SFC-SAM3: SPECIALIZING FOUNDATION PRIORS VIA CROSS-MODAL COORDINATION FOR REFERRING IMAGE SEGMENTATION IN AGRICULTURAL REMOTE SENSING

Anonymous authors
Paper under double-blind review

ABSTRACT

Referring image segmentation (RIS) grounds free-form language in pixel-level masks, but extending it to agricultural remote sensing exposes a substantial mismatch between generic foundation-model priors and domain-specific visual and linguistic structure. Agricultural scenes contain large and weakly bounded regions, while referring expressions combine phenology, morphology, and spatial relations. To address these challenges, we introduce SFC-SAM3, a parameter-efficient foundation-prior framework for agricultural referring remote sensing image segmentation (RRSIS) that specializes complementary structural, visual, and linguistic priors. Rather than retraining foundation backbones, SFC-SAM3 coordinates SAM3, DINOv3, and CLIP through a frequency-aware PromptGenerator, a multi-layer DINOv3 FPN bridge, a CLIP token-level text bridge, and a residual mask refinement head, together with LoRA residuals on the SAM3 encoder. Training follows two stages: joint fine-tuning of the adapters and SAM3 mask decoder, followed by parameter-efficient continued training with the decoder frozen, reducing the trainable parameters from 6.58M (0.54%) in Stage 1 to 4.28M (0.35%) in Stage 2. On FarmSeg-VL, SFC-SAM3 reaches 87.04% mIoU, compared with 86.11% for RMSIN trained on the same data. The advantage becomes clearer off the training distribution: without training on the target datasets, SFC-SAM3 reaches 41.14% mIoU on the RRSIS-D land subset and 19.81% on the RefSegRS land/ground subset, compared with 5.65% and 13.99%, respectively, for the strongest evaluated baselines. These results point to a practical lesson for agricultural RRSIS: task-structured specialization of complementary foundation priors can preserve strong in-domain performance while substantially improving transfer to unseen land-oriented remote-sensing scenes.

1 INTRODUCTION

Agriculture covers a third of Earth's habitable land and supports over a billion livelihoods, making language-driven farmland monitoring important for food security and precision agriculture. Referring remote sensing image segmentation (RRSIS) asks a model to segment the region described by a free-form expression, e.g., the irrigated wheat plot adjacent to the riverbank. Agricultural RRSIS is particularly challenging because farmland targets are often large and weakly bounded, while expressions combine crop state, attributes, and spatial relations. FarmSeg-VL Tao et al. (2025) exposes these challenges through farmland masks paired with dense contextual descriptions.

Current RRSIS work either adapts referring image segmentation (RIS) encoders to aerial imagery or wraps segmentation around large vision language models. LAVT Yang et al. (2022) and RMSIN Liu et al. (2024b) represent the first route, with RMSIN adding intra-/cross-scale interaction and rotated convolutions on RRSIS-D. LISA Lai et al. (2024), GLaMM Rasheed et al. (2024), GeoChat Kuckreja et al. (2024), and RSRefSeg Chen et al. (2025; 2026) move toward MLLM or CLIP–SAM hybrids, gaining language capacity but often at high cost and with coarse boundaries. Meanwhile, foundation models such as SAM/SAM2/SAM3 Kirillov et al. (2023); Ravi et al. (2024); Carion et al. (2026), DINOv2/v3 Oquab et al. (2024); Siméoni et al. (2025), and RemoteCLIP Liu et al. (2024a) provide

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

strong priors, but off-the-shelf foundation models remain poorly matched to agricultural RRSIS, where crop-state descriptions, texture-poor fields, and spatial grounding remain challenging. Adapter works such as SAM-Adapter Chen et al. (2023) and Earth-Adapter Hu et al. (2026) show that these priors can be redirected with small modules, but existing designs do not jointly cover cross-modal grounding, multi-layer spatial context and boundary fidelity for dominant agricultural land regions.

We present SFC-SAM3, a specialised foundation-prior framework for agricultural RRSIS. It coordinates three largely frozen backbones, SAM3, DINOv3 and CLIP, through four lightweight adapters: a frequency-aware PromptGenerator, a multi-layer DINO–FPN bridge over DINOv3 layers $\{6, 12, 18, 24\}$, a text-conditioned prompt module using CLIP token features, and a residual mask refinement head, together with LoRA on the SAM3 image encoder. Training follows a two-stage curriculum on FarmSeg-VL: joint adapter–decoder fine-tuning, followed by parameter-efficient continued training with the decoder frozen, reducing the trainable parameters from 6.58M (0.54%) in Stage 1 to 4.28M (0.35%) in Stage 2.

On the FarmSeg-VL test set, SFC-SAM3 reaches 87.04% mIoU, modestly surpassing RMSIN trained on the same data (86.11%). The advantage is larger under cross-domain evaluation, where SFC-SAM3 reaches 41.14% and 19.81% mIoU on the RRSIS-D and RefSegRS subsets, respectively, compared with 5.65% and 13.99% for the strongest evaluated baselines. These results suggest that carefully specialising foundation priors can preserve in-domain accuracy while improving transfer across land remote-sensing scenes.

Our contributions are threefold:

- We introduce SFC-SAM3, a specialised foundation-prior framework for agricultural RRSIS that integrates frequency, multi-scale visual, token-level language and boundary priors through lightweight adapters.
- We develop a two-stage parameter-efficient training recipe that reaches 87.04% mIoU on FarmSeg-VL, with 6.58M (0.54%) trainable parameters in Stage 1 and 4.28M (0.35%) in Stage 2.
- We show substantially improved cross-domain transfer on land-related RRSIS-D and RefSegRS subsets.

2 Related Work

2.1 Referring Image Segmentation

RIS segments the target specified by a free-form language expression. Recent progress is largely driven by vision–language transformers: LAVT Yang et al. (2022) injects pixel–word attention into Swin stages, CRIS Wang et al. (2022) transfers CLIP knowledge through text-to-pixel contrastive learning, ReferFormer Wu et al. (2022) and VLT Ding et al. (2023) explore query-based formulations, and GRES/ReLA Liu et al. (2023) extends RIS to multi-target and no-target expressions. A parallel trend uses large language models for generative segmentation, including LISA Lai et al. (2024), GLaMM Rasheed et al. (2024), and PixelLM Ren et al. (2024). X-Decoder Zou et al. (2023) and MaskCLIP Zhou et al. (2022) further show the value of unified segmentation–VL decoding and CLIP's latent pixel signal. These methods are developed mainly for internet imagery, and transfer imperfectly to remote sensing, where scale, viewpoint and vocabulary differ substantially.

2.2 Vision–Language Understanding for Remote Sensing

RRSIS extends RIS to aerial imagery. RefSegRS Yuan et al. (2024) formalises the task, while RMSIN Liu et al. (2024b) scales it with the 17K-image RRSIS-D benchmark and introduces intra-/cross-scale interaction with adaptive rotated convolutions. RSRefSeg Chen et al. (2025) and RSRefSeg 2 Chen et al. (2026) combine CLIP-aligned features with a SAM decoder to decouple grounding from mask generation. Broader remote-sensing vision–language work includes RemoteCLIP Liu et al. (2024a), RingMo Sun et al. (2023), GeoChat Kuckreja et al. (2024), and SkyEyeGPT Zhan et al. (2025). Closer to agriculture, AgriCLIP Nawaz et al. (2025) specialises CLIP for agricultural imagery, CrossEarth Gong et al. (2024) studies cross-domain RS segmentation, and FarmSeg-VL Tao et al. (2025) introduces referring-style farmland captions. Despite these advances, existing methods are not

specialised end-to-end for agricultural RRSIS, whose large land regions, fine-grained descriptions and fuzzy boundaries require a different design.

2.3 Frequency-Domain Adapters and Boundary-Aware Losses

Two smaller lines of work motivate our adapter and training design. First, frequency-domain adaptation is useful in remote sensing because spatial encoders often smooth periodic field textures. Earth-Adapter Hu et al. (2026) shows that frequency-guided adapters can improve geospatial PEFT, which motivates the FFT branch in our PromptGen-FFT module. Second, agricultural masks are highly boundary-sensitive: regional losses such as Dice or BCE alone often miss fuzzy, elongated field contours. Lovász loss Berman et al. (2018) provides an mIoU-oriented surrogate, while boundary loss Kervadec et al. (2021) directly penalises contour errors. We use these ideas as lightweight training priors rather than separate model components.

3 Method

3.1 Overview

We address agricultural RRSIS: given an aerial image $\mathbf{I} \in \mathbb{R}^{3\times H\times W}$ and a free-form expression E, the model predicts a binary mask $\hat{\mathbf{M}} \in \{0,1\}^{H\times W}$ for the referred region. Compared with natural-image RIS, agricultural RRSIS is harder because targets often dominate the scene, expressions mix category, phenology, attributes and spatial relations, and crop boundaries are typically fuzzy.

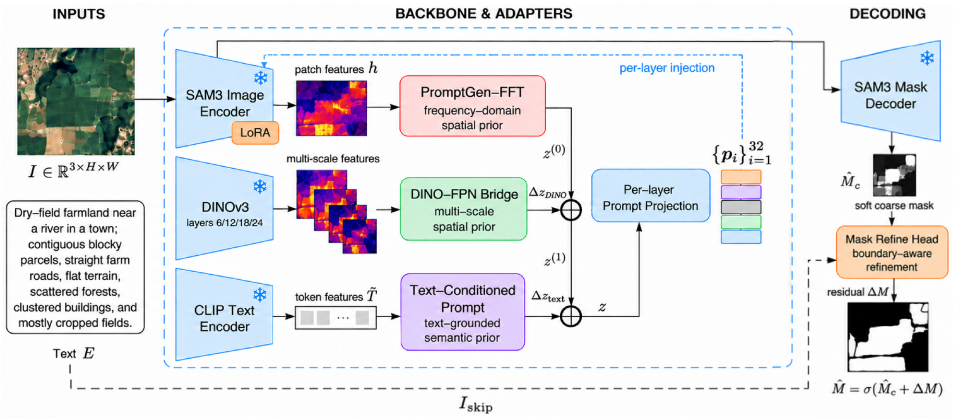

Figure 1: **Overview of SFC-SAM3.** Image $\mathbf{I}$ and expression E are encoded by three foundation backbones. PromptGen-FFT initializes the shared prompt representation, which is sequentially refined by the DINO–FPN Bridge and Text-Conditioned Prompt module before per-layer projection into the SAM3 image encoder. The SAM3 decoder produces coarse-mask logits $\hat{\mathbf{M}}_c$, and the Mask Refine Head predicts an image-conditioned residual $\Delta\mathbf{M}$ via $\mathbf{I}_{\text{skip}}$, yielding $\hat{\mathbf{M}}$.

Rather than retraining a foundation model, we coordinate three largely frozen backbones—SAM3 Carion et al. (2026), DINOv3 Siméoni et al. (2025) and CLIP Radford et al. (2021)—with lightweight adapters. Our design follows three empirical observations: SAM3 mainly needs a better prompt, DINOv3 carries useful semantics at multiple intermediate layers, and CLIP's pooled text feature loses fine-grained agricultural descriptors. We therefore specialize three complementary prompt-side priors—frequency-aware structural, multi-scale visual, and token-level linguistic cues—and complement them with boundary-aware residual mask refinement. Figure 1 summarises the pipeline. The image and expression are encoded by the SAM3 image encoder with LoRA, DINOv3 layers $\{6, 12, 18, 24\}$, and a frozen CLIP text encoder. PromptGen-FFT (§3.2) initializes a shared prompt representation, which is sequentially refined by the DINO–FPN Bridge (§3.3) and the Text-Conditioned Prompt module (§3.4). The resulting representation is projected into layer-specific prompts and injected into the SAM3 image encoder. The SAM3 decoder outputs coarse-mask

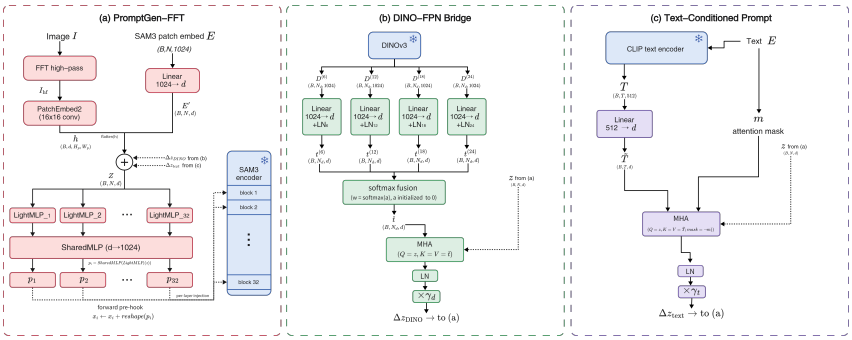

Figure 2: Internal structure of the three cross-modal adapters. (a) PromptGen-FFT fuses the FFT prior and SAM3 patch embeddings into a shared prompt, followed by DINO/text refinement and projection into $L = 32$ per-block prompts. (b) DINO–FPN Bridge softmax-fuses four DINOv3 layers into $\Delta\mathbf{z}_{\text{DINO}}$. (c) Text-Conditioned Prompt injects CLIP token-level features through padding-masked cross-attention as $\Delta\mathbf{z}_{\text{text}}$.

logits $\hat{\mathbf{M}}_c$, and the Mask Refine Head predicts an image-conditioned residual $\Delta\mathbf{M}$ through the skip connection $\mathbf{I}_{\text{skip}}$, giving $\hat{\mathbf{M}} = \sigma(\hat{\mathbf{M}}_c + \Delta\mathbf{M})$.

3.2 Frequency-Aware Prompt Generation

Purely learned prompts easily overfit on the limited-scale RRSIS training set, whereas agricultural fields bear repeated high-frequency boundary textures—furrows, tillage lines, plot borders—that are data-agnostic structural cues. Inspired by SAM-Adapter Chen et al. (2023) and frequency-domain adapters for remote sensing Hu et al. (2026), we pair the learned stream with a handcrafted Fourier high-pass prior that regularises the adapter from the first iteration.

Let $\mathbf{I} \in \mathbb{R}^{B\times 3\times H\times W}$ denote the input batch. The two-branch layout is sketched in Figure 2(a). We compute its Fourier high-pass residual with a centred rectangular low-pass mask $\mathbf{M}_r$ of relative area $r = 0.25$,

$$\mathbf{I}_{\text{hf}} = \left|\mathcal{F}^{-1}\big((1 - \mathbf{M}_r) \odot \mathcal{F}(\mathbf{I})\big)\right|, \tag{1}$$

and embed it with a strided 16×16 convolution into a bottleneck map $\mathbf{h} \in \mathbb{R}^{B\times d\times H_p\times W_p}$ with $d = 128$. In parallel, the SAM3 patch embeddings $\mathbf{E} \in \mathbb{R}^{B\times N\times 1024}$ ($N = H_pW_p$) are projected to the same bottleneck via a single linear layer, $\mathbf{E}' = \text{Linear}_{1024\to d}(\mathbf{E})$. The two branches are first fused additively:

$$\mathbf{z}^{(0)} = \text{flatten}(\mathbf{h}) + \mathbf{E}' \in \mathbb{R}^{B\times N\times d}. \tag{2}$$

SAM3's image encoder comprises $L = 32$ transformer blocks whose attention statistics differ layer by layer. After the sequential DINO and text refinement described in Sections 3.3–3.4, the resulting shared representation $\mathbf{z}$ is projected into layer-specific prompts. To respect the layer-wise heterogeneity without paying a full per-layer parameter cost, we instantiate L independent lightweight MLPs followed by a single shared up-projection,

$$\mathbf{c}_i = \text{LightMLP}_i(\mathbf{z}), \qquad \mathbf{p}_i = \text{SharedMLP}(\mathbf{c}_i), \qquad i = 1, \ldots, L, \tag{3}$$

yielding per-layer prompts $\mathbf{p}_i \in \mathbb{R}^{B\times N\times 1024}$. Each $\mathbf{p}_i$ is added to the input of block i through a `forward_pre_hook`, $\mathbf{x}_i \leftarrow \mathbf{x}_i + \text{reshape}(\mathbf{p}_i)$, without directly modifying the pretrained SAM3 encoder weights.

The $8\times$ bottleneck ($d = 128$, compressing SAM3's 1024-dimensional embedding) keeps the prompt generator lightweight, while the layer-specific LightMLPs let different encoder blocks receive distinct residuals before a shared up-projection. Since these prompts are injected additively through hooks, the branch acts as a bias generator on top of the pretrained encoder rather than a weight modifier. Together with the FFT prior, this design gives the adapter a stable low-cost route for learning agricultural boundary cues.

3.3 MULTI-LAYER DINOV3 FPN BRIDGE

A single DINOv3 layer encodes semantics at one granularity: early layers capture local geometry, while late layers compress global class abstraction Siméoni et al. (2025). Agricultural referents demand both at once—fine field boundaries and coarse land-cover categories—so relying on a single layer underuses the representation. Following the spirit of FPN Lin et al. (2017a), we aggregate four evenly spaced DINOv3 layers into one multi-scale bridge and route it into the PromptGen-FFT bottleneck via cross-attention.

Given a frozen DINOv3 ViT-L/16, we perform one forward pass with hidden states exposed and harvest patch-token sequences $\mathbf{D}^{(\ell)} \in \mathbb{R}^{B\times N_d\times 1024}$ from the four layers $\mathcal{L} = \{6, 12, 18, 24\}$. Figure 2(b) summarises the bridge. Each level is projected to the shared bottleneck and LayerNormed independently,

$$\mathbf{t}^{(\ell)} = \mathrm{LN}_\ell\big(\mathrm{Linear}_{\ell,\,1024\to d}(\mathbf{D}^{(\ell)})\big), \qquad \ell \in \mathcal{L}, \tag{4}$$

then combined by a learnable softmax fusion:

$$\bar{\mathbf{t}} = \sum_{\ell\in\mathcal{L}} w_\ell\, \mathbf{t}^{(\ell)}, \qquad \mathbf{w} = \mathrm{softmax}(\mathbf{a}), \qquad \mathbf{a} \in \mathbb{R}^{|\mathcal{L}|}, \tag{5}$$

where the logits $\mathbf{a}$ are initialised to zero so step 0 corresponds to a uniform mix. The fused DINO tokens serve as key/value for cross-attention on the initialPromptGen features $\mathbf{z}^{(0)}$, producing a gated residual with γ_d initialised to zero:

$$\Delta\mathbf{z}_{\text{DINO}} = \gamma_d \cdot \mathrm{LN}\big(\mathrm{MHA}(\mathbf{z}^{(0)}, \bar{\mathbf{t}}, \bar{\mathbf{t}})\big), \qquad \mathbf{z}^{(1)} = \mathbf{z}^{(0)} + \Delta\mathbf{z}_{\text{DINO}}. \tag{6}$$

Because cross-attention imposes no constraint on key/value length, the SAM3 patch grid (N_s) and the DINOv3 patch grid (N_d) are allowed to differ without any explicit resampling.

We prefer softmax fusion to hard concatenation because key/value memory stays independent of $|\mathcal{L}|$ and the network learns layer-wise importance rather than a handcrafted schedule—useful on agricultural imagery, where the relevant semantic scale depends on the referent. The residual gate γ_d is zero-initialised, so the DINO branch is an exact no-op at the start of training and is gradually incorporated as the level projections and fusion logits are optimised.

3.4 TEXT-CONDITIONED PROMPT MODULE

SAM3 accepts a text prompt, but its internal tokenizer truncates the input to 32 tokens (`max_length=32` in the SAM3 text processor), while its text encoder only exposes a pooled sentence embedding. Agricultural referring expressions routinely carry fine-grained attributes—phenology (*matured*, *harvested*), land-use modifiers (*irrigated*, *terraced*), and spatial relations (*north of the road*, *adjacent to the river*)—that either overflow the 32-token budget or are averaged away by pooling. To recover the lost signal, we reroute a separate CLIP text-encoder pass with a longer token context into the adapter via cross-attention.

We independently call a frozen CLIP ViT-B/32 text encoder with the full 77-token context and take the token-level hidden states $\mathbf{T} \in \mathbb{R}^{B\times T\times 512}$. Figure 2(c) summarises the bridge. A single linear layer projects them to the shared bottleneck, $\tilde{\mathbf{T}} = \mathrm{Linear}_{512\to d}(\mathbf{T})$ with $d = 128$. The DINO-refined PromptGen features $\mathbf{z}^{(1)}$ then attend over $\tilde{\mathbf{T}}$,

$$\Delta\mathbf{z}_{\text{text}} = \gamma_t \cdot \mathrm{LN}\big(\mathrm{MHA}(\mathbf{z}^{(1)}, \tilde{\mathbf{T}}, \tilde{\mathbf{T}}; \neg\mathbf{m})\big), \qquad \mathbf{z} = \mathbf{z}^{(1)} + \Delta\mathbf{z}_{\text{text}}. \tag{7}$$

where $m \in \{0, 1\}^{B\times T}$ is CLIP's attention mask and γ_t is a learnable scalar gate initialised to zero. Masking out the padded positions prevents the attention layer from spending capacity on filler tokens, which we found critical for short captions.

The DINO–FPN and text bridges are complementary: the former injects a multi-scale visual prior, while the latter provides token-level linguistic guidance. They sequentially refine the PromptGen features through gated residual updates, with both γ_d and γ_t zero-initialised to preserve the original PromptGen pathway at the start of training.

3.5 ENCODER ADAPTATION AND MASK REFINEMENT

This subsection closes the model with two backbone-side additions: LoRA on the SAM3 image encoder and a residual refinement head after the decoder.

LoRA on the SAM3 image encoder. PromptGen-FFT perturbs each encoder block only at its input, but the attention statistics of SAM3 remain those of web-domain pretraining. We therefore attach LoRA Hu et al. (2022) to the q_{proj} and v_{proj} layers of all 32 encoder blocks. For a frozen projection W_0, the adapted forward is

$$y = W_0 x + (\alpha/r)BAx, \qquad A \in \mathbb{R}^{r\times d},\ B \in \mathbb{R}^{d\times r}. \tag{8}$$

We use r=16 and α=32, A is Kaiming-initialised and B is zero-initialised, so the LoRA branch is an exact no-op at the start of training.

Mask refinement head. The SAM3 decoder provides strong semantic masks, but its $16\times$ patch stride limits boundary precision on thin irrigation channels, tillage lines and fragmented post-harvest fields. We add a lightweight residual head that uses full-resolution image cues. A two-layer Conv–GN–GELU stem encodes the input image $\mathbf{I}$ (I_{skip} in Figure 1) into $\mathbf{F}_I \in \mathbb{R}^{B\times 64\times H\times W}$. The upsampled coarse-mask logits $\hat{\mathbf{M}}_c$ are concatenated with $\mathbf{F}_I$, passed through three Conv–GN–GELU blocks, and projected by a zero-initialised 1×1 convolution into a residual $\Delta\mathbf{M}$:

$$\hat{\mathbf{M}} = \sigma(\hat{\mathbf{M}}_c + \Delta\mathbf{M}). \tag{9}$$

Thus, the zero-initialised refinement branch initially preserves the coarse decoder prediction and learns only the residual needed for boundary correction.

Joint objective. The two training stages share the same mask supervision but differ in the active loss terms. Stage 1 uses the full mask objective:

$$\mathcal{L}_1^{\text{mask}} = 20\mathcal{L}_{\text{Dice}} + 20\mathcal{L}_{\text{Focal}} + 3\mathcal{L}_{\text{Lovász}} + 30\mathcal{L}_{\text{Boundary}} + 8\mathcal{L}_{\text{OHEM}}, \tag{10}$$

where Dice Milletari et al. (2016), Focal Lin et al. (2017b), Lovász Berman et al. (2018), Boundary loss Kervadec et al. (2021) and OHEM address region overlap, imbalance, IoU optimisation, contour fidelity and hard pixels, respectively. Stage 2 retains only the Dice and Focal terms:

$$\mathcal{L}_2^{\text{mask}} = 20\mathcal{L}_{\text{Dice}} + 20\mathcal{L}_{\text{Focal}}. \tag{11}$$

In both stages, the final prediction and coarse-mask logits are jointly supervised, together with the CAPA presence head:

$$\mathcal{L}^{\text{seg}} = \mathcal{L}_s^{\text{mask}}(\hat{\mathbf{M}}) + 0.5\,\mathcal{L}_s^{\text{mask}}(\hat{\mathbf{M}}_c) + \mathcal{L}_{\text{pres}}, \qquad s \in \{1, 2\}. \tag{12}$$

where $\mathcal{L}_{\text{pres}}$ is the binary cross-entropy loss of the CAPA presence head.

4 Experiments

4.1 Datasets and Evaluation Protocol

Training set. All experiments use FarmSeg-VL Tao et al. (2025) as the sole training source. The benchmark provides 15,820 training, 4,530 validation, and 2,272 test image–expression–mask triplets covering mainland-China farmland, with referring expressions drawn from its `img_description_eg` field and combining shape, season, terrain, and surrounding-environment descriptors in a single sentence. Inputs are processed at 896×896 resolution.

Cross-domain evaluation. We evaluate every baseline and SFC-SAM3 under the same protocol on three evaluation sets: FarmSeg-VL, RRSIS-D Liu et al. (2024b), and RefSegRS Yuan et al. (2024). FarmSeg-VL is the in-domain test set. For cross-domain analysis, we construct a land-related subset from RRSIS-D and a land-/ground-related subset from RefSegRS by selecting official labels whose referents correspond to extended land or ground regions rather than small isolated objects. These subsets are fixed for all methods, so each model is compared on the same FarmSeg-VL, RRSIS-D land, and RefSegRS land/ground splits.

Metrics. We adopt the standard RIS protocol Yang et al. (2022); Liu et al. (2024b): mean Intersection-over-Union (mIoU, averaged over per-sample IoU) and overall IoU (oIoU, computed from cumulative intersection and union at the pixel level). Distributional behaviour is summarised by precision at IoU thresholds $\text{Pr@}\{0.5, 0.7, 0.9\}$. All reported results are computed on the corresponding evaluation splits using 4-way flip test time augmentation (TTA).

4.2 Implementation Details

Experiments are conducted on a single NVIDIA RTX 5090 GPU. We use SAM3, satellite-pretrained DINOv3 ViT-L/16 and CLIP ViT-B/32 as the three foundation backbones, with images resized to 896×896. DINOv3 and CLIP remain frozen throughout training, while the SAM3 image encoder is adapted through LoRA in both stages and its mask decoder is fine-tuned in Stage 1 and frozen in Stage 2. Models are trained with AdamW, cosine decay, weight decay 10^{-2}, mixed precision, standard geometric/photometric augmentation, and EMA decay 0.9995. The final checkpoint is selected by FarmSeg-VL validation mIoU. FarmSeg-VL captions contain template metadata such as acquisition month and geographic coordinates in addition to the farmland description. We retain the original captions during training and evaluation.

Training uses a two-stage optimisation schedule on FarmSeg-VL. Stage 1 jointly fine-tunes PromptGen-FFT, the DINO-FPN and text-conditioned bridges, the CAPA presence head, LoRA residuals, mask refinement head and SAM3 mask decoder for 25 epochs with the full objective in Section 3.5. Stage 2 continues from the best Stage 1 checkpoint in a stricter PEFT regime: the decoder is frozen, and only the adapters, LoRA and refinement head are updated. Differential learning rates are used for the text bridge (1×10^{-4}), core/DINO adapters (3×10^{-5}), refinement head (5×10^{-5}), and LoRA (1.5×10^{-5}). In Stage 2, Lovász, Boundary and OHEM losses are disabled, while Dice/Focal and coarse-mask supervision are retained, leaving 4.28M trainable parameters (0.35%).

4.3 Quantitative and Qualitative Results

Table 1(a) reports the quantitative comparison on FarmSeg-VL and two cross-domain land-oriented subsets. The baselines cover two regimes: training-free foundation-model pipelines (Grounded-SAM2, Florence-2, and LISA) and RIS models trained on FarmSeg-VL (LAVT and RMSIN). This protocol separates two questions: whether a model remains competitive on the in-domain agricultural benchmark, and whether the learned representation transfers to land-related scenes in other remote-sensing RIS datasets.

Table 1: Quantitative comparison on FarmSeg-VL and land-related cross-domain subsets.

(a) Overall comparison on in-domain and cross-domain benchmarks.

#	Method	Training config	FarmSeg-VL		RRSIS-D		RefSegRS	
			mIoU	oIoU	mIoU	oIoU	mIoU	oIoU
1	Grounded-SAM2	zero-shot	47.01	48.55	0.04	0.03	4.02	5.49
2	Florence-2	zero-shot	47.09	48.64	0.08	0.04	7.74	7.52
3	LISA	zero-shot	35.71	38.11	5.65	8.62	13.99	**27.63**
4	LAVT	FarmSeg-VL	86.46	**90.20**	1.16	1.72	0.20	0.67
5	RMSIN	FarmSeg-VL	86.11	89.94	2.20	2.99	1.34	3.04
6	**SFC-SAM3 (ours)**	FarmSeg-VL	**87.04**	89.72	**41.14**	**51.26**	**19.81**	26.48

(b) Detailed FarmSeg-VL metrics.

#	Method	mIoU	oIoU	Pr@0.5	Pr@0.7	Pr@0.9
1	Grounded-SAM2	47.01	48.55	47.93	30.37	8.71
2	Florence-2	47.09	48.64	48.15	30.33	8.71
3	LISA	35.71	38.11	33.49	15.49	3.08
4	LAVT	86.46	**90.20**	**95.95**	86.97	57.88
5	RMSIN	86.11	89.94	95.51	86.40	56.65
6	**SFC-SAM3 (ours)**	**87.04**	89.72	93.64	**87.05**	**62.73**

On FarmSeg-VL, SFC-SAM3 reaches 87.04% mIoU, matching the strongest dataset-specific RIS models and modestly improving over RMSIN (86.11%) and LAVT (86.46%). This is important because SFC-SAM3 is not built as a fully dataset-specific encoder. It starts from broadly transferable foundation priors from SAM3, DINOv3 and CLIP, and specializes them with lightweight adapters. In other words, the model keeps the generalization capacity of its backbones without sacrificing in-domain accuracy. The detailed metrics in Table 1(b) show the same trend: LAVT obtains the highest oIoU and Pr@0.5, indicating strong foreground coverage on FarmSeg-VL, while SFC-SAM3 gives the best mIoU, Pr@0.7, and Pr@0.9. The gain at the strict Pr@0.9 threshold (62.73% vs. 57.88%

for LAVT and 56.65% for RMSIN) suggests that our model more often produces high-overlap masks rather than only coarse foreground coverage.

The difference becomes much larger outside the training distribution. On the RRSIS-D land subset, SFC-SAM3 obtains 41.14% mIoU and 51.26% oIoU, while the best non-ours mIoU is only 5.65% from LISA. FarmSeg-trained LAVT and RMSIN drop to 1.16% and 2.20% mIoU, showing that strong in-domain fitting does not by itself yield transferable land-region grounding. On the RefSegRS land/ground subset, SFC-SAM3 again achieves the highest mIoU (19.81%), ahead of LISA (13.99%) and the training-free cascades. Overall, SFC-SAM3 remains competitive with dataset-specific RIS models in-domain while transferring substantially better to land-related cross-domain samples. This supports our central claim that specializing foundation priors through lightweight cross-modal adapters provides a more robust route to land-oriented remote-sensing RIS than either direct foundation-model prompting or conventional RIS fine-tuning alone.

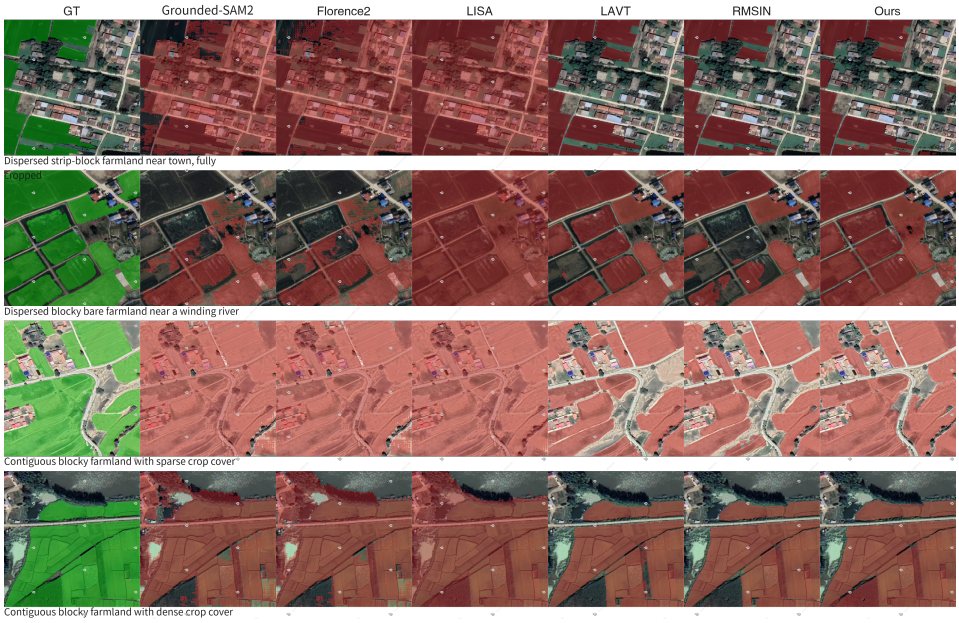

Figure 3: Qualitative comparison. Green denotes ground truth and red denotes predictions.

Figure 3 provides a visual check on these trends. The training-free pipelines often either cover large irrelevant areas or fail to isolate the referred farmland parcel. LAVT and RMSIN are much stronger in-domain, but their masks still tend to leak into neighbouring fields or miss thin boundary structures. SFC-SAM3 produces masks that more closely follow the target parcel extent, especially where field boundaries are weak or broken. This behaviour is consistent with the quantitative pattern in Table 1(b): the gains are not only in average mIoU, but also in the stricter Pr@0.7 and Pr@0.9 metrics.

4.4 Ablation Studies

Table 2 presents a progressive ablation showing how SFC-SAM3 is built from a general foundation-model baseline. The study starts from a SAM3-only configuration and then gradually adds PromptGen-FFT, decoder fine-tuning, DINO guidance, stronger augmentation, text conditioning, LoRA, high-resolution training, boundary-aware supervision, OHEM, DINO-FPN and the mask refinement head. This is not a leave-one-out ablation: each row keeps the changes introduced in the previous rows, so the table should be read as the cumulative construction of the final model.

SAM3 alone reaches 1.01% mIoU, whereas adding PromptGen-FFT and CAPA raises it to 63.19%, establishing prompt-space adaptation as the main entry point for agricultural specialisation. Decoder fine-tuning and the DINO semantic bridge add smaller gains, suggesting that the prompt generator still benefits from a trainable decoder and dense visual priors.

Table 2: Progressive ablation on FarmSeg-VL. Each row adds components on top of the previous configuration.

	Cumulative setting	mIoU	Δ
0	SAM3 only	1.01	–
(a)	PromptGen-FFT + CAPA	63.19	+62.18
(b)	+ decoder fine-tuning	63.90	+0.71
(c)	+ DINO semantic bridge	64.67	+0.77
(d)	+ augmentation + wider bottleneck	69.09	+4.42
(e)	+ text bridge + EMA + TTA	71.34	+2.25
(f)	+ LoRA + Lovász	77.14	+5.80
(g)	+ high-res + boundary supervision	81.25	+4.11
(h)	+ OHEM + DINO-FPN + mask refinement	87.04	+5.79

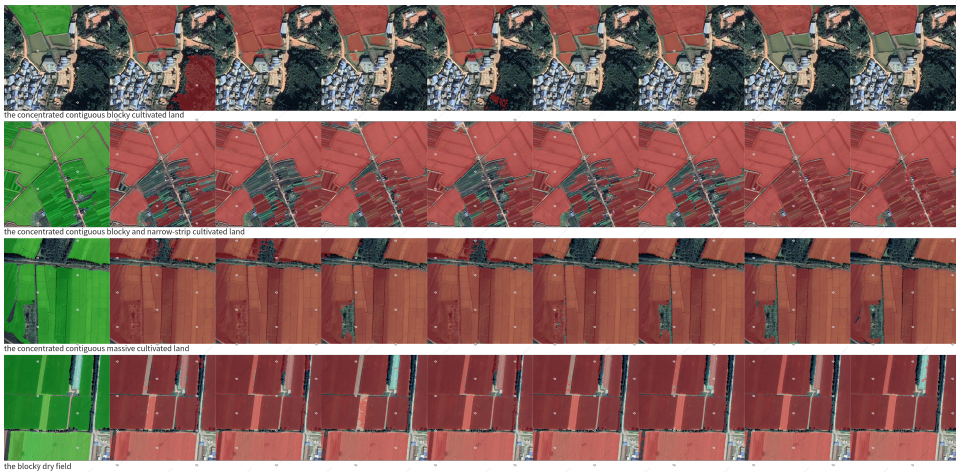

Figure 4: Qualitative progression of the ablation. Columns show the ground truth and (a)–(h).

The visual progression in Figure 4 follows the same trend: early masks recover coarse farmland but leak into adjacent plots, while later stages become tighter after DINO guidance, token-level text conditioning and LoRA are introduced. The middle and late rows add training and refinement gains: augmentation with a wider bottleneck reaches 69.09%, text conditioning with EMA/TTA reaches 71.34%, LoRA with Lovász supervision reaches 77.14%, and high-resolution boundary training reaches 81.25%. The final OHEM+DINO-FPN+refinement setting reaches 87.04% mIoU. Overall, the improvement comes from the accumulation of complementary signals: the frequency branch captures field structure, DINO-FPN supplies multi-level visual semantics, the text bridge strengthens referring-expression grounding, and the Mask Refine Head improves boundary details.

5 Conclusion

We present SFC-SAM3, a parameter-efficient framework that coordinates complementary structural, visual, and linguistic foundation priors for agricultural RRSIS. Its two-stage training schedule uses only 6.58M (0.54%) and 4.28M (0.35%) trainable parameters in Stages 1 and 2, respectively, while achieving 87.04% mIoU on FarmSeg-VL. SFC-SAM3 remains competitive with RIS models trained directly on FarmSeg-VL while transferring substantially better to land- and ground-related subsets of RRSIS-D and RefSegRS, suggesting that agricultural RRSIS benefits from task-structured specialization of foundation priors rather than simply scaling generic models.

Future work will extend SFC-SAM3 to multi-temporal and multimodal remote sensing (e.g., SAR and multispectral imagery) for improved robustness to seasonal and crop-state variations, while exploring lighter LoRA, distillation, and edge-deployment variants for large-scale applications. These directions may further extend SFC-SAM3 to diverse land-cover domains while preserving the general priors of foundation models.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation and to provide feedback on research methodology, experimental design, and the interpretation of experimental results. We have not used generative AI tools to generate synthetic datasets, formulate mathematical claims, provide critical ingredients for proving mathematical claims, or assist in the writing of mathematical proofs; the remaining tasks with required disclosure are not applicable to this work. Additionally, we used generative AI tools for language editing and readability improvement, manuscript organization, literature-related assistance, code editing, and suggestions for the presentation and modification of scientific figures. We have reviewed all AI-assisted work. In particular, AI-assisted code was manually reviewed and tested for correctness, literature-related suggestions were checked against the original sources, and AI-assisted text, figures, and interpretations were verified against the actual methodology and experimental results. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Maxim Berman, Amal Rannen Triki, and Matthew B. Blaschko. The Lovász-Softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4413–4421, 2018.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2511.16719.

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

Keyan Chen, Chenyang Liu, Bowen Chen, Jiafan Zhang, Zhengxia Zou, and Zhenwei Shi. RSRefSeg 2: Decoupling referring remote sensing image segmentation with foundation models. *IEEE Transactions on Geoscience and Remote Sensing*, 64:1–20, 2026. doi: 10.1109/TGRS.2026.4700420.

Tianrun Chen, Lanyun Zhu, Chaotao Ding, Runlong Cao, Yan Wang, Shangzhan Zhang, Zejian Li, Lingyun Sun, Ying Zang, and Papa Mao. SAM-Adapter: Adapting segment anything in underperformed scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 3367–3375, 2023.

Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. VLT: Vision-language transformer and query generation for referring segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(6):7900–7916, 2023. doi: 10.1109/TPAMI.2022.3217852.

Ziyang Gong, Zhixiang Wei, Di Wang, Xianzheng Ma, Hongruixuan Chen, Yuru Jia, Yupeng Deng, Zhenming Ji, Xiangwei Zhu, Naoto Yokoya, Jing Zhang, Bo Du, and Liangpei Zhang. CrossEarth: Geospatial vision foundation model for domain generalizable remote sensing semantic segmentation. *arXiv preprint arXiv:2410.22629*, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Xiaoxing Hu, Ziyang Gong, Yupei Wang, Yuru Jia, Gen Luo, and Xue Yang. Earth-Adapter: Bridge the geospatial domain gaps with mixture of frequency adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. arXiv:2504.06220.

Hoel Kervadec, Jihene Bouchtiba, Christian Desrosiers, Eric Granger, Jose Dolz, and Ismail Ben Ayed. Boundary loss for highly unbalanced segmentation. *Medical Image Analysis*, 67:101851, 2021. doi: 10.1016/j.media.2020.101851.

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

Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. GeoChat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27831–27840, 2024.

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

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 936–944, 2017a. doi: 10.1109/CVPR.2017.106.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017b. doi: 10.1109/ICCV.2017.324.

Chang Liu, Henghui Ding, and Xudong Jiang. GRES: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23592–23601, 2023.

Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. RemoteCLIP: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024a. doi: 10.1109/TGRS.2024.3390838.

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

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *Fourth International Conference on 3D Vision (3DV)*, pp. 565–571, 2016. doi: 10.1109/3DV.2016.79.

Umair Nawaz, Muhammad Awais, Hanan Gani, Muzammal Naseer, Fahad Shahbaz Khan, Salman Khan, and Rao Muhammad Anwer. AgriCLIP: Adapting CLIP for agriculture and livestock via domain-specialized cross-model alignment. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, pp. 9630–9639, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *PMLR*, pp. 8748–8763, 2021.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S. Khan. GLaMM: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13009–13018, 2024.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. PixelLM: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26374–26383, 2024.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Xian Sun, Peijin Wang, Wanxuan Lu, Zicong Zhu, Xiaonan Lu, Qibin He, Junxi Li, Xuee Rong, Zhujun Yang, Hao Chang, Qinglin He, Guang Yang, Ruiping Wang, Jiwen Lu, and Kun Fu. RingMo: A remote sensing foundation model with masked image modeling. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–22, 2023. doi: 10.1109/TGRS.2022.3194732.

Chao Tao, Dandan Zhong, Weiliang Mu, Zhuofei Du, and Haiyang Wu. A large-scale image-text dataset benchmark for farmland segmentation. *Earth System Science Data*, 17:4835–4858, 2025. doi: 10.5194/essd-17-4835-2025.

Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. CRIS: CLIP-driven referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11676–11685, 2022.

Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4974–4984, 2022.

Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip H. S. Torr. LAVT: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18155–18165, 2022.

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

Yang Zhan, Zhitong Xiong, and Yuan Yuan. SkyEyeGPT: Unifying remote sensing vision-language tasks via instruction tuning with large language model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 221:64–77, 2025. doi: 10.1016/j.isprsjprs.2025.01.020.

Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from CLIP. In *European Conference on Computer Vision (ECCV)*, volume 13688 of *LNCS*, pp. 696–712, 2022.

Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, Nanyun Peng, Lijuan Wang, Yong Jae Lee, and Jianfeng Gao. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15116–15127, 2023.