetation, trees, cars, and clutter/background. For comparability with the established comparison protocol, we use the common 16/17 tile split:training IDs 1, 3, 5, 7, 11, 13, 15, 17, 21, 23, 26, 28, 30, 32, 34, and 37; test IDs 2, 4, 6, 8, 10, 12, 14, 16, 20, 22, 24, 27, 29, 31, 33, 35, and 38. The IRRG channels are mapped to three input channels. Patches of 1024 × 1024 pixels are extracted with 50% overlap for training; overlapping logits are averaged for tile reconstruction. The five foreground categories are used for the primary mIoU and mF1 comparison, matching the cited table, while clutter/background is still predicted and the present table reports the five-class protocol only.

4.1.3. ISPRS Potsdam

Potsdam comprises 38 georeferenced 6000 × 6000 tiles at 5 cm ground sampling distance and uses the same six labels as Vaihingen [56]. RGB, IRRG, DSM, and normalized DSM products exist, but the primary experiment uses RGB only to match the established comparison protocol.Test IDs are 2_13, 2_14, 3_13, 3_14, 4_13, 4_14, 4_15, 5_13, 5_14, 5_15, 6_13, 6_14, 6_15, and 7_13. Tile 7_10 is excluded because of the documented annotation issue in the comparison protocol; the remaining 23 tiles are used for training. Patch extraction and reconstruction follow the Vaihingen procedure.

4.1.4. RefSegRS

RefSegRS is built on high-resolution SkyScapes aerial scenes and contains 4420 image–expression–mask triplets from 285 scenes [8,43]. The official scene-disjoint split contains 2172 training triplets from 151 scenes, 431 validation triplets from 31 scenes, and 1817 test triplets from 103 scenes. Images are standardized to 512 × 512 pixels, and the annotation includes 14 hierarchical categories and five attribute groups. Scene-level separation is preserved to prevent nearly identical crops from leaking across splits. The primary model receives the image at 1024 × 1024 in the SAM branch and 512 × 512 in the SigLIP2 branch; the ground-truth mask is resized only with nearest-neighbor interpolation.

4.1.5. RRSIS-D

RRSIS-D was introduced with RMSIN and is derived from the RSVGD remote sensing visual-grounding data [9,44]. It contains 17,402 image–expression–mask triplets over 20 categories, divided into 12,181 training, 1740 validation, and 3481 test samples. Images are represented at 800 × 800 pixels and span spatial resolutions from approximately 0.5 to 30 m per pixel. Targets range from small vehicles and windmills to large airports and fields, and many classes occur at arbitrary orientation. For SAM, each image is letterboxed to 1024 × 1024 with a recorded valid-region mask; for SigLIP2 it is resized with aspect-ratio-preserving padding to 512 × 512. Predictions are mapped back to the original 800 × 800 grid before metrics are computed.

4.1.6. RISBench

RISBench, introduced with CroBIM, contains 52,472 image–expression–mask triplets built from VRSBench annotations, with imagery originating from DOTA-v2 and DIOR [57]. It covers 26 object categories, eight attribute types, 512 × 512 image patches, and ground sampling distances of 0.1–30 m. We conduct an additional supervised benchmark exper-

iment using its official training partition and report results on its official test partition. The SAM/SigLIP 2 input sizes and mask restoration follow the referring configuration above. This experiment extends dataset coverage; the original five-benchmark joint-training controls and their checkpoint-selection protocol are retained.

4.2. Data Processing and Augmentation

Image normalization follows the pretrained encoder associated with each branch. The SAM input uses SAM's pixel mean and standard deviation; the SigLIP 2 input uses the checkpoint processor's normalization. A single decoded RGB image is transformed into the two resolutions to avoid inconsistent color conversion. Masks always use nearest-neighbor resampling.

Semantic training uses random scale in $[0.5, 2.0]$, random 1024×1024 crop, horizontal and vertical flips, and rotations by multiples of 90°. Mild brightness, contrast, and saturation jitter is applied with probability 0.5. Referring training uses photometric jitter and random resize only in the primary protocol. Naive flips or rotations are not used because they would invalidate expressions such as "upper-left building" or "vehicle east of the road." This distinction preserves annotation validity: a geometric transform changes both the mask and the spatial reference frame of directional language. Semantic geometry augmentation broadens orientation and scale coverage; mild photometric jitter broadens appearance coverage without changing the target. Referring resizing preserves relative position and photometric jitter leaves the expression unchanged. Primary results use this policy; Referring resizing preserves relative position and photometric jitter leaves the expression unchanged. Primary results use this policy; the effect of training without augmentation is evaluated separately in the Results section.

4.3. Evaluation Metrics

For semantic segmentation, we report per-class intersection over union (IoU), mean IoU (mIoU), per-class F1, mean F1 (mF1), and overall pixel accuracy (OA). The primary Vaihingen and Potsdam metrics average the five foreground classes in accordance with the cited comparison source. The reported semantic comparisons use the region metrics above; no additional boundary metric is inferred from the qualitative panels.

For referring segmentation, generalized IoU (gIoU) is the arithmetic mean of per-sample mask IoUs. Cumulative IoU (cIoU) sums intersections over all test samples and divides by the summed unions; it therefore gives larger targets greater influence. Precision at threshold x, denoted Pr@x, is the percentage of samples whose IoU exceeds x, with $x \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$. The present referring tables report these overlap and threshold metrics; boundary F1 and median IoU are not included. Empty predictions and empty targets follow the official dataset evaluator; no post hoc removal of difficult examples is permitted.

4.4. Implementation Settings

The primary configuration uses SAM ViT-B and `siglip2-so400m-patch16-512`. New decoder, CNN, adapter, and projection parameters use AdamW with learning rate 1×10^{-4}; SAM adapters use 5×10^{-5}; and SigLIP2 LoRA parameters use 1×10^{-5}. Weight decay is 0.01 except for normalization, bias, prompt, and temperature parameters, which use zero decay. Gradients are clipped to a global norm of 1.0. A linear warm-up of 1,500 optimizer steps is followed by cosine decay to 1×10^{-6}.

Stage 1 trains newly initialized modules for 30 epochs with both foundation encoders frozen. Stage 2 trains adapters and selected LoRA parameters for 80 epochs. Stage 3 performs task-balanced joint training for 100 epochs at one-tenth of the Stage 2 learning rates. The effective batch size is 4 through gradient accumulation; each optimizer step draws

one task, then one dataset within that task, with equal task probability and inverse-square-root dataset-size sampling. Automatic mixed precision uses bfloat16 when supported and float16 otherwise. Gradient checkpointing is enabled for both foundation encoders. We have trained our model on a single RTX 5090 GPU.

Primary models are trained with seeds 0, 1, and 2, and model selection uses the arithmetic mean of normalized validation mIoU on the three semantic datasets and validation gIoU on the two referring datasets. No test metric participates in checkpoint selection. The internal UniRS-Seg configurations follow the same implementation setting. Published baseline values retain their source protocols; they are not described as reruns in this environment.

Computational Complexity

The full UniRS-Seg implementation contains 1140.7 M parameters, including 540.7 M in the vision encoders. Of the full model, 32.1 M parameters are trainable (2.81%); this is a subset of the total, not an additional parameter count. A referring forward pass for one image–expression pair requires 2305.16 GFLOPs with a 1024 × 1024 SAM input and a 512 × 512 SigLIP 2 image input. On a single RTX 5090, mean forward-pass latency is 312 ms, with 28.4 GB peak training memory and 14.8 GB peak inference memory. Frozen foundation weights reduce the optimization burden but remain active in the image and expression pathways at inference.

Table 1 compares these measurements with published referring-segmentation statistics from Dong et al.,[47]. The 32.1 M trainable count is smaller than the listed baseline counts, while the full parameter count and 2305.16 GFLOPs make the deployment cost explicit. Parameter-efficient adaptation therefore addresses optimization cost rather than eliminating foundation-model inference. The baseline rows retain their source checkpoints, inputs, and profiling environment; the latency values provide published context rather than a common-hardware speed ranking.

Table 1. Computational statistics for referring segmentation. Baseline measurements are from Dong et al., [47]. UniRS-Seg is profiled for one image–expression pair with 1024 × 1024 SAM and 512 × 512 SigLIP 2 inputs; its latency is measured on RTX 5090. M denotes millions of parameters. Best values are shown in bold.

Method	Total (M)	Trainable (M)	GFLOPs	Latency (ms)
Published baseline measurements				
RMSIN [9]	240.04	240.04	1712.19	71.30
LGCE [8]	276.26	276.26	1745.23	66.23
FIANet [10]	251.92	251.92	1712.90	73.49
RSRefSeg 2 [19]	1447.13	97.87	2563.51	267.29
SGSRF [47]	1407.38	58.12	2570.23	263.20
UniRS-Seg measurements				
UniRS-Seg	**1140.70**	**32.10**	**2305.16**	**312.00**

4.5. Comparison Protocol and Provenance

The benchmark splits specified in Section 4 are fixed for all internal UniRS-Seg comparisons and ablations. Checkpoint selection and architecture decisions use validation data, with no test-set retuning. The primary tables compare the corresponding benchmark evaluation partitions; however, published methods retain their original training, validation, and preprocessing choices. Thus, identical sample splits and settings can be asserted for our internal controls, but not independently verified for every source-reported baseline.

Tables 2, 3 and 4 include the comparison reported by WASP-Mamba [29]. These baseline values are interpreted as reported or reproduced in that study rather than as

independently rerun results here. Its protocol uses 1024 × 1024 crops or resized inputs with flip-based test-time augmentation; the UniRS-Seg entries use the matched test-time setting. Table 2 additionally includes the published LoveDA results of STAR-Net [59], a Swin-T-based dual-encoder model that fuses global semantic context and local spatial detail for agricultural land-cover parsing. STAR-Net reports 56.62% mIoU, compared with 57.30% for UniRS-Seg, and its row retains the original article's evaluation setting.

Table 2. Semantic segmentation results on LoveDA (%). Best values are shown in bold.

Method	Backbone	Bac.	Bui.	Roa.	Wat.	Bar.	For.	Agr.	mIoU
BANet [25]	ResT-Lite	46.40	57.71	53.74	80.20	15.75	45.36	63.72	51.84
DC-Swin [26]	Swin-T	44.36	58.65	56.41	80.72	18.18	45.79	59.38	51.93
UNetFormer [2]	ResNet-18	45.29	56.36	57.70	80.72	21.27	46.37	61.13	52.69
RS3Mamba [3]	R18–VMamba-T	45.64	59.13	57.84	79.45	15.11	45.75	63.36	52.32
UNetMamba [4]	ResT-Lite	46.63	61.08	56.15	81.18	17.89	46.79	62.62	53.19
UMFormer [27]	ResNet-18	44.87	57.40	55.44	80.45	14.84	45.39	63.01	51.63
E-PyramidMamba [5]	ResNet-18	44.93	57.81	55.13	80.74	17.46	47.03	62.89	52.28
PPMamba [28]	ResNet-18	45.86	60.10	57.08	81.86	19.41	46.34	63.61	53.47
RECNet [58]	ResNet-18	44.42	56.93	56.80	81.00	16.02	46.19	61.76	51.87
WASP-Mamba [29]	ResT-Lite	48.39	60.10	58.77	80.68	17.08	47.98	64.30	53.90
STAR-Net [59]	Swin-T	48.70	58.78	56.07	75.61	**37.77**	**55.15**	64.25	56.62
UniRS-Seg (ours)	SAM ViT-B	**51.03**	**63.12**	**61.87**	**83.65**	23.01	51.08	**67.31**	**57.30**

Table 3. Semantic segmentation results on ISPRS Vaihingen (%). mIoU and mF1 average the five foreground classes. Best values are shown in bold.

Method	Backbone	Imp.	Bui.	Low.	Tre.	Car.	mF1	mIoU	OA
BANet [25]	ResT-Lite	93.31	90.75	72.72	82.04	77.95	90.73	83.35	93.01
DC-Swin [26]	Swin-T	93.84	91.72	73.39	82.48	75.51	90.72	83.39	93.32
UNetFormer [2]	ResNet-18	93.79	91.37	73.28	82.13	75.61	90.63	83.24	93.27
RS3Mamba [3]	R18–VMamba-T	93.77	91.69	72.94	82.11	77.53	90.86	83.61	93.27
UNetMamba [4]	ResT-Lite	93.62	91.47	73.11	81.99	78.66	90.98	83.77	93.19
UMFormer [27]	ResNet-18	93.37	90.41	72.91	81.53	75.75	90.38	82.80	92.92
E-PyramidMamba [5]	ResNet-18	93.59	91.20	72.87	81.73	74.75	90.38	82.83	93.08
PPMamba [28]	ResNet-18	93.74	91.66	73.46	82.42	75.47	90.70	83.35	93.29
RECNet [58]	ResNet-18	94.13	92.08	74.22	82.17	78.29	91.22	84.18	93.52
WASP-Mamba [29]	ResT-Lite	93.78	91.78	74.00	**82.53**	79.51	91.31	84.32	93.39
UniRS-Seg (ours)	SAM ViT-B	**95.21**	**92.10**	**76.32**	81.08	**81.82**	**91.91**	**85.31**	**94.22**

Table 4. Semantic segmentation results on ISPRS Potsdam (%). mIoU and mF1 average the five foreground classes. Best values are shown in bold.

Method	Backbone	Imp.	Bui.	Low.	Tre.	Car.	mF1	mIoU	OA
BANet [25]	ResT-Lite	87.33	92.51	77.17	79.72	90.46	92.03	85.44	90.86
DC-Swin [26]	Swin-T	89.37	93.60	78.17	80.88	92.43	92.86	86.89	91.68
UNetFormer [2]	ResNet-18	88.27	92.81	77.86	79.78	92.39	92.48	86.22	91.21
RS3Mamba [3]	R18–VMamba-T	87.71	93.14	77.81	80.12	92.75	92.53	86.32	91.21
UNetMamba [4]	ResT-Lite	88.51	93.44	77.88	80.04	91.40	92.50	86.25	91.34
UMFormer [27]	ResNet-18	88.80	93.78	78.31	79.48	**93.14**	92.74	86.70	91.51
E-PyramidMamba [5]	ResNet-18	88.41	93.09	77.72	80.42	92.98	92.65	86.53	91.36
PPMamba [28]	ResNet-18	88.48	92.99	77.80	81.02	92.96	92.73	86.65	91.40
RECNet [58]	ResNet-18	88.45	92.31	76.65	79.26	92.79	92.27	85.89	90.90
WASP-Mamba [29]	ResT-Lite	89.57	94.45	78.74	80.17	92.36	92.95	87.06	91.91
UniRS-Seg (ours)	SAM ViT-B	**90.31**	**96.07**	**80.11**	**81.72**	93.00	**93.63**	**88.24**	**92.07**

Tables 5 and 6 use the RefSegRS and RRSIS-D test comparisons reported by RSRefSeg 2 [19]. They include original RRSIS methods, natural-image referring baselines retrained for aerial imagery, and foundation-model approaches. The additional RISBench baseline results in Table 7 follow RSRefSeg 2 [19], with each compared method cited individually. Baseline values are retained as source-reported percentages; the UniRS-Seg values are produced by the evaluation protocol described above.

Table 5. Referring segmentation results on the RefSegRS test set (%). Best values are shown in bold.

Method	Publication	Pr@0.5	Pr@0.6	Pr@0.7	Pr@0.8	Pr@0.9	cIoU	gIoU
CMPC [36]	CVPR 2020	32.36	14.14	6.55	1.76	0.22	55.39	40.63
BRINet [34]	CVPR 2020	20.72	14.26	9.87	2.98	1.14	58.22	31.51
LSCM [35]	ECCV 2020	31.54	20.41	9.51	5.29	0.84	61.27	35.54
LAVT [38]	CVPR 2022	51.84	30.27	17.34	9.52	2.09	71.86	47.40
RIS-DMMI [40]	ICCV 2023	63.89	44.30	19.81	6.49	1.00	68.58	52.15
CrossVLT [42]	IEEE TMM	71.16	58.28	34.51	16.35	5.06	77.44	58.84
LGCE [8]	IEEE TGRS 2024	73.75	61.14	39.46	16.02	5.45	76.81	59.96
RMSIN [9]	CVPR 2024	79.20	65.99	42.98	16.51	3.25	75.72	62.58
FIANet [10]	IEEE TGRS 2025	84.09	77.05	61.86	33.41	7.10	78.32	68.67
EVF-SAM [51]	arXiv 2024	35.17	22.34	9.36	2.86	0.39	55.51	36.64
RS2-SAM2 [18]	AAAI 2026	84.31	79.42	70.89	55.70	21.19	80.87	73.90
RSRefSeg 2 [19]	IEEE TGRS 2026	**88.22**	82.99	**73.97**	60.92	34.40	81.24	77.39
UniRS-Seg (ours)	This work	87.43	**83.12**	72.65	**62.41**	**35.12**	**81.65**	**78.18**

Table 6. Referring segmentation results on the RRSIS-D test set (%). Best values are shown in bold.

Method	Publication	Pr@0.5	Pr@0.6	Pr@0.7	Pr@0.8	Pr@0.9	cIoU	gIoU
CMPC [36]	CVPR 2020	55.83	47.40	36.94	25.45	9.19	69.22	49.24
BRINet [34]	CVPR 2020	56.90	48.77	39.12	27.03	8.73	69.88	49.65
LSCM [35]	ECCV 2020	56.02	46.25	37.70	25.28	8.27	69.05	49.92
LAVT [38]	CVPR 2022	69.52	63.63	53.29	41.60	24.94	77.19	61.04
RIS-DMMI [40]	ICCV 2023	68.74	60.96	50.33	38.38	21.63	76.20	60.12
CrossVLT [42]	IEEE TMM	70.38	63.83	52.86	42.11	25.02	76.32	61.00
LGCE [8]	IEEE TGRS 2024	67.65	61.53	51.45	39.62	23.33	76.34	59.37
RMSIN [9]	CVPR 2024	74.26	67.25	55.93	42.55	24.53	77.79	64.20
FIANet [10]	IEEE TGRS 2025	74.46	66.96	56.31	42.83	24.13	76.91	64.01
EVF-SAM [51]	arXiv 2024	72.16	66.50	56.59	43.92	25.48	76.77	62.75
RS2-SAM2 [18]	AAAI 2026	77.56	72.34	61.76	47.92	29.73	78.99	66.72
RSRefSeg 2 [19]	IEEE TGRS 2026	**80.23**	75.78	65.41	**50.65**	**31.05**	79.45	69.17
UniRS-Seg (ours)	This work	78.44	**76.87**	**67.05**	48.61	30.01	**80.32**	**70.88**

Table 7. Referring segmentation results on the RISBench test partition (%). Baseline results follow RSRefSeg 2 [19]. UniRS-Seg is trained using the official RISBench training partition. Best values among the listed methods are bold.

Method	Publication	Pr@0.5	Pr@0.6	Pr@0.7	Pr@0.8	Pr@0.9	cIoU	gIoU
LGCE [8]	TGRS 2024	69.64	64.07	56.26	44.92	25.74	73.87	62.13
RMSIN [9]	CVPR 2024	71.01	65.46	57.69	45.50	25.92	74.09	63.07
CroBIM-Swin [57]	arXiv 2024 (v1)	75.75	70.34	63.12	51.12	28.45	73.61	67.32
CroBIM-ConvNeXt [57]	arXiv 2024 (v1)	77.55	72.83	66.38	55.93	34.07	73.04	69.33
LSCF [60]	TGRS 2025	76.08	71.29	64.96	55.13	36.73	**74.88**	68.53
RSRefSeg 2 [19]	TGRS 2026	**79.08**	75.36	**70.62**	62.75	48.75	74.77	72.57
UniRS-Seg (ours)	This work	78.94	**75.51**	70.43	**63.01**	**49.01**	74.85	**72.91**

4.6. Ablation Design

All ablations use the same data splits, seed set, training budget, and input resolution as the primary model unless the varied factor is explicitly named. Architecture components are selected on LoveDA and RRSIS-D validation sets and subsequently checked on Vaihingen and RefSegRS, avoiding test-set selection. The confidence study compares fixed $\beta \in \{0, 0.25, 0.5, 1, 2\}$ with the full model's learnable confidence at $\epsilon = 10^{-6}$; $\beta = 0$ removes the mask prior while retaining ordinary cross-attention. The LoveDA augmentation control varies the training augmentation policy. These studies retain the benchmark splits and evaluation metrics described above.

Cumulative module control. Table 8 isolates the contribution of each architectural group. Variant A receives deterministic class or expression prompts but contains no oracle points or boxes. Adding SAM adapters and the CNN/CBA pathway improves both tasks. The

sparse–dense interface produces the largest single gain on RRSIS-D (5.34 gIoU points from C to D), while the decoder-attention group and hierarchical decoder provide further improvements. The LMCAM/CMAttn column enables LMCAM for both tasks and CMAttn for semantic samples only, following Section 3.6.3. Joint training raises the final LoveDA mIoU from 56.98% to 57.30% and RRSIS-D gIoU from 70.01% to 70.88%.

Table 8. Component ablation on LoveDA and RRSIS-D.

Variant	Adapter	CNN/CBA	SigLIP2 SPI	LMCAM/CMAttn	HPD	Joint	LoveDA mIoU	RRSIS-D gIoU
A: frozen SAM decoder	–	–	–	–	–	–	50.05	60.76
B	✓	–	–	–	–	–	53.87	62.43
C	✓	✓	–	–	–	–	54.91	62.98
D	✓	✓	✓	–	–	–	55.12	68.32
E	✓	✓	✓	✓	–	–	56.21	68.76
F	✓	✓	✓	✓	✓	–	56.98	70.01
G: full UniRS-Seg	✓	✓	✓	✓	✓	✓	57.30	70.88

Sparse and dense prompting. Table 9 compares sparse, dense, and joint prompting on RefSegRS. The prompt ablation distinguishes semantic identity from spatial support. Sparse and dense prompts together outperform either input alone. Adding LMCAM to the joint prompt increases RefSegRS gIoU from 77.32% to 78.18%, cIoU from 80.12% to 81.65%, and Pr@0.9 from 34.22% to 35.12%, indicating that soft iterative spatial guidance improves both mean overlap and strict-mask accuracy.

Table 9. Sparse–dense prompt ablation on RefSegRS.

Prompt Design	Pr@0.5	Pr@0.6	Pr@0.7	Pr@0.8	Pr@0.9	cIoU	gIoU
Sparse query only	85.67	81.98	71.03	61.55	34.21	79.12	76.35
Dense activation only	86.02	82.00	71.12	61.43	34.02	79.53	76.54
Sparse + dense	86.61	82.31	71.77	62.00	34.22	80.12	77.32
Sparse + dense + LMCAM	87.43	83.12	72.65	62.41	35.12	81.65	78.18

Joint and task-specific training. Table 10 evaluates the effect of shared training under equal optimizer-step counts. The semantic-only and referring-only controls retain the SAM ViT-B/SigLIP 2 backbone pair, prompt pathway, decoder, input sizes, optimizer settings, augmentation policy, and applicable mask losses; the available task supervision changes. The two task-specific models have independent copies of the corresponding architecture, with no weight sharing or cross-task consistency. This is an accuracy control, not an equal-total-parameter deployment comparison: storing both models duplicates the shared pathways. Equal total optimizer steps also do not imply equal task-specific exposure under alternated joint batches. Relative to semantic-only training, the full model gains 0.45 and 0.89 mIoU points on LoveDA and Vaihingen; relative to referring-only training it gains 1.40 and 0.97 gIoU points on RefSegRS and RRSIS-D. The full model exceeds the two task-specific models by 0.83 and 1.19 mIoU points on LoveDA and Vaihingen and by 1.94 and 1.76 gIoU points on RefSegRS and RRSIS-D. Consistency supervision also improves every dataset relative to task-balanced sampling alone, with gains ranging from 0.12 to 0.67 points.

Table 10. Joint-training and single-task controls under the fixed benchmark splits and common UniRS-Seg backbone/decoder setting. Metrics are percentages. Dashes indicate that a task is not trained or evaluated by that control.

Training Regime	LoveDA mIoU	Vaihingen mIoU	RefSegRS gIoU	RRSIS-D gIoU
Two task-specific models	56.47	84.12	76.24	69.12
Shared, semantic only	56.85	84.42	–	–
Shared, referring only	–	–	76.78	69.91
Shared, task-balanced sampling	57.03	85.10	78.06	70.21
Shared + consistency (full)	57.30	85.31	78.18	70.88

4.7. Reproducibility

Each experiment stores the dataset version and split, checkpoint identifiers, input processors, class prompt templates, random seed, optimizer groups, learning-rate schedule, loss weights, and software versions. Semantic label mappings are stored in query order, including ignore indices. Referring records retain the exact expression, tokenizer revision, original image size, padding transform, and inverse transform. Metrics are computed from saved per-sample masks rather than copied from training logs.

5. Results

5.1. Semantic Segmentation Performance

The semantic comparison in Tables 2, 3 and 4 includes the source-reported WASP-Mamba comparison [29], with the published STAR-Net result [59] added to Table 2. In Table 2, Bac., Bui., Roa., Wat., Bar., For., and Agr. denote background, building, road, water, barren land, forest, and agriculture, respectively. In Tables 3 and 4, Imp., Bui., Low., Tre., and Car. denote impervious surface, building, low vegetation, tree, and car; the mean scores average these five foreground categories, and the Potsdam comparison uses RGB input. On LoveDA, UniRS-Seg achieves 57.30% mIoU, exceeding the strongest listed baseline, STAR-Net (56.62%), by 0.68 percentage points and WASP-Mamba by 3.40 points. It obtains the highest listed IoU in five of the seven classes. Relative to the best baseline value in each of these columns, the gains are 2.33 points for background, 2.04 for building, 3.10 for road, 1.79 for water, and 3.01 for agriculture. STAR-Net performs better on barren land and forest, scoring 37.77% and 55.15% versus 23.01% and 51.08% for UniRS-Seg. The overall mIoU advantage therefore coexists with class-dependent weaknesses, while the road and agriculture gains support the value of complementary spatial and semantic features. Figure 6 provides the corresponding qualitative comparison.

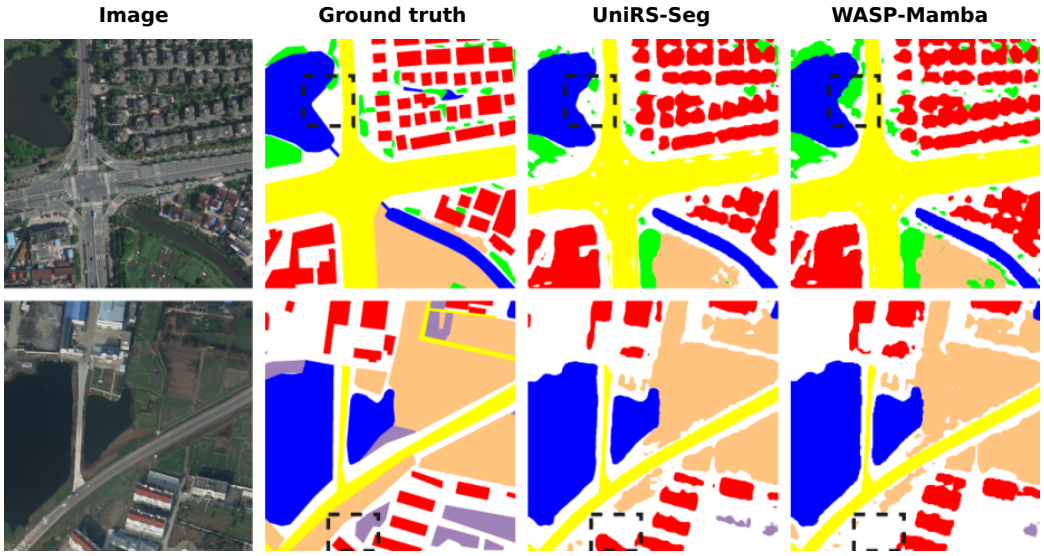


Figure 6. Qualitative semantic segmentation on LoveDA. Columns show image, ground truth, UniRS-Seg, and WASP-Mamba. Label colors are shared across predictions and ground truth; dashed boxes mark the original regions of interest. Residual errors from the upper scene are examined in Figure 10.

On Vaihingen, UniRS-Seg obtains 85.31% mIoU, 91.91% mF1, and 94.22% OA (Table 3), improving the strongest listed values by 0.99, 0.60, and 0.70 points, respectively. It provides the highest IoU for impervious surfaces, buildings, low vegetation, and cars, while its tree IoU is 1.45 points below the strongest baseline. On Potsdam, the model reaches 88.24%

mIoU, 93.63% mF1, and 92.07% OA (Table 4). These values exceed the strongest listed mean scores by 1.18, 0.68, and 0.16 points. The model leads four of the five class IoUs; its car IoU is 0.14 points below the best listed result. Figures 7 and 8 show representative predictions from the two ISPRS benchmarks. In both montages, the shared class palette exposes boundary disagreement around buildings and vegetation, while the dashed boxes emphasize cars and other compact structures for which the CNN/CBA and hierarchical decoder are intended to preserve detail.

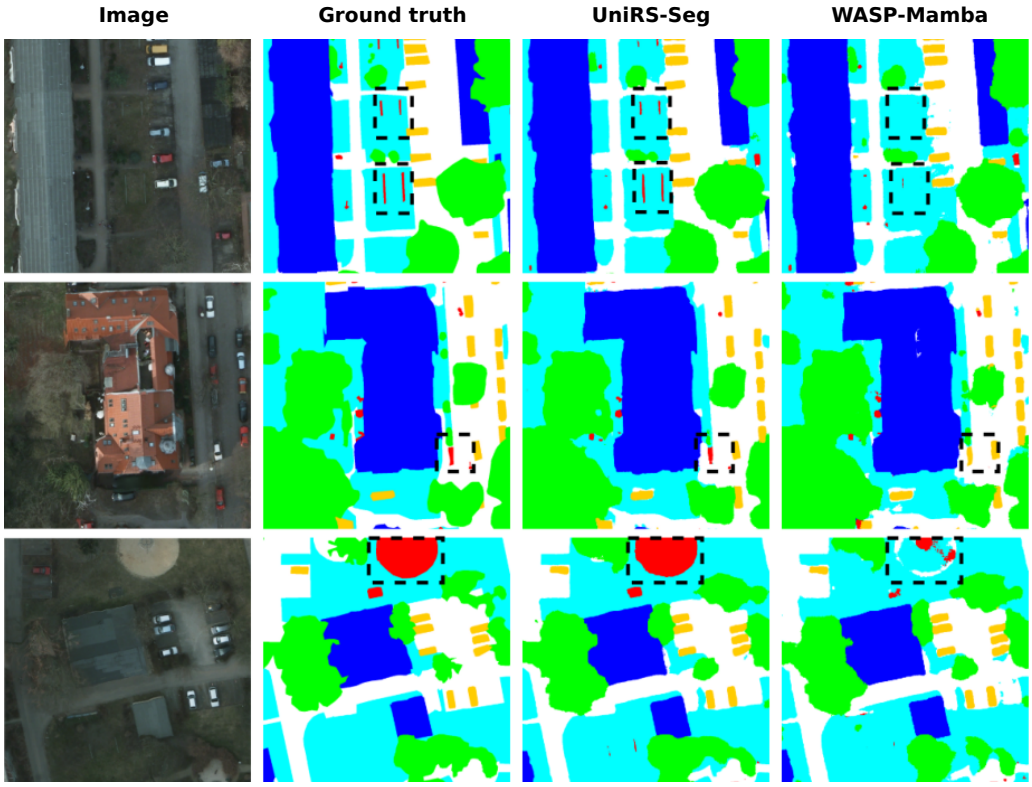


Figure 7. Qualitative semantic segmentation on ISPRS Potsdam. Columns show image, ground truth, UniRS-Seg, and WASP-Mamba. A common label palette supports boundary and small-object comparison. Dashed boxes highlight regions for comparing boundary and small-target delineation.

5.2. Referring Remote Sensing Segmentation Performance

Tables 5 and 6 use the source-reported comparisons from RSRefSeg 2 [19]; Pr@x denotes the percentage of test samples whose mask IoU exceeds threshold x. On RefSegRS, Table 5 shows a clear progression from early cross-modal methods to remote-sensing and foundation-model approaches. Early methods report 31.51–40.63% gIoU, followed by 58.84% for CrossVLT, 62.58% for RMSIN, and 68.67% for FIANet. Incorporating stronger foundation-model priors raises the reported value to 73.90% for RS2-SAM2 and 77.39% for RSRefSeg 2. UniRS-Seg reaches 78.18% gIoU and 81.65% cIoU, improving the strongest listed baseline by 0.79 and 0.41 points, respectively. Its cIoU–gIoU gap is 3.47 points, compared with 3.85 points for RSRefSeg 2 and 6.97 points for RS2-SAM2. Because cIoU weights samples in proportion to their union area whereas gIoU weights every expression equally, the simultaneous gains indicate that the improvement is retained under both size-weighted and sample-balanced aggregation. The precision profile is nevertheless non-uniform. Relative to RSRefSeg 2, UniRS-Seg changes Pr@0.5, Pr@0.6, Pr@0.7, Pr@0.8, and Pr@0.9 by -0.79, $+0.13$, -1.32, $+1.49$, and $+0.72$ points, respectively. Thus, the principal advantage appears at the stricter 0.8 and 0.9 thresholds rather than as dominance at every cutoff.

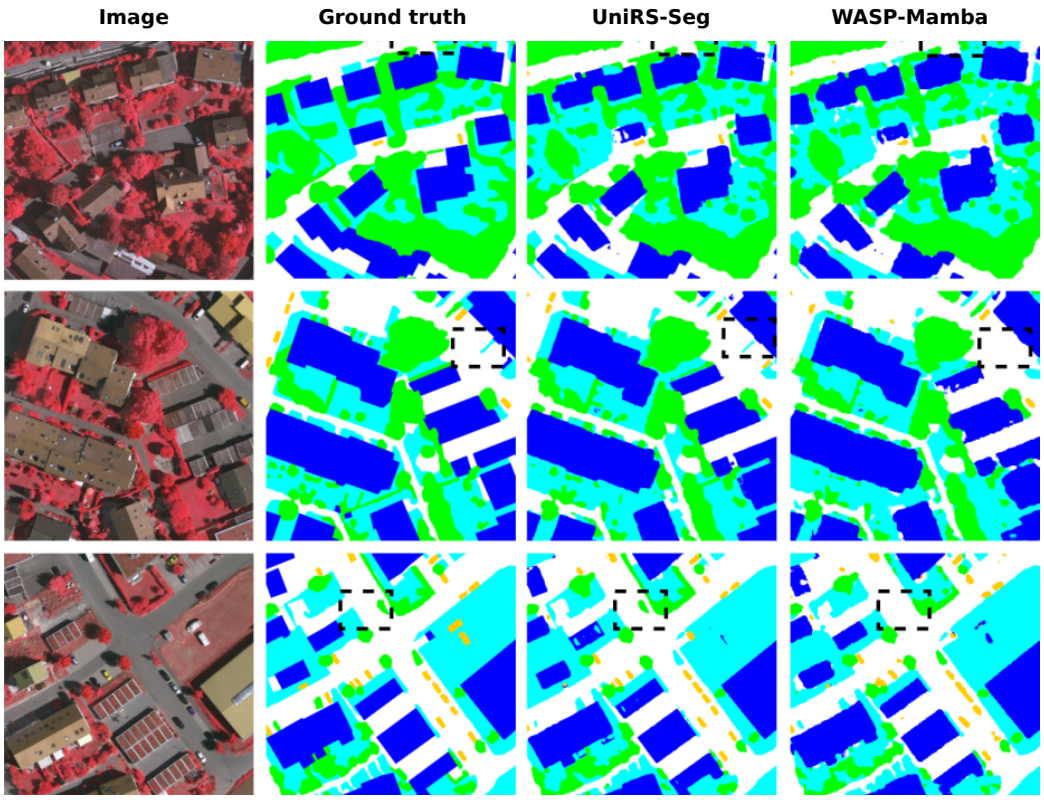


Figure 8. Qualitative semantic segmentation on ISPRS Vaihingen. Columns show image, ground truth, UniRS-Seg, and WASP-Mamba. A common label palette supports boundary and small-object comparison. Dashed boxes highlight regions for comparing boundary and small-target delineation.

RRSIS-D presents a complementary test because it contains a wider range of target sizes, ground resolutions, and object orientations. As reported in Table 6, the early CMPC, BRINet, and LSCM baselines remain near 49–50% gIoU, whereas LAVT, RIS-DMMI, CrossVLT, LGCE, RMSIN, FIANet, and EVF-SAM occupy the 59.37–64.20% range. RS2-SAM2 advances to 66.72%, and RSRefSeg 2 reaches 69.17%. UniRS-Seg obtains 70.88% gIoU and 80.32% cIoU, which are 1.71 and 0.87 points above the strongest listed values. The cIoU–gIoU gap decreases from 12.27 points for RS2-SAM2 and 10.28 points for RSRefSeg 2 to 9.44 points for UniRS-Seg, which is consistent with less concentration of the aggregate score in larger regions, although it does not by itself establish scale invariance. The threshold results again require a qualified interpretation. Compared with RSRefSeg 2, UniRS-Seg improves Pr@0.6 from 75.78% to 76.87% and Pr@0.7 from 65.41% to 67.05%, corresponding to gains of 1.09 and 1.64 points. Conversely, it is lower by 1.79, 2.04, and 1.04 points at Pr@0.5, Pr@0.8, and Pr@0.9. The higher gIoU and cIoU therefore reflect a favorable change in the overall overlap distribution, particularly around the moderate thresholds, rather than a uniform increase in the number of near-perfect masks.

Taken together, the two benchmarks show that the unified query formulation remains competitive under different expression and scale distributions. UniRS-Seg is the only entry in the reported tables that attains the highest gIoU and cIoU on both datasets, but its threshold behavior differs: RefSegRS shows its clearest gains at Pr@0.8 and Pr@0.9, whereas RRSIS-D shows gains at Pr@0.6 and Pr@0.7 and small regressions at the strictest cutoffs. This distinction is important because an aggregate IoU can improve through many moderate corrections even when the count of almost perfectly segmented samples does not. Figure 9 supports the corresponding qualitative diagnosis by placing RSRefSeg 2 and UniRS-Seg predictions beside the same ground truth. Both methods delineate the stadium and chimney accurately, whereas the relation-dependent overpass and airplane examples

reveal incorrect extent or instance selection in the comparison column and closer agreement by UniRS-Seg. These examples are consistent with the intended division of labor between language-aligned localization and SAM-based delineation, although the quantitative tables, rather than selected visual examples, provide the primary evidence. Since the baseline entries are source reported rather than independently rerun, all differences in this subsection should be read as descriptive comparisons under the cited protocols; paired significance tests or per-category uncertainty estimates are not available for the published baseline predictions.

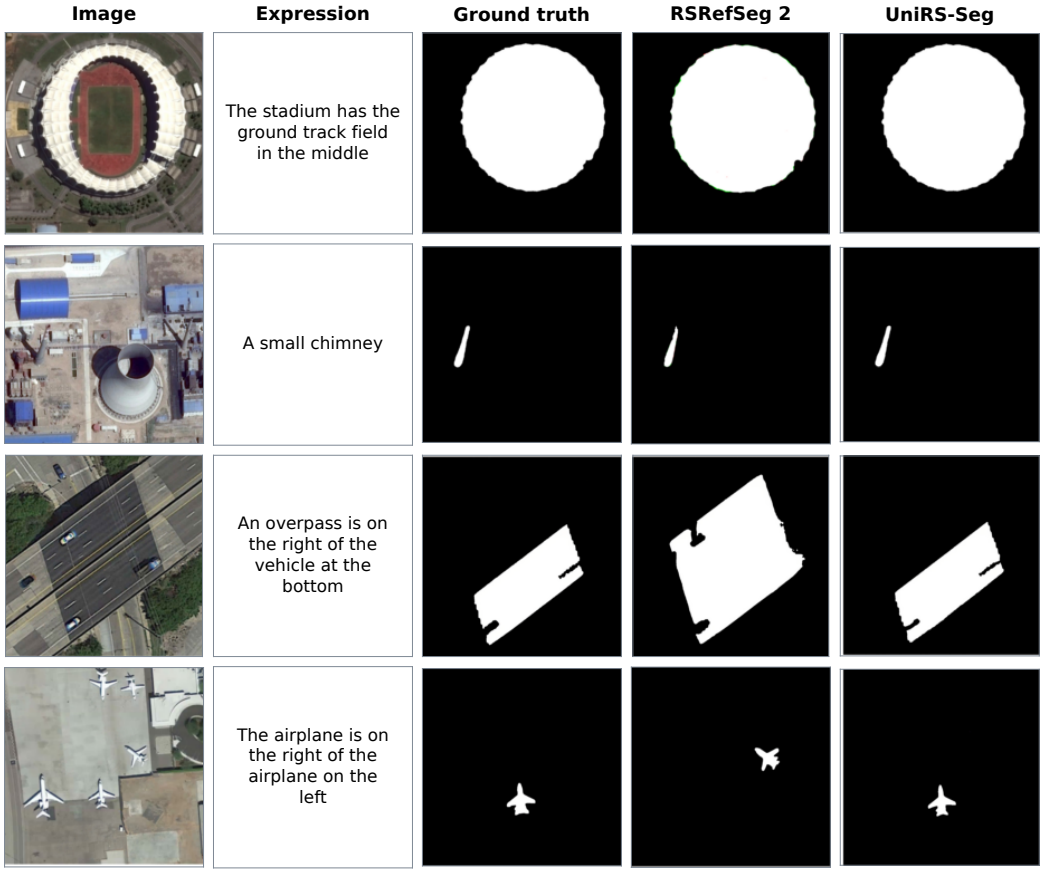


Figure 9. Qualitative referring segmentation. Columns show the image, expression, ground truth, RSRefSeg 2, and UniRS-Seg. The four examples depict a stadium, small chimney, relation-dependent overpass, and relation-dependent airplane. White denotes target and black denotes background in all masks. The shared display format supports direct comparison of target selection and mask extent.

5.3. Additional Evaluation on RISBench

Table 7 extends the referring evaluation to RISBench. UniRS-Seg obtains 72.91% gIoU, exceeding RSRefSeg 2 by 0.34 percentage points and CroBIM-ConvNeXt by 3.58 points. Its 74.85% cIoU is 0.08 points above RSRefSeg 2 and 0.03 points below LSCF, so the highest sample-averaged overlap does not imply the highest cumulative overlap. Relative to RSRefSeg 2, Pr@0.6, Pr@0.8, and Pr@0.9 increase by 0.15, 0.26, and 0.26 points, respectively; Pr@0.5 and Pr@0.7 decrease by 0.14 and 0.19 points. The performance profile is therefore competitive across thresholds while remaining non-uniform, consistent with the two primary referring comparisons.

RISBench broadens the evaluation to a larger expression–mask collection with more categories and attribute types than RefSegRS or RRSIS-D. The additional result supports the applicability of the shared query-conditioned architecture under another supervised benchmark protocol. Because RISBench training samples are used, this experiment evaluates

benchmark coverage rather than zero-shot transfer to an unseen domain. It complements the original joint-training controls without attributing a RISBench-specific gain to task sharing, for which no separate RISBench single-task ablation is reported.

5.4. Ablation Results

Domain adaptation and local detail. Table 8 isolates the architectural additions in a cumulative sequence; a checkmark indicates that the corresponding component is enabled. CNN/CBA denotes the local convolutional and cross-branch attention pathway, SPI denotes the SigLIP 2 sparse–dense prompt interface, LMCAM/CMAttn denotes the decoder-attention group described above, and HPD denotes the hierarchical pixel decoder. CMAttn is active only for semantic samples, whereas LMCAM is shared by both task modes. Starting from Variant A, the full configuration raises LoveDA mIoU from 50.05% to 57.30% and RRSIS-D gIoU from 60.76% to 70.88%, for total gains of 7.25 and 10.12 points. SAM adapters account for the first 3.82-point LoveDA gain and 1.67-point RRSIS-D gain, showing that lightweight domain adaptation is important before cross-modal fusion. Adding CNN/CBA provides a further 1.04 and 0.55 points, respectively, which supports the use of an explicit local-detail pathway alongside SAM's token representation.

Vision–language prompting. The largest task-specific change occurs from Variant C to D: introducing the SigLIP 2 SPI adds only 0.21 LoveDA points but 5.34 RRSIS-D points. This asymmetric result is expected because the referring task directly depends on language-conditioned localization, whereas LoveDA receives a fixed semantic query bank.

Soft attention and hierarchical reconstruction. Enabling the decoder-attention group then contributes 1.09 LoveDA points and 0.44 RRSIS-D points; the RRSIS-D change is attributable to LMCAM because CMAttn is bypassed in referring mode. HPD contributes 0.77 and 1.25 points, and joint training contributes the final 0.32 and 0.87 points. The semantic attention gain combines LMCAM and CMAttn, so this experiment does not isolate CMAttn alone; the RefSegRS prompt study separately isolates LMCAM. The positive increments in both columns indicate complementarity among domain adaptation, local detail, prompt construction, mask-conditioned fusion, and multi-scale decoding; because the variants are cumulative, however, these increments should not be interpreted as order-independent main effects.

Prompt composition and LMCAM. The prompt and training studies clarify which parts of the unified interface produce these gains. In Table 9, dense activation alone slightly exceeds the sparse query alone in cIoU and gIoU (79.53% versus 79.12% cIoU and 76.54% versus 76.35% gIoU), but neither representation is sufficient by itself. Combining them raises the scores to 80.12% cIoU and 77.32% gIoU, gains of 0.59 and 0.78 points over the stronger single-prompt result. LMCAM then increases the joint design by another 1.53 cIoU points and 0.86 gIoU points, reaching 81.65% and 78.18%. Relative to the sparse-only variant, the complete prompt pathway also raises Pr@0.5 by 1.76 points and Pr@0.9 by 0.91 points, suggesting benefits at both permissive and strict overlap criteria.

Task sharing and consistency. Table 10 evaluates unification under equal optimizer-step counts. The "two task-specific models" row uses separate networks with no shared weights, whereas the "shared" rows use the same architecture and parameter budget as UniRS-Seg. The full shared model improves over the two-model setting by 0.83 and 1.19 mIoU points on LoveDA and Vaihingen and by 1.94 and 1.76 gIoU points on RefSegRS and RRSIS-D. Compared with single-task training of the shared architecture, task-balanced training with consistency supervision gains 0.45 and 0.89 semantic points and 1.40 and 0.97 referring points. Consistency supervision itself provides smaller but uniform improvements over task-balanced sampling alone: 0.27, 0.21, 0.12, and 0.67 points across the four datasets. These results are consistent with positive transfer under the reported training budget. They

do not isolate encoder sharing from decoder sharing, and equal-step joint and single-task runs differ in the mixture of task exposures.

Parameter-efficient adaptation. Table 11 finally examines parameter-efficient tuning while all other components are held fixed within that controlled setting. With both foundation encoders frozen, the configuration obtains 79.04% mIoU on Vaihingen, 82.22% mIoU on Potsdam, 50.05% mIoU on LoveDA, and 60.76% gIoU on RRSIS-D. Adding a bottleneck adapter to SAM raises these values to 81.99%, 84.54%, 52.98%, and 61.65%, confirming that adapting mask geometry is useful across both task families. SAM LoRA with rank 16 produces a similar semantic result on Vaihingen and Potsdam, a higher LoveDA score of 54.91%, and a comparable RRSIS-D score of 61.54%; thus, neither SAM-only tuning method dominates on every dataset. The selected combination of a SAM adapter and SigLIP 2 LoRA achieves the best value in all four columns: 82.76%, 86.11%, 54.98%, and 67.43%. Relative to the fully frozen setting, the gains are 3.72, 3.89, 4.93, and 6.67 points, respectively. More importantly for referring segmentation, tuning the vision–language pathway raises RRSIS-D by 5.78 points over the SAM-adapter-only configuration, whereas the corresponding semantic gains are 0.77, 1.57, and 2.00 points. This pattern agrees with the role assigned to SigLIP 2 in the architecture: its adaptation primarily improves alignment between remote-sensing imagery and language queries, while the SAM pathway supplies the mask prior and boundary refinement. The absolute scores in this table belong to the controlled adaptation study and are not the final full-model results, so they should be used to compare tuning choices rather than to replace the benchmark values in Tables 2–4.

Table 11. Parameter-efficient adaptation ablation.

SAM Tuning	SigLIP2 Tuning	Vaihingen mIoU	Potsdam mIoU	LoveDA mIoU	RRSIS-D gIoU
Frozen	Frozen	79.04	82.22	50.05	60.76
Adapter (1:4)	Frozen	81.99	84.54	52.98	61.65
LoRA ($r = 16$)	Frozen	81.87	84.31	54.91	61.54
Adapter (1:4)	LoRA ($r_v = 16, r_t = 8$)	82.76	86.11	54.98	67.43

LMCAM confidence sensitivity. Table 12 evaluates the mask-confidence parameter in Equation (10) on LoveDA and RefSegRS. With $\beta = 0$, the prior is disabled and LMCAM reduces to ordinary cross-attention. Increasing fixed confidence from 0 to 1 raises LoveDA mIoU from 56.43% to 57.05%, RefSegRS gIoU from 77.32% to 77.92%, and cIoU from 80.12% to 80.98%. Sharpening the prior to $\beta = 2$ lowers these scores by 0.17, 0.22, and 0.28 points relative to $\beta = 1$. This pattern is consistent with excessive reliance on a potentially incorrect spatial prior, although aggregate scores do not directly establish the cause of individual localization errors. Learnable confidence performs best among the evaluated settings, reaching 57.30% mIoU, 78.18% gIoU, and 81.65% cIoU. It exceeds the strongest fixed setting by 0.25, 0.26, and 0.67 points, respectively. The RefSegRS endpoints reproduce the "Sparse + dense" and "Sparse + dense + LMCAM" values in Table 9. The attention-odds column is an analytical illustration of Equation (11); the segmentation metrics are measured results. This comparison supports adaptive confidence within the tested range, without establishing a globally optimal initialization or stabilizer.

Table 12. LMCAM confidence sensitivity on LoveDA and RefSegRS at $\epsilon = 10^{-6}$. $\beta = 0$ disables the mask prior; learnable β is the full-model setting. LoveDA reports mIoU and RefSegRS reports gIoU/cIoU. Attention odds are analytical illustrations for $M_a = 0.8$, $M_b = 0.2$, and equal feature logits (Equation (11)), rather than measured attention statistics.

β Setting	Interpretation	Attention Odds a_{ja}/a_{jb}	LoveDA mIoU (%)	RefSegRS gIoU (%)	RefSegRS cIoU (%)
$\beta = 0$	Mask prior disabled	1.00	56.43	77.32	80.12
$\beta = 0.25$	Weak preference	1.41	56.72	77.55	80.45
$\beta = 0.5$	Soft preference	2.00	56.91	77.78	80.78
$\beta = 1.0$	Proportional preference	4.00	57.05	77.92	80.98
$\beta = 2.0$	Sharpened preference	16.00	56.88	77.70	80.70
Learnable β	**Adaptive (full model)**	–	**57.30**	**78.18**	**81.65**

Effect of data augmentation. Table 13 compares UniRS-Seg on LoveDA with and without the training augmentation policy described in Section 4.2. Augmentation raises mIoU from 57.13% to 57.30%, a modest 0.17-point improvement. Background IoU increases by 1.04 points, while each of the six remaining classes improves by 0.01–0.03 points. Thus, most of the aggregate gain comes from the heterogeneous background category; the comparison supports the stated policy but does not indicate a large improvement for every land-cover class. The augmented row matches the primary LoveDA result in Table 2. Geometric and photometric transformations broaden training variation while preserving semantic labels, and the conservative referring policy preserves directional language. This LoveDA comparison evaluates the combined policy, without separating individual transforms or measuring its effect on referring segmentation or corruption robustness.

Table 13. Semantic segmentation results on LoveDA (%) with and without the training augmentation policy. Both rows use UniRS-Seg with SAM ViT-B. The augmented row is the primary configuration reported in Table 2; bold denotes the higher value in each metric.

Training Augmentation	Backbone	Bac.	Bui.	Roa.	Wat.	Bar.	For.	Agr.	mIoU
Without augmentation	SAM ViT-B	49.99	63.10	61.84	83.63	22.98	51.05	67.30	57.13
With augmentation	SAM ViT-B	**51.03**	**63.12**	**61.87**	**83.65**	**23.01**	**51.08**	**67.31**	**57.30**

5.5. Robustness Analysis

Controlled image corruptions. We evaluate the trained full UniRS-Seg model on LoveDA and RefSegRS under Gaussian noise, resolution degradation, and illumination changes (Table 14). The model components, training configuration, and evaluation protocol remain those of the main experiments; only the evaluation images are corrupted. Mild, moderate, and severe Gaussian noise use standard deviations $\sigma \in \{0.05, 0.10, 0.20\}$ on RGB intensities scaled to $[0, 1]$. Resolution degradation resizes images to $\{0.75, 0.50, 0.25\}$ of their original width and height, respectively, and then bilinearly upsamples them to the original size. Illumination changes multiply pixel intensities by $\{0.8, 1.2\}$, $\{0.6, 1.4\}$, and $\{0.4, 1.6\}$ at the three severities; each reported illumination score is the arithmetic mean of the separate dimming and brightening evaluations. Labels and referring expressions remain unchanged. The clean row reproduces the main LoveDA mIoU and RefSegRS gIoU results.

Performance under increasing severity. Scores decrease monotonically with severity for all three corruption types. Illumination changes have the smallest effect: the severe setting yields 55.19% LoveDA mIoU and 76.15% RefSegRS gIoU, decreases of 2.11 and 2.03 percentage points from clean performance. Severe Gaussian noise reduces the scores to 54.16% and 75.23%, corresponding to decreases of 3.14 and 2.95 points. Resolution degradation causes the largest losses, reaching 51.24% and 72.08% at the severe setting, or 6.06 and 6.10 points below the clean results. Nevertheless, these scores retain 89.4% and

92.2% of their respective clean values. These percentages describe retained overlap scores, rather than the proportion of unaffected objects or pixels.

Interpretation and remaining limits. The greater sensitivity to resolution loss is consistent with removal of the fine spatial evidence needed to delineate small targets, thin structures, and boundaries. The smaller illumination-related drops are compatible with the photometric variation seen during training. Local CNN features, adapted SAM geometry, language-aligned localization, soft mask guidance, and hierarchical reconstruction provide complementary processing pathways, but this full-model study does not isolate their individual effects under corruption. The measured results support tolerance to the specified perturbations within the two evaluated datasets; they do not establish superiority over corruption-tested baselines, robustness to combined corruptions, or transfer to unseen sensors and scenes. Resolution loss remains the most damaging tested condition, and fully occluded regions are outside the model's visible-mask objective.

Table 14. Robustness of UniRS-Seg under controlled image corruptions. LoveDA reports mIoU (%) and RefSegRS reports gIoU (%). Severity settings are defined in Section 5.5; illumination scores average the paired dimming and brightening evaluations. Clean performance is included for reference.

Corruption	Severity	LoveDA mIoU (%)	RefSegRS gIoU (%)
Clean	—	57.30	78.18
Gaussian noise	Mild	56.84	77.91
	Moderate	55.92	76.97
	Severe	54.16	75.23
Resolution degradation	Mild	56.73	77.65
	Moderate	54.81	75.74
	Severe	51.24	72.08
Illumination change	Mild	57.05	77.96
	Moderate	56.38	77.34
	Severe	55.19	76.15

6. Discussion

6.1. Unified Query Interface

The results support the central formulation: semantic segmentation can be represented by a bank of class-name queries, whereas referring segmentation can be represented by one expression query. Once both are mapped to the same decoder space, the difference between the tasks lies primarily in query construction, output normalization, and supervision. The training-regime ablation is important because it separates this architectural claim from a simple increase in model capacity. Under the reported setting, the shared model improves both semantic and referring metrics relative to two task-specific networks.

The sparse–dense interface also provides an interpretable separation between localization and delineation. The dense prompt records where the vision–language pathway finds compatible evidence, while the sparse token identifies what the decoder should extract. A localization error appears in the dense activation map before boundary refinement; a geometry error appears when the activation is correct but the final contour is inaccurate. This separation is useful for analyzing small targets, repeated instances, and relation-heavy expressions.

6.2. Role of SigLIP 2 and SAM

SigLIP 2 and SAM contribute different priors. SigLIP 2 supplies a multilingual image–text embedding space and unpooled patch features whose pretraining explicitly emphasizes localization and dense prediction [15]. SAM supplies prompt-conditioned mask geometry and strong boundary priors [11]. The query compiler and prompt interface align these

representations without assuming that their token spaces or spatial resolutions are directly interchangeable. The marked RRSIS-D gain obtained when the sparse–dense interface is added is consistent with this division of labor.

Combining a vision–language encoder with SAM is not itself the novelty of UniRS-Seg. RSRefSeg, EVF-SAM, RS2-SAM2, and RSRefSeg 2 already use foundation models for language-guided segmentation [17–19,51]. The contribution here is the task-level interface: one class-or-expression query compiler, one generalized mask decoder, and joint training across semantic and referring annotations. The present evaluation supports unified supervised segmentation, but it does not establish open-vocabulary or zero-shot segmentation on unseen categories.

6.3. Correlation and Complementarity of the Two Tasks

Both tasks estimate regions in the same visual scene, so they share category evidence, local contours, scale cues, and contextual structure. Their supervision nevertheless differs. Semantic masks teach exhaustive class coverage, including background and adjacent land-cover transitions. Referring masks teach which instance or region satisfies attributes and relations while excluding other regions of the same class. In a dense settlement, class supervision helps distinguish roof from vegetation, while referring supervision distinguishes the roof next to a road from neighboring roofs. Class-union consistency connects only the compatible category-level outputs; it does not equate a semantic union with a relation-specific instance mask.

This complementarity is relevant to small targets and ambiguous boundaries. Semantic supervision supplies many labeled boundary pixels even when referring masks are sparse. Referring supervision rewards selective retrieval amid repeated targets and complex backgrounds. The joint-training improvements over the single-task shared-architecture controls are 0.45/0.89 semantic points and 1.40/0.97 referring points (Table 10). These aggregate results support complementary supervision in the evaluated setting, but do not identify which scene type or target-size group benefits most. CMAttn addresses semantic frequency imbalance only; the shared LMCAM and reconstruction pathway carry the common spatial processing.

6.4. Language Grounding in Remote Sensing Scenes

Description diversity and background selection. The referring compiler combines pooled expression semantics with token-level modifiers before retrieving image context (Equation (6)). This provides a route for category, size, appearance, and directional cues to influence the same mask query. Cosine matching reduces feature-norm bias, the dense prompt supplies candidate spatial support, and LMCAM reweights it using the current mask. These operations can reduce irrelevant background aggregation; however, attention can still favor a visually similar distractor if language grounding is wrong. The four referring examples in Figure 9 illustrate size and relational variation, rather than exhaustive linguistic robustness or multilingual grounding.

Scale diversity, occlusion, and heterogeneity. The coarse SigLIP 2 grid offers language-aligned context, while intermediate SAM states, local CNN features, and hierarchical fusion retain finer spatial evidence. This division explains why a small chimney and a broad stadium can use the same prompt interface. Domain adapters and dataset-conditioned class queries provide trainable responses to overhead appearance and taxonomy differences. They do not make the fixed-resolution language grid scale invariant. Severe occlusion removes observable contours, and shadows or vegetation may obscure the target even when its category is understood. The model predicts visible benchmark masks; it has

neither an amodal reconstruction objective nor an explicit mechanism for recovering fully hidden objects.

Qualitative successes and residual failures. Figure 9 shows correct selection of a compact chimney and relation-dependent overpass/airplane targets in the shown panels. Figure 10 complements these successes with residual errors taken from the LoveDA prediction: roof outlines lose regularity, and a narrow canal has mismatched width and contact with adjacent land cover. The prediction–ground-truth differences expose incomplete contour recovery and foreground extent errors. They illustrate the distinction between recognizing a region and reconstructing its contour. The lower Vaihingen tree IoU and lower strict-threshold RRSIS-D precision relative to the strongest listed baseline likewise show that aggregate gains do not eliminate class- or difficulty-dependent failures.

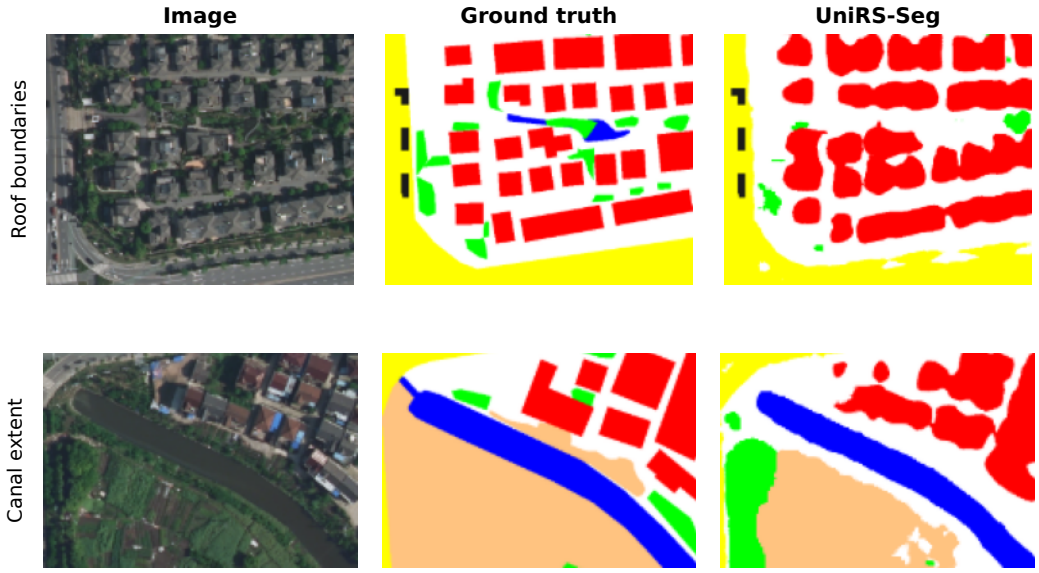


Figure 10. Residual UniRS-Seg errors from the LoveDA qualitative montage (Figure 6, upper scene). Each row shows the same RGB crop, ground truth, and UniRS-Seg prediction. The roof crop exposes softened/irregular outlines; the canal crop exposes extent and boundary disagreement. Matched crop coordinates preserve spatial correspondence among the three columns.

6.5. Evaluation Scope and Generalization

The six benchmarks cover urban tiles, urban/rural land cover, and expression-guided aerial targets across different resolutions and taxonomies. The additional RISBench experiment extends supervised dataset coverage (Section 5.3); it is distinct from transfer without target-dataset training. RefSegRS preserves its scene-disjoint split, and LoveDA contains both urban and rural imagery, but the reported scores are aggregate.

The controlled corruption study in Section 5.5 quantifies Gaussian-noise, resolution-degradation, and illumination sensitivity on LoveDA and RefSegRS. It shows progressive degradation within these benchmark distributions, with resolution loss producing the largest reduction among the tested conditions. These measurements complement the clean-benchmark evaluation but do not establish unseen-domain transfer or the corruption-specific contribution of each module. The present studies also do not include a loss-weight search, a controlled shared-versus-task-specific-encoder factorial experiment, or occlusion-stratified metrics. The cumulative module ablation and single-task controls address narrower questions and should be interpreted within that scope.

6.6. Limitations

The dual-foundation-image-pathway design is computationally demanding (Table 1). Parameter-efficient tuning reduces optimization cost but leaves both image encoders active during inference. Cached class queries remove repeated semantic text encoding; smaller vision–language checkpoints, distillation, and token reduction could reduce deployment cost. These alternatives are not evaluated here. LMCAM depends on the quality of its coarse spatial prior, and the grouped semantic-attention ablation does not independently separate CMAttn from LMCAM. The confidence sweep compares five fixed values with learned confidence on LoveDA and RefSegRS; initialization and ϵ sensitivity remain unevaluated. The augmentation control is limited to LoveDA and does not isolate individual transformations.

The corruption study covers three separate perturbation types at three severities on two datasets; combined corruptions and matched robustness comparisons with other models remain unevaluated. Severe resolution degradation still lowers LoveDA mIoU and RefSegRS gIoU by 6.06 and 6.10 percentage points. The input is three-channel optical imagery; multispectral, SAR, hyperspectral, and temporal transfer remain outside the present evaluation. The referring benchmarks are English-only, so multilingual pretraining does not demonstrate multilingual geospatial grounding. Severe occlusion, very small objects, ambiguous referring expressions, and scene-domain shifts can still produce incorrect masks. The work establishes supervised dual-task decoding and measured transfer in the reported benchmark experiments, with no claim of universal scene robustness or unseen-class recognition.

7. Conclusions

We presented UniRS-Seg, a supervised framework that connects full-scene semantic parsing and expression-guided remote sensing segmentation through a shared sparse–dense query interface and hierarchical mask decoder. Class-vocabulary queries and referring expressions use compatible decoder representations, while adapted SAM geometry, local CNN detail, and SigLIP 2 alignment supply complementary evidence. Across three semantic and three referring benchmarks, the model achieves the highest listed mIoU/gIoU values, including the additional supervised RISBench evaluation. Joint training improves the single-task shared-architecture controls on all four datasets in the original training-regime study. The prompt ablation separately supports the contribution of soft mask cross-attention, while cumulative component results reveal the importance of domain adaptation and language-aligned localization. The confidence study favors learnable guidance over the tested fixed values, and augmentation provides a modest LoveDA improvement. These findings support dual-task weight sharing under the reported supervised protocols. The controlled corruption study further measures tolerance to noise, resolution degradation, and illumination changes on LoveDA and RefSegRS, with resolution loss causing the largest degradation. These results do not establish universal corruption robustness or unseen-scene transfer. Smaller shared models and explicit evaluations of scene shift, difficult targets, and augmentation across additional datasets are useful directions for extending the framework.

Author Contributions: Conceptualization, A.T.; methodology, A.T.; software, A.T.; resources, H.H.; writing—original draft preparation, A.T.; writing—review and editing, A.T. and H.H.; visualization, H.H.; supervision, H.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: LoveDA, ISPRS Vaihingen, ISPRS Potsdam, RefSegRS, RRSIS-D, and RISBench are publicly available from their respective benchmark providers, subject to the original licenses and access conditions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36. https://doi.org/10.1109/mgrs.2017.2762307.
2. Wang, L.; Li, R.; Zhang, C.; Fang, S.; Duan, C.; Meng, X.; Atkinson, P.M. UNetFormer: A UNet-Like Transformer for Efficient Semantic Segmentation of Remote Sensing Urban Scene Imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 196–214. https://doi.org/10.1016/j.isprsjprs.2022.06.008.
3. Ma, X.; Zhang, X.; Pun, M.O. RS3Mamba: Visual State Space Model for Remote Sensing Image Semantic Segmentation. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 6011405. https://doi.org/10.1109/lgrs.2024.3414293.
4. Zhu, E.; Chen, Z.; Wang, D.; Shi, H.; Liu, X.; Wang, L. UNetMamba: An Efficient UNet-Like Mamba for Semantic Segmentation of High-Resolution Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2025**, *22*, 6001205.https://doi.org/10.1109/lgrs.2024.3505193.
5. Wang, L.; Li, D.; Dong, S.; Meng, X.; Zhang, X.; Hong, D. PyramidMamba: Rethinking Pyramid Feature Fusion with Selective Space State Model for Semantic Segmentation of Remote Sensing Imagery. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *144*, 104884. https://doi.org/10.1016/j.jag.2025.104884.
6. Tong, Q.; Zhu, Z.; Zhang, M.; Cao, K.; Xing, H. CrossFormer Embedding DeepLabv3+ for Remote Sensing Images Semantic Segmentation. *Comput. Mater. Contin.* **2024**, *79*, 1353–1375. https://doi.org/10.32604/cmc.2024.049187.
7. Zhu, P.; Yan, J.; Wang, E.; Hou, Y.; Liu, J.; Zhou, L.; Luo, X. Advancements in Semisupervised Remote Sensing Segmentation Using Adaptive Patch Augmentation and Class Ranking Weight. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5645418. https://doi.org/10.1109/tgrs.2025.3615725.
8. Yuan, Z.; Mou, L.; Hua, Y.; Zhu, X.X. RRSIS: Referring Remote Sensing Image Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5613312.https://doi.org/10.1109/tgrs.2024.3369720.
9. 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, Seattle, WA, USA, 16–22 June 2024; pp. 26648–26658. https://doi.org/10.1109/cvpr52733.2024.02517.
10. 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 Trans. Geosci. Remote Sens.* **2025**, *63*, 5604611. https://doi.org/10.1109/tgrs.2024.3522293.
11. 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, Paris, France, 1–6 October 2023; pp. 3992–4003. https://doi.org/10.1109/iccv51070.2023.00371.
12. Chen, T.; Zhu, L.; Deng, C.; Cao, R.; Wang, Y.; Zhang, S.; Li, Z.; Sun, L.; Zang, Y.; Mao, P. SAM-Adapter: Adapting Segment Anything in Underperformed Scenes. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Paris, France, 2–6 October 2023; pp. 3367–3375.
13. Ren, S.; Luzi, F.; Lahrichi, S.; Kassaw, K.; Collins, L.M.; Bradbury, K.; Malof, J.M. Segment Anything, From Space? In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2024; pp. 8340–8350. https://doi.org/10.1109/wacv57701.2024.00817.
14. 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 Trans. Geosci. Remote Sens.* **2024**, *62*, 5622216. https://doi.org/10.1109/TGRS.2024.3390838.
15. 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* **2025**, arXiv:2502.14786.
16. Guo, X.; Lao, J.; Dang, B.; Zhang, Y.; Yu, L.; Ru, L.; Zhong, L.; Huang, Z.; Wu, K.; Hu, D.; et al. SkySense: A Multi-Modal Remote Sensing Foundation Model Towards Universal Interpretation for Earth Observation Imagery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 27662–27673. https://doi.org/10.1109/cvpr52733.2024.02613.
17. Chen, K.; Zhang, J.; Liu, C.; Zou, Z.; Shi, Z. RSRefSeg: Referring Remote Sensing Image Segmentation with Foundation Models. In Proceedings of the 2025 IEEE International Geoscience and Remote Sensing Symposium, Brisbane, Australia, 3–8 August 2025; pp. 1070–1074. https://doi.org/10.1109/igarss55030.2025.11243338.

18. Rong, F.; Lan, M.; Zhang, Q.; Zhang, L. RS2-SAM2: Customized SAM2 for Referring Remote Sensing Image Segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, Singapore, 20–27 January 2026; Volume 40, pp. 8751–8759. https://doi.org/10.1609/aaai.v40i11.37828.
19. Chen, K.; Liu, C.; Chen, B.; Zhang, J.; Zou, Z.; Shi, Z. RSRefSeg 2: Decoupling Referring Remote Sensing Image Segmentation with Foundation Models. *IEEE Trans. Geosci. Remote Sens.* **2026**, *64*, 4700420. https://doi.org/10.1109/tgrs.2025.3647535.
20. Ni, S.; Wang, D.; Chen, H.; Guo, H.; Zhang, N.; Zhang, J. UniGeoSeg: Towards Unified Open-World Segmentation for Geospatial Scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Denver, CO, USA, 3–7 June 2026; pp. 34865–34876.
21. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440. https://doi.org/10.1109/cvpr.2015.7298965.
22. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.
23. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587.
24. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 9992–10002. https://doi.org/10.1109/iccv48922.2021.00986.
25. Wang, L.; Li, R.; Wang, D.; Duan, C.; Wang, T.; Meng, X. Transformer Meets Convolution: A Bilateral Awareness Network for Semantic Segmentation of Very Fine Resolution Urban Scene Images. *Remote Sens.* **2021**, *13*, 3065. https://doi.org/10.3390/rs13163065.
26. Wang, L.; Li, R.; Duan, C.; Zhang, C.; Meng, X.; Fang, S. A Novel Transformer Based Semantic Segmentation Scheme for Fine-Resolution Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6506105. https://doi.org/10.1109/lgrs.2022.3143368.
27. Li, L.; Yi, J.; Fan, H.; Lin, H. A Lightweight Semantic Segmentation Network Based on Self-Attention Mechanism and State Space Model for Efficient Urban Scene Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 4703215. https://doi.org/10.1109/tgrs.2025.3562185.
28. Mu, J.; Zhou, S.; Sun, X. PPMamba: Enhancing Semantic Segmentation in Remote Sensing Imagery by SS2D. *IEEE Geosci. Remote Sens. Lett.* **2025**, *22*, 6001705. https://doi.org/10.1109/lgrs.2024.3507033.
29. Zhang, Y.; Zhang, G.; Li, Y.; Wu, Y.; Zhou, Z. WASP-Mamba: A Wavelet-Enhanced Mamba Framework for Remote Sensing Semantic Segmentation. *Remote Sens.* **2026**, *18*, 2188. https://doi.org/10.3390/rs18132188.
30. Zou, X.; Li, Y.; Zhang, S.; Li, K.; Wang, S.; Tao, P.; Xing, J.; Lang, C. Dynamic Dictionary Learning for Remote Sensing Image Segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 19–25 October 2025; pp. 22457–22466.
31. Zhang, Y.; Luo, Z.; Yang, B.; Chen, Z.; Zou, Y. Spatially-adaptive dynamic dictionary learning for fine-grained remote sensing image segmentation. *Sci. Rep.* **2026**, ahead of print. https://doi.org/10.1038/s41598-026-63993-z.
32. Li, R.; Li, K.; Kuo, Y.C.; Shu, M.; Qi, X.; Shen, X.; Jia, J. Referring Image Segmentation via Recurrent Refinement Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5745–5753. https://doi.org/10.1109/cvpr.2018.00602.
33. Ye, L.; Rochan, M.; Liu, Z.; Wang, Y. Cross-Modal Self-Attention Network for Referring Image Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10494–10503. https://doi.org/10.1109/cvpr.2019.01075.
34. 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, Seattle, WA, USA, 13–19 June 2020; pp. 4423–4432. https://doi.org/10.1109/cvpr42600.2020.00448.
35. Hui, T.; Liu, S.; Huang, S.; Li, G.; Yu, S.; Zhang, F.; Han, J. Linguistic Structure Guided Context Modeling for Referring Image Segmentation. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 59–75. https://doi.org/10.1007/978-3-030-58607-2_4.
36. 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, Seattle, WA, USA, 13–19 June 2020; pp. 10485–10494. https://doi.org/10.1109/cvpr42600.2020.01050.

37. Liu, S.; Hui, T.; Huang, S.; Wei, Y.; Li, B.; Li, G. Cross-Modal Progressive Comprehension for Referring Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 4761–4775. https://doi.org/10.1109/tpami.2021.3079993.
38. 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, New Orleans, LA, USA, 18–24 June 2022; pp. 18134–18144. https://doi.org/10.1109/cvpr52688.2022.01762.
39. 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, New Orleans, LA, USA, 18–24 June 2022; pp. 11676–11685. https://doi.org/10.1109/cvpr52688.2022.01139.
40. 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, Paris, France, 1–6 October 2023; pp. 4044–4054. https://doi.org/10.1109/iccv51070.2023.00376.
41. 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, Ottawa, ON, Canada, 29 October–3 November 2023; pp. 779–788. https://doi.org/10.1145/3581783.3612117.
42. Cho, Y.; Yu, H.; Kang, S.J. Cross-Aware Early Fusion with Stage-Divided Vision and Language Transformer Encoders for Referring Image Segmentation. *IEEE Trans. Multimed.* **2024**, *26*, 5823–5833. https://doi.org/10.1109/tmm.2023.3340062.
43. Azimi, S.M.; Henry, C.; Sommer, L.; Schumann, A.; Vig, E. SkyScapes: Fine-Grained Semantic Understanding of Aerial Scenes. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 7392–7402. https://doi.org/10.1109/iccv.2019.00749.
44. Zhan, Y.; Xiong, Z.; Yuan, Y. RSVG: Exploring Data and Models for Visual Grounding on Remote Sensing Data. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5604513. https://doi.org/10.1109/tgrs.2023.3250471.
45. Lu, Q.; Xie, Y.; Zhang, J.; Guo, Y.; Wei, Y.; Jiang, J.; Luan, X. CLIP-Driven with Dynamic Feature Selection and Alignment Network for Referring Remote Sensing Image Segmentation. *Remote Sens.* **2025**, *17*, 3675. https://doi.org/10.3390/rs17223675.
46. Zhao, Z.; Xu, X.; Huang, B.; Chen, H.; Pu, F. RSAM: Vision-Language Two-Way Guidance for Referring Remote Sensing Image Segmentation. *Remote Sens.* **2025**, *17*, 3960. https://doi.org/10.3390/rs17243960.
47. Dong, S.; Xie, J.; Chen, L.; Chen, H.; Qi, B.; Ge, Y. Structurally Consistent and Grounding-Aware Stagewise Reasoning for Referring Remote Sensing Image Segmentation. *Remote Sens.* **2026**, *18*, 1015. https://doi.org/10.3390/rs18071015.
48. Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; de Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; Gelly, S. Parameter-Efficient Transfer Learning for NLP. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; Volume 97, pp. 2790–2799.
49. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. In Proceedings of the International Conference on Learning Representations, Virtual, 25–29 April 2022.
50. 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 Proceedings of the 38th International Conference on Machine Learning, Virtual, 18–24 July 2021; Volume 139, pp. 8748–8763.
51. Zhang, Y.; Cheng, T.; Zhu, L.; 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* **2024**, arXiv:2406.20076.
52. Cheng, B.; Misra, I.; Schwing, A.G.; Kirillov, A.; Girdhar, R. Masked-Attention Mask Transformer for Universal Image Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1280–1289. https://doi.org/10.1109/cvpr52688.2022.00135.
53. Gong, Z.; Wei, Z.; Wang, D.; Hu, X.; Ma, X.; Chen, H.; Jia, Y.; Deng, Y.; Ji, Z.; Zhu, X.; et al. CrossEarth: Geospatial Vision Foundation Model for Domain Generalizable Remote Sensing Semantic Segmentation. *arXiv* **2024**, arXiv:2410.22629v3.
54. 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, Paris, France, 1–6 October 2023; pp. 11975–11986.
55. Wang, J.; Zheng, Z.; Ma, A.; Lu, X.; Zhong, Y. LoveDA: A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation. In Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, Virtual, 6–14 December 2021; Volume 1. https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/4e732ced3463d06de0ca9a15b6153677-Paper-round2.pdf.
56. ISPRS Working Group III/4. ISPRS 2D Semantic Labeling Benchmark: Vaihingen and Potsdam. 2026. Available online: https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/semantic-labeling.aspx (accessed on 7 October 2026).
57. Dong, Z.; Sun, Y.; Gu, Y.; Liu, T. Cross-Modal Bidirectional Interaction Model for Referring Remote Sensing Image Segmentation. *arXiv* **2024**, arXiv:2410.08613v1.
58. Li, Y.; Liu, G.; Bai, Y.; Zheng, C. Rotation Equivariant Convolution-Based Network for Semantic Segmentation of High Spatial Resolution Remote Sensing Imagery. *IEEE Geosci. Remote Sens. Lett.* **2026**, *23*, 6004705. https://doi.org/10.1109/lgrs.2026.3662255.

59. Yang, B.; Xu, P.; Han, H.; Wu, C.; Tang, J.; Feng, Z.; Cao, C.; Qiao, L. STAR-Net: Dual-Encoder Network with Global-Local Fusion for Agricultural Land Cover Parsing. *Remote Sens.* **2026**, *18*, 1314. https://doi.org/10.3390/rs18091314.
60. 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 Trans. Geosci. Remote Sens.* **2025**, *63*, 5628313. https://doi.org/10.1109/TGRS.2025.3578515.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.