DISP: DECOUPLED IMAGE-SEMANTIC PROMPTING WITH MULTI-SCALE VISUAL ALIGNMENT FOR REMOTE SENSING REFERRING SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In remote sensing referring segmentation, models must localize socially defined regions—such as parks, schools, and commercial zones—whose boundaries depend not only on visual appearance but also on functional and contextual semantics. This makes socio-semantic segmentation especially challenging, since visually similar regions may correspond to different social meanings and require fine-grained cross-modal reasoning. Existing vision-language segmentation paradigms still face two key limitations: single-token methods entangle semantic reasoning and spatial localization within one representation, while multi-stage VLM-to-SAM pipelines compress rich socio-semantic understanding into coarse geometric prompts; moreover, they insufficiently model multi-scale regional visual statistics needed for precise masks. To address these challenges, we propose **DISP** (Decoupled Image-Semantic Prompting with Multi-scale Visual Alignment), a unified framework that decouples the language-semantic role and visual-spatial role into two dedicated tokens, `[SEG]` and `[IMG]`. These tokens are projected through separate pathways and used as independent sparse prompts for the SAM2 decoder under a split-path dual-image architecture, enabling collaborative yet disentangled semantic-spatial segmentation. We further introduce Multi-scale Masked Feature Matching (MS-MFM), a training-only supervision module that aligns the `[IMG]` token with attention-pooled target-region representations from multiple ViT layers, allowing it to capture hierarchical visual cues with zero inference overhead. Experiments on socio-semantic segmentation benchmarks demonstrate that DISP achieves consistent SOTA performance.

1 INTRODUCTION

Multi-modal Large Language Models (MLLMs) have recently achieved remarkable progress, largely driven by advances in Large Language Models (LLMs) Touvron et al. (2023); Hugo Touvron et al. (2023); Yang et al. (2024) and large-scale vision–language alignment. Beyond image-level recognition and question answering, MLLMs are increasingly being extended to fine-grained visual understanding and pixel-level prediction, with models such as Sa2VA Yuan et al. (2025) demonstrating the potential to unify multimodal reasoning and segmentation. This capability is particularly appealing for remote sensing, where interpreting a scene often requires combining visual evidence with geographic knowledge and natural-language instructions. Nevertheless, directly applying existing MLLM-based segmentation methods to socio-semantic regions remains challenging. Unlike conventional objects with relatively stable visual signatures, socially defined regions—such as schools, hospitals, parks, and commercial zones—are determined jointly by appearance, spatial organization, surrounding infrastructure, and human function. Consequently, visually similar areas may carry different social meanings, while regions sharing the same function may exhibit substantially different textures, layouts, and scales.

Several recent works have begun to address these characteristics through geospatial reasoning and remote sensing segmentation (Li et al., 2025b; Yao et al., 2025). Among them, SocioReasoner Wang et al. (2026) introduces a three-level progressive task hierarchy, ranging from socio-entity name segmentation to socio-category segmentation and ultimately socio-function segmentation. This hierarchy encourages VLMs to move from explicit entity recognition toward increasingly abstract func-

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.

tional reasoning, substantially advancing socio-semantic scene understanding. However, accurate segmentation requires more than identifying the intended social concept: the model must also translate that concept into precise and complete spatial support across paired satellite and digital-map observations. Existing methods still struggle to preserve rich semantic evidence during localization and to represent the diverse regional appearance needed for boundary-sensitive prediction, as illustrated in Figure 1.

Existing reasoning-segmentation methods mainly follow two paradigms. The first is the LISA-style paradigm Lai et al. (2024), in which a single special token (`[SEG]`) must simultaneously encode the target semantics and its spatial extent. Although compact and end-to-end trainable, this design forces "what to segment" and "where to segment" into a shared representation, allowing global semantic reasoning and local boundary cues to interfere with one another. The second is the VLM-to-SAM paradigm adopted by methods such as SocioReasoner Wang et al. (2026), which predicts bounding boxes and subsequently refines them into masks through sampled point prompts. While this decomposition benefits from the strong mask prior of SAM, it compresses the VLM's rich socio-semantic understanding into a small set of coarse geometric prompts. Ambiguous or incomplete boxes can therefore discard contextual evidence before mask decoding, and the multi-stage, non-differentiable prompt construction also weakens direct semantic-to-mask supervision.

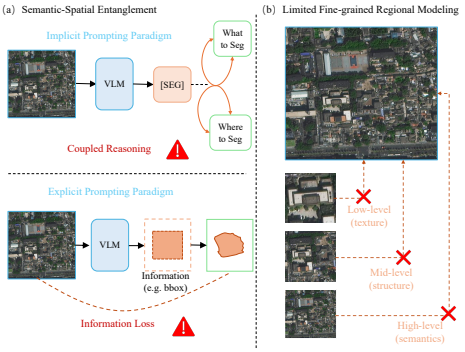


Figure 1: Existing methods either couple semantic reasoning with localization or compress semantics into coarse prompts, while insufficient multi-scale modeling limits precise regional alignment.

To overcome this limitation, we propose **DISP**, a unified decoupled reasoning framework with two complementary special tokens. The global semantic token `[SEG]` is responsible for interpreting the language query and integrating social context from the satellite–map pair, whereas the image-oriented token `[IMG]` focuses on the target region's local appearance, spatial structure, and boundary evidence. Their hidden states are transformed by separate projection pathways and supplied to the SAM2 decoder as independent sparse prompts. This design preserves task-specific information in each representation while allowing the decoder to jointly exploit semantic and spatial cues, providing a direct end-to-end connection from multimodal reasoning to mask generation.

Decoupling alone, however, does not guarantee that the `[IMG]` token contains sufficiently detailed regional information. Socio-semantic targets may span a few buildings or an entire functional zone, and their boundaries may be indicated by low-level textures, mid-level geometric patterns, or high-level contextual structure. We therefore introduce Multi-scale Masked Feature Matching (MS-MFM), a training-time supervision module that explicitly strengthens the `[IMG]` token with hierarchical target-region features. MS-MFM extracts patch representations from shallow, intermediate, and deep ViT layers, covering visual evidence from local textures to semantic organization. At each scale, a learnable masked attention-pooling head adaptively selects discriminative features inside the ground-truth region, reducing the information dilution caused by naive average pooling Liu et al. (2024); Aasan et al. (2026). The resulting representations supervise both the first-order feature content and second-order relational statistics of the `[IMG]` token. Because MS-MFM is used only as an auxiliary training objective and removed at inference, it improves regional alignment without introducing additional inference overhead.

The main contributions of this work are summarized as follows:

- We identify a key issue in socio-semantic segmentation: the entanglement of semantic reasoning and spatial localization, which limits the clear separation of "what to segment" and "where to segment." To address this, we propose a decoupled dual-prompt mechanism for semantic-spatial collaboration.

- We find that existing methods poorly model fine-grained regional visual information, making it difficult to capture multi-scale structural variations in complex socio-semantic regions. To this end, we introduce MS-MFM, a multi-scale supervision module that enhances the `[IMG]` token via hierarchical feature alignment.
- Experiments on socio-semantic segmentation benchmarks show that our DISP framework achieves consistent state-of-the-art performance across multiple metrics, demonstrating its effectiveness.

2 Related Work

2.1 Semantic Segmentation

Semantic segmentation is a fundamental task in computer vision (Voulodimos et al., 2018). Early deep learning methods relied on fully supervised training and could only recognize a fixed set of predefined categories (Ronneberger et al., 2015; Xie et al., 2021; Zhu et al., 2023). With the rise of large-scale pre-trained models, open-vocabulary segmentation (Ghiasi et al., 2022) and referring segmentation (Wang et al., 2022) have enabled recognition of unseen categories or segmentation based on textual descriptions. More recently, reasoning segmentation (Lai et al., 2024) extends this paradigm by requiring models to infer targets from functional or relational descriptions rather than visual appearance. Many recent approaches adopt VLM-based pipelines (Liu et al., 2025a; You & Wu, 2025; Liu et al., 2025b), where visual prompts (e.g., boxes or points) generated by VLM reasoning are fed into SAM, often with reinforcement learning to enhance reasoning. A similar evolution appears in satellite imagery segmentation (Kotaridis & Lazaridou, 2021), from supervised extraction of buildings (Cheng et al., 2019) and roads (Sun et al., 2019), to open-vocabulary (Li et al., 2025a; Zhu et al., 2025b) and referring segmentation (Chen et al., 2025). Recent work further explores reasoning segmentation in remote sensing using closed-source VLMs to rephrase categories into reasoning-based queries (Li et al., 2025b), and some studies combine MLLMs with SAM in multi-stage frameworks for urban semantic understanding Wang et al. (2026). However, these methods rely on complex pipelines and do not fully exploit the end-to-end capability of MLLMs and SAM. In contrast, our work focuses on a unified end-to-end framework that fully leverages both MLLMs and SAM for semantic understanding and segmentation.

2.2 Multi-Modal Approaches for Urban Understanding

The task of segmenting urban socio-semantic entities (i.e., urban socio-semantic segmentation) is an emerging research direction, and existing work remains limited. A representative study is SocioReasoner Wang et al. (2026), which adopts a multi-stage training pipeline to first enhance MLLMs' understanding of socio-semantic concepts and then leverages SAM for segmentation. Beyond this, related research mainly comes from urban science, such as land-use classification (Xiong et al., 2025) and urban functional zone analysis (Yao et al., 2018). These methods typically integrate multimodal data, including Points of Interest (POIs), road networks, and satellite imagery, and use separate encoders for each modality followed by feature fusion for classification or segmentation (Xiong et al., 2025; Zhang et al., 2017). However, they face clear limitations when handling fine-grained socio-semantic attributes and entity-name-based segmentation, as they struggle to effectively model the correspondence between social semantics and visual information, resulting in suboptimal segmentation performance. In contrast, our method fully leverages the reasoning and semantic understanding capabilities of MLLMs, and explicitly decouples semantic reasoning from visual perception, enabling more accurate and higher-quality end-to-end segmentation.

3 Method

3.1 Problem Formulation and Overview

Given a training sample (I^r, I^t, q, Y), where I^r denotes the reference image used by the MLLM for semantic reasoning, I^t denotes the target remote sensing image to be segmented by SAM2Ravi et al. (2025), q is the referring expression, and $Y \in \{0,1\}^{H\times W}$ is the ground-truth mask, our goal is to predict a binary mask $\hat{Y}$ that is both semantically consistent with q and spatially aligned with

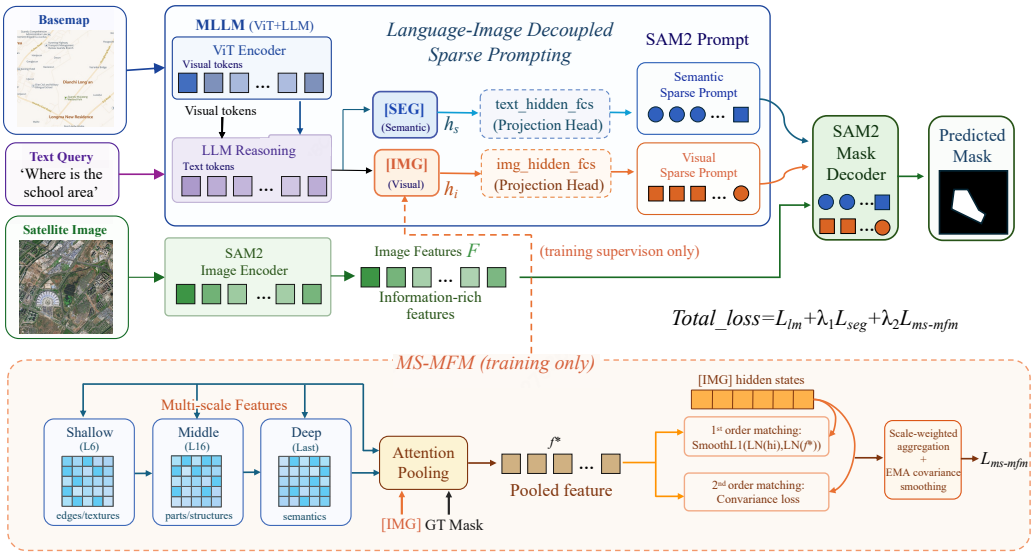


Figure 2: Overview of DISP. The MLLM generates decoupled `[SEG]` and `[IMG]` tokens from the reference image and referring expression, which are projected as semantic and visual sparse prompts for SAM2 mask decoding on the target image. MS-MFM supervises the `[IMG]` token with multi-scale masked ViT features during training and is removed at inference.

the target region in I^t. Existing VLM-to-SAM paradigms usually compress the entire cross-modal reasoning result into a single hidden state, making one token responsible for both semantic selection and spatial localization. We argue that this single-channel interface is suboptimal for socio-semantic segmentation, where the target is often defined by functional semantics while its visual evidence is distributed across multi-scale regional structures.

From an interface perspective, the VLM-to-SAM2Ravi et al. (2025) pipeline can be written as learning a prompt generator Π_θ that converts the reasoning context into sparse prompts for a frozen or fine-tuned mask decoder:

$$\hat{Y} = \sigma\left(\mathcal{D}_{\text{sam}}\left(\mathcal{E}_{\text{sam}}(I^t), \Pi_\theta(I^r, q)\right)\right). \tag{1}$$

The difficulty is that Π_θ must simultaneously produce a prompt that is discriminative in the language space and geometrically useful in the SAM2Ravi et al. (2025) mask space. In a single-token design, these two requirements are forced to share the same latent vector. This creates an optimization conflict: gradients that make the token more abstract and category-level for semantic reasoning may not be aligned with gradients that make it more sensitive to local visual boundaries and regional textures. DISP addresses this conflict by explicitly factorizing the prompt generator into two complementary components,

$$\Pi_\theta(I^r, q) = \left[\Pi_\theta^{\text{sem}}(I^r, q); \Pi_\theta^{\text{vis}}(I^r, q)\right], \tag{2}$$

where the first component carries language-grounded semantics and the second component is regularized to carry target-region visual statistics.

To address this issue, we propose **DISP**, a unified framework with two coupled designs. First, we introduce **Language-Image Decoupled Sparse Prompting (LID-SP)**, which decomposes the MLLM-to-SAM2 interface into a semantic prompt token `[SEG]` and a visual-spatial prompt token `[IMG]`. Second, we introduce **Multi-scale Masked Feature Matching (MS-MFM)**, a training-only objective that transfers multi-scale target-region statistics from the MLLM ViT encoder to the `[IMG]` token. During inference, DISP only keeps the learned MLLM, two lightweight projection heads, and the SAM2 decoder, while MS-MFM is discarded and introduces no extra cost.

3.2 LANGUAGE-IMAGE DECOUPLED SPARSE PROMPTING

Let $\mathcal{F}_\theta$ be the MLLM with InternVL3-2BZhu et al. (2025a) as the backbone, let $\mathcal{E}_{\text{vit}}^{\text{mllm}}$ denote its ViT visual encoder, and let $\mathcal{E}_{\text{sam}}$, $\mathcal{D}_{\text{sam}}$ denote the SAM2 image encoder and mask decoder, respectively. For a reference image I^r and query q, we append two special tokens to the language sequence and obtain the hidden states

$$\mathbf{H} = \mathcal{F}_\theta\left(\mathcal{V}(I^r), \mathcal{T}(q), \texttt{[SEG]}, \texttt{[IMG]}\right), \tag{3}$$

where $\mathcal{V}(\cdot)$ and $\mathcal{T}(\cdot)$ are the visual and textual tokenizers of the MLLM. The hidden states at the two special-token positions are extracted as

$$\mathbf{h}_s = \mathbf{H}_{\tau_s}, \qquad \mathbf{h}_i = \mathbf{H}_{\tau_i}, \tag{4}$$

where τ_s and τ_i correspond to `[SEG]` and `[IMG]`, respectively. Instead of using a shared projection, DISP maps the two hidden states into the sparse-prompt space of SAM2 through two independent projection heads:

$$\mathbf{p}_s = \phi_s(\mathbf{h}_s), \qquad \mathbf{p}_i = \phi_i(\mathbf{h}_i), \tag{5}$$

$$\phi_\bullet(\mathbf{h}) = \mathbf{W}_\bullet^{(2)}\delta\left(\mathbf{W}_\bullet^{(1)}\,\mathrm{LN}(\mathbf{h})\right), \quad \bullet \in \{s, i\}, \tag{6}$$

where $\delta(\cdot)$ is a nonlinear activation and $\mathrm{LN}(\cdot)$ denotes layer normalization. The semantic prompt $\mathbf{p}_s$ mainly preserves the language-grounded decision of "what to segment", while the visual prompt $\mathbf{p}_i$ is encouraged to encode "where and how the target appears" in the target image.

The two projection heads are deliberately not shared. The reason is that the two tokens are expected to occupy different subspaces before entering SAM2: $\mathbf{p}_s$ should remain stable under local appearance variations as long as the referred semantics are unchanged, whereas $\mathbf{p}_i$ should be sensitive to regional appearance, texture, and boundary evidence. A shared mapping would implicitly assume that semantic abstraction and visual localization are linearly compatible in the same prompt coordinate system. By contrast, the decoupled projections allow the model to learn two anisotropic transformations from the MLLM hidden space to the SAM2 sparse-prompt space. The resulting prompt set can be viewed as a two-vector basis for mask decoding, where $\mathbf{p}_s$ determines the semantic direction and $\mathbf{p}_i$ provides visual anchoring.

Under the split-path dual-image design, the MLLM reasons over I^r, while SAM2 receives the target image I^t for high-resolution mask decoding. The predicted mask is obtained by

$$\mathbf{Z}^t = \mathcal{E}_{\text{sam}}(I^t), \tag{7}$$

$$\hat{Y} = \sigma\left(\mathcal{D}_{\text{sam}}\left(\mathbf{Z}^t, [\mathbf{p}_s; \mathbf{p}_i]\right)\right), \tag{8}$$

where $\mathbf{Z}^t$ is the final SAM2 image embedding and $[\cdot;\cdot]$ denotes sparse-prompt concatenation. Compared with the single-token formulation, this design expands the VLM-to-SAM2 interface from one mixed prompt to two role-specific prompts, reducing semantic-spatial interference while preserving end-to-end training.

This split-path design is important for remote sensing inputs. The reference image I^r provides a context in which the MLLM can associate the referring expression with socio-semantic concepts, while the target image I^t preserves the spatial resolution required by SAM2. Therefore, the MLLM is not required to memorize pixel-level details that are better handled by the SAM2 encoder. Instead, it only needs to generate compact semantic and visual prompts that condition the target-image decoder. This division of labor keeps the reasoning pathway lightweight and leaves dense geometric modeling to the segmentation pathway.

3.3 MULTI-SCALE MASKED FEATURE MATCHING

Although LID-SP provides a dedicated `[IMG]` token, the token will not automatically internalize detailed regional visual statistics without explicit supervision. We therefore propose MS-MFM to align $\mathbf{h}_i$ with target-region representations extracted from multiple layers of the MLLM ViT. Let $\mathcal{S} = \{\ell_1, \ell_2, \ell_3\}$ denote shallow, middle, and deep layers. The selected MLLM ViT features are obtained as $\{\mathbf{U}_\ell^t\}_{\ell\in\mathcal{S}} = \mathcal{E}_{\text{vit}}^{\text{mllm}}(I^t)$. For each layer ℓ, the patch feature map $\mathbf{U}_\ell^t = \{\mathbf{u}_{\ell,j}\}_{j=1}^{N_\ell}$ is projected to a common D-dimensional space:

$$\mathbf{x}_{\ell,j} = \psi_\ell(\mathbf{u}_{\ell,j}) \in \mathbb{R}^D, \qquad j = 1, \ldots, N_\ell. \tag{9}$$

The use of multiple layers is motivated by the hierarchical nature of region evidence. Shallow features are more responsive to edges, textures, and local contrast; middle features describe recurring parts and spatial layouts; deep features encode semantic regularities. A single alignment layer would bias the `[IMG]` token toward only one of these aspects. MS-MFM instead uses the MLLM ViT as a multi-scale visual teacher and distills its masked regional statistics into the MLLM token.

The ground-truth mask Y is downsampled to the patch grid of layer ℓ, producing a binary patch mask $m_{\ell,j} \in \{0,1\}$. Let $\Omega_\ell = \{j \mid m_{\ell,j} = 1\}$ be the foreground patch set. A naive region descriptor $|\Omega_\ell|^{-1} \sum_{j\in\Omega_\ell} \mathbf{x}_{\ell,j}$ treats all foreground patches equally, which is undesirable when the mask contains mixed materials, shadows, roads, or low-texture areas. Instead, we use the `[IMG]` hidden state as a query to perform masked attention pooling:

$$e_{\ell,j} = \frac{\left(\mathbf{W}_\ell^q \,\mathrm{LN}(\mathbf{h}_i)\right)^\top \left(\mathbf{W}_\ell^k \mathbf{x}_{\ell,j}\right)}{\sqrt{D}}, \tag{10}$$

$$a_{\ell,j} = \frac{m_{\ell,j} \exp(e_{\ell,j})}{\sum_{r=1}^{N_\ell} m_{\ell,r} \exp(e_{\ell,r}) + \epsilon}, \tag{11}$$

$$\mathbf{f}_\ell^* = \sum_{j=1}^{N_\ell} a_{\ell,j} \mathbf{W}_\ell^v \mathbf{x}_{\ell,j}. \tag{12}$$

Here, $\mathbf{f}_\ell^*$ is an attention-pooled region descriptor that adaptively emphasizes discriminative patches inside the target mask, avoiding the signal dilution caused by naive foreground averaging.

The attention weights can be interpreted as a token-conditioned soft selection over the ground-truth region. Unlike class-agnostic pooling, this operation makes the teacher feature depend on the current `[IMG]` representation, so the supervision is not a fixed visual prototype but a dynamically selected regional summary. This is useful when different queries refer to regions with similar appearances but different social meanings, because the token itself participates in deciding which visual evidence should be emphasized.

To supervise the `[IMG]` token at each scale, we first project it to the same feature space as $\mathbf{f}_\ell^*$:

$$\bar{\mathbf{h}}_\ell = \rho_\ell(\mathbf{h}_i). \tag{13}$$

The first-order alignment term matches normalized appearance statistics using a robust Smooth-L1 loss:

$$\mathcal{L}_\ell^{(1)} = \mathrm{SmoothL1}\left(\mathrm{LN}(\bar{\mathbf{h}}_\ell), \mathrm{LN}(\mathbf{f}_\ell^*)\right). \tag{14}$$

However, first-order matching mainly constrains the regional mean and is insufficient to describe internal texture and structural correlations. Therefore, we further introduce a second-order covariance alignment term. For a mini-batch of size B, let $\bar{\mathbf{h}}_\ell^{(b)}$ and $\mathbf{f}_\ell^{*(b)}$ be the token and region descriptors of sample b. Their batch covariance estimates are

$$\widehat{\mathbf{\Sigma}}_\ell^h = \frac{1}{B-1} \sum_{b=1}^{B} \left(\bar{\mathbf{h}}_\ell^{(b)} - \boldsymbol{\mu}_\ell^h\right) \left(\bar{\mathbf{h}}_\ell^{(b)} - \boldsymbol{\mu}_\ell^h\right)^\top, \tag{15}$$

$$\widehat{\mathbf{\Sigma}}_\ell^f = \frac{1}{B-1} \sum_{b=1}^{B} \left(\mathbf{f}_\ell^{*(b)} - \boldsymbol{\mu}_\ell^f\right) \left(\mathbf{f}_\ell^{*(b)} - \boldsymbol{\mu}_\ell^f\right)^\top, \tag{16}$$

where $\boldsymbol{\mu}_\ell^h$ and $\boldsymbol{\mu}_\ell^f$ are batch means. Since remote sensing segmentation is often trained with small batches, direct covariance estimation can be noisy. We maintain exponential moving average (EMA) covariance statistics:

$$\mathbf{\Sigma}_{\ell,t}^h = \beta \mathbf{\Sigma}_{\ell,t-1}^h + (1-\beta)\widehat{\mathbf{\Sigma}}_{\ell,t}^h, \tag{17}$$

$$\mathbf{\Sigma}_{\ell,t}^f = \beta \mathbf{\Sigma}_{\ell,t-1}^f + (1-\beta)\widehat{\mathbf{\Sigma}}_{\ell,t}^f, \tag{18}$$

and define the second-order loss as

$$\mathcal{L}_\ell^{(2)} = \frac{1}{D^2} \left\| \mathbf{\Sigma}_{\ell,t}^h - \mathbf{\Sigma}_{\ell,t}^f \right\|_F^2. \tag{19}$$

This covariance term provides information that cannot be captured by the mean descriptor alone. For example, two urban regions may have similar average color or average deep feature responses, but

differ in the co-occurrence pattern between vegetation, buildings, roads, and open space. Matching second-order statistics encourages the [IMG] token to preserve such internal correlation patterns. The normalization by D^2 makes the scale of the loss less sensitive to the feature dimension, while the EMA update reduces the variance of covariance estimates under small-batch training.

The multi-scale feature matching objective is then

$$\mathcal{L}_{\text{MS-MFM}} = \sum_{\ell \in \mathcal{S}} \alpha_\ell \left(\mathcal{L}_\ell^{(1)} + \lambda_{\text{cov}} \mathcal{L}_\ell^{(2)} \right), \tag{20}$$

where α_ℓ balances shallow texture, middle structure, and deep semantic features. In this way, MS-MFM acts as a visual-statistics distillation objective: it teaches the [IMG] token to absorb multi-scale and multi-order target-region information during training, while requiring no additional module at inference.

3.4 Training Objective and Inference

The segmentation loss combines binary cross-entropy and Dice loss:

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{BCE}}(\hat{Y}, Y) + \lambda_{\text{dice}} \left(1 - \frac{2 \sum_u \hat{Y}_u Y_u + \epsilon}{\sum_u \hat{Y}_u + \sum_u Y_u + \epsilon} \right). \tag{21}$$

When language responses are supervised, we also keep the standard autoregressive language modeling objective $\mathcal{L}_{\text{LM}}$. The final training objective is

$$\mathcal{L} = \mathcal{L}_{\text{LM}} + \lambda_1 \mathcal{L}_{\text{seg}} + \lambda_2 \mathcal{L}_{\text{MS-MFM}}. \tag{22}$$

At inference time, only $\mathbf{h}_s$, $\mathbf{h}_i$, their projection heads, and the SAM2 decoder are used. All mask-dependent operations in MS-MFM, including masked attention pooling and covariance matching, are removed. Therefore, DISP improves the learned visual-spatial prompt representation without increasing inference latency or memory consumption.

4 Experiments

4.1 Datasets.

SocioSegWang et al. (2026) a comprehensive benchmark for social-semantic-driven remote sensing segmentation, consisting of urban entities defined by social functions and semantics, such as schools, hospitals, parks, and commercial areas. These targets often lack stable visual appearances, and their boundaries are determined more by high-level semantic reasoning than by local texture cues. As a result, the dataset requires models to perform cross-modal understanding between satellite imagery and digital map information, enabling inference of both "what" the target is and "where" it is. Furthermore, SocioSeg introduces a three-level hierarchical task design (Socio-name, Socio-class, and Socio-function), progressively increasing semantic abstraction from specific instances to categories and functional concepts. Socio-name contains about 5,000+ fine-grained entity instances, Socio-class includes around 90+ social semantic categories, and Socio-function comprises about 10+ high-level functional categories. This hierarchical structure enables a systematic evaluation of a model's capability.

4.2 Baselines and Metrics.

We compare our method against three categories of approaches. First, standard semantic segmentation models, including the CNN-based UNet (Ronneberger et al., 2015) and the Transformer-based SegFormer (Xie et al., 2021). Second, recent reasoning-based segmentation methods for natural images, including VisionReasoner (Liu et al., 2025b), Seg-R1 (You & Wu, 2025), and SAM-R1 (Huang et al., 2025). Third, state-of-the-art segmentation methods in remote sensing, including open-vocabulary segmentation methods such as SegEarth-OV (Li et al., 2025a), referring segmentation methods such as RSRefSeg (Chen et al., 2025), and reasoning-based approaches such as SegEarth-R1 (Li et al., 2025b), RemoteReasoner (Yao et al., 2025), and SocioReasoner Wang et al. (2026). Since SocioSeg involves two input modalities (satellite images and digital maps), we adapt

all VLM-based baselines to accept dual-image inputs; for methods that do not support multi-image inputs (e.g., RSRefSeg and SegEarth-R1), we provide only the satellite image as input. *All baselines are re-trained on the SocioSeg training split to ensure fair comparison.* For evaluation, we follow prior work (Lai et al., 2024) and report cIoU and gIoU, while additionally introducing the F1 score to evaluate instance-level segmentation performance.

Table 1: Comparison with state-of-the-art methods on SocioSeg test set. The best performance in each column is highlighted in **bold**. The second best is underlined. Baselines are re-trained on the SocioSeg training split to ensure fair comparison.

Method	Socio-name			Socio-class			Socio-function			All dataset		
	cIoU	gIoU	F1	cIoU	gIoU	F1	cIoU	gIoU	F1	cIoU	gIoU	F1
UNet	11.0	9.3	8.1	12.5	12.0	11.1	11.2	10.5	10.5	11.6	10.8	10.1
Segformer	22.1	19.5	18.2	22.3	21.5	19.4	21.5	20.1	18.0	22.0	20.6	18.6
VisionReasoner	48.6	50.8	58.5	44.5	49.2	55.4	36.2	41.9	44.9	44.1	48.4	54.4
Seg-R1	45.9	48.2	50.3	40.5	44.6	45.3	34.4	39.6	36.4	41.1	45.1	45.1
SAM-R1	25.7	25.3	37.1	22.4	23.7	32.2	17.8	19.8	25.1	22.4	23.8	32.3
SegEarth-OV	3.4	3.2	0.0	3.7	3.9	0.0	4.1	4.3	0.0	3.6	3.8	0.0
RSRefSeg	27.2	25.3	31.0	30.6	30.7	35.2	28.8	28.7	30.7	28.9	28.4	32.9
SegEarth-R1	36.8	42.0	47.0	39.0	45.0	50.1	39.6	45.5	47.3	38.2	44.2	48.5
RemoteReasoner	46.7	49.4	56.0	42.8	48.1	53.8	38.1	43.4	47.3	43.3	47.6	53.4
SocioReasoner	52.7	55.6	64.5	47.5	52.9	60.0	40.7	46.8	50.4	47.8	52.9	59.6
Ours	**53.3**	**56.1**	**65.3**	**50.4**	**55.6**	**63.5**	**45.6**	**53.0**	**58.6**	**50.7**	**55.3**	**63.3**

4.3 Main Results.

Table 1 compares DISP under the same three-level SocioSeg hierarchy and evaluation protocol as SocioReasoner Wang et al. (2026). DISP ranks first across all 12 metrics, consistently outperforming standard segmentation models, natural-image reasoning methods, and remote-sensing baselines. Compared with SocioReasoner, the strongest prior method, it improves the overall cIoU/gIoU/F1 by 2.9/2.4/3.7 points. The gains increase from 2.9/2.7/3.5 on Socio-class to 4.9/6.2/8.2 on the more abstract Socio-function task, showing that decoupled `[SEG]` and `[IMG]` prompts are particularly effective when appearance alone is insufficient. The F1 gains also indicate more reliable instance-level masks.

General reasoning models remain notably weaker: VisionReasoner and RemoteReasoner trail DISP overall by 6.6/6.9/8.9 and 7.4/7.7/9.9 cIoU/gIoU/F1 points, respectively. This confirms the importance of the dual-image remote-sensing design and explicit semantic–spatial separation. The strong Socio-function results further support MS-MFM, which improves visual-spatial prompting through training-time multi-scale alignment without inference overhead.

4.4 Ablation Results

Table 2: Ablation study on SocioSeg test set. The table follows the same evaluation protocol and metric layout as Table 1.

Variant	Socio-name			Socio-class			Socio-function			All dataset		
	cIoU	gIoU	F1	cIoU	gIoU	F1	cIoU	gIoU	F1	cIoU	gIoU	F1
w/o `[IMG]`	51.9	55.1	64.5	48.9	54.0	63.1	42.4	50.1	57.2	48.9	53.0	62.5
w/o MS-MFM	52.1	54.8	63.9	49.3	53.9	62.4	44.7	52.6	58.0	49.8	54.4	62.7
Full DISP	**53.3**	**56.1**	**65.3**	**50.4**	**55.6**	**63.5**	**45.6**	**53.0**	**58.6**	**50.7**	**55.3**	**63.3**

Table 2 verifies both components across all task levels. Removing `[IMG]` lowers the overall cIoU/gIoU/F1 from 50.7/55.3/63.3 to 48.9/53.0/62.5, with particularly clear Socio-function drops of 3.2/2.9/1.4 points. Thus, semantic reasoning through `[SEG]` alone cannot preserve the image-grounded "where" information required for functional regions; `[IMG]` provides this complementary localization prompt to SAM2.



Figure 3: Qualitative visualization of different ablation variants. From left to right: the ground-truth mask, the prediction without [IMG], the prediction without MS-MFM, and the prediction of the full model.

Removing MS-MFM retains [IMG] but reduces the overall scores by 0.9/0.9/0.6 points. Its consistent effect across all three task levels shows that explicit multi-scale region supervision helps [IMG] encode cues from fine textures to semantic structures. The larger degradation without [IMG] identifies it as the core visual-spatial prompt, while MS-MFM further improves its alignment.

4.5 Visual Analysis

Figure 3 provides qualitative evidence consistent with the ablation results. Without [IMG], the model can generally infer the target category from the semantic prompt, but its predictions may cover only part of the target or extend toward nearby regions with similar appearance. This behavior shows that semantic reasoning alone is insufficient to determine complete, context-dependent boundaries. Introducing [IMG] provides complementary image-grounded cues, improving both target localization and regional coverage. MS-MFM further reduces fragmented responses and boundary deviations by aligning [IMG] with hierarchical target-region features during training. Consequently, the full model produces cleaner and more coherent masks that follow the ground-truth extent more closely, supporting the complementary roles of decoupled prompting and multi-scale visual alignment.

5 Conclusion

In this paper, we presented DISP, a decoupled image-semantic prompting framework for remote sensing referring segmentation. By introducing separate [SEG] and [IMG] tokens, DISP disentangles language-semantic reasoning from visual-spatial localization and provides complementary sparse prompts to the SAM2 decoder under a dual-image architecture. We further proposed MS-MFM, a training-only multi-scale masked feature matching objective that strengthens the [IMG] token with hierarchical target-region visual statistics without adding inference overhead. Extensive experiments on SocioSeg show that DISP consistently outperforms existing segmentation, reasoning-based, and remote-sensing baselines, with especially clear gains on semantically abstract socio-functional targets. Ablation studies and visualizations further verify the importance of the [IMG] token for capturing target-region information and the effectiveness of MS-MFM in refining fine-grained mask boundaries. These results suggest that explicitly separating semantic interpretation from spatial prompting is a promising direction for complex remote-sensing segmentation tasks where visual appearance alone is ambiguous. More details are provided in the Appendix.

AI USE STATEMENT

Generative AI tools were used solely to improve the clarity, grammar, fluency, and academic style of the manuscript. They were not used to generate research ideas or hypotheses, develop the proposed method, design or conduct experiments, implement the method, create or process data, analyze or interpret results, formulate scientific claims, produce figures, or identify and verify references. All AI-assisted edits were carefully reviewed by the authors to ensure that they preserved the intended technical meaning and accurately reflected the conducted research. The authors take full responsibility for the final content, claims, and presentation of this paper.

REFERENCES

Marius Aasan, Martine Hjelkrem Tan, Nico Catalano, Changkyu Choi, and Adín Ramírez Rivera. Differentiable hierarchical visual tokenization. *Advances in Neural Information Processing Systems*, 38:87605–87634, 2026.

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

Dominic Cheng, Renjie Liao, Sanja Fidler, and Raquel Urtasun. Darnet: Deep active ray network for building segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7431–7439, 2019.

Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European conference on computer vision*, pp. 540–557. Springer, 2022.

Jiaqi Huang, Zunnan Xu, Jun Zhou, Ting Liu, Yicheng Xiao, Mingwen Ou, Bowen Ji, Xiu Li, and Kehong Yuan. Sam-r1: Leveraging sam for reward feedback in multimodal segmentation via reinforcement learning. In *Advances in Neural Information Processing Systems*, 2025.

Louis Martin Hugo Touvron, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*, 2023.

Ioannis Kotaridis and Maria Lazaridou. Remote sensing image segmentation advances: A meta-analysis. *ISPRS Journal of Photogrammetry and Remote Sensing*, 173:309–322, 2021.

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*, pp. 9579–9589, 2024.

Kaiyu Li, Ruixun Liu, Xiangyong Cao, Xueru Bai, Feng Zhou, Deyu Meng, and Zhi Wang. Segearth-ov: Towards training-free open-vocabulary segmentation for remote sensing images. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10545–10556, 2025a.

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

Yuang Liu, Zhiheng Qiu, and Xiaokai Qin. Nested-tnt: Hierarchical vision transformers with multi-scale feature processing. *arXiv preprint arXiv:2404.13434*, 2024.

Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*, 2025a.

Yuqi Liu, Tianyuan Qu, Zhisheng Zhong, Bohao Peng, Shu Liu, Bei Yu, and Jiaya Jia. Visionreasoner: Unified visual perception and reasoning via reinforcement learning. *arXiv preprint arXiv:2505.12081*, 2025b.

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

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Jee-Young Sun, Seung-Wook Kim, Sang-Won Lee, Ye-Won Kim, and Sung-Jea Ko. Reverse and boundary attention network for road segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pp. 0–0, 2019.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv:2302.13971*, 2023.

Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018(1):7068349, 2018.

Yu Wang, Yi Wang, Rui Dai, Yujie Wang, Kaikui Liu, Xiangxiang Chu, and Yansheng Li. Urban socio-semantic segmentation with vision-language reasoning. *arXiv preprint arXiv:2601.10477*, 2026.

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*, pp. 11686–11695, 2022.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

Shuping Xiong, Xiuyuan Zhang, Haoyu Wang, Yichen Lei, Ge Tan, and Shihong Du. Mapping the first dataset of global urban land uses with sentinel-2 imagery and poi prompt. *Remote Sensing of Environment*, 327:114824, 2025.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Liang Yao, Fan Liu, Hongbo Lu, Chuanyi Zhang, Rui Min, Shengxiang Xu, Shimin Di, and Pai Peng. Remotereasoner: Towards unifying geospatial reasoning workflow. *arXiv preprint arXiv:2507.19280*, 2025.

Zijun Yao, Yanjie Fu, Bin Liu, Wangsu Hu, and Hui Xiong. Representing urban functions through zone embedding with human mobility patterns. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*, 2018.

Zuyao You and Zuxuan Wu. Seg-r1: Segmentation can be surprisingly simple with reinforcement learning. *arXiv preprint arXiv:2506.22624*, 2025.

Haobo Yuan, Xiangtai Li, Tao Zhang, Yueyi Sun, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, et al. Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. *arXiv preprint arXiv:2501.04001*, 2025.

Xiuyuan Zhang, Shihong Du, and Qiao Wang. Hierarchical semantic cognition for urban functional zones with vhr satellite images and poi data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 132:170–184, 2017.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025a.

Qi Zhu, Jiangwei Lao, Deyi Ji, Junwei Luo, Kang Wu, Yingying Zhang, Lixiang Ru, Jian Wang, Jingdong Chen, Ming Yang, et al. Skysense-o: Towards open-world remote sensing interpretation with vision-centric visual-language modeling. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14733–14744, 2025b.

Yida Zhu, Liying Chen, Daping Xiong, Shuiping Chen, Fangxiao Du, Jinghua Hao, Renqing He, and Zhizhao Sun. C-aoi: Contour-based instance segmentation for high-quality areas-of-interest in online food delivery platform. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5750–5759, 2023.

A Dataset Introduction

SocioSeg Wang et al. (2026) is an urban socio-semantic segmentation benchmark designed to fill the gap between conventional remote sensing segmentation and socially defined urban understanding. Unlike traditional datasets that mainly focus on physical entities such as buildings, roads, water bodies, and vegetation, SocioSeg targets social-semantic entities such as schools, hospitals, parks, residential communities, and commercial areas. These entities often do not have stable or distinctive visual appearances in satellite imagery, and their boundaries are shaped by social function, map context, and human-defined semantics rather than by low-level texture alone. Therefore, SocioSeg requires models to jointly reason over satellite images and digital maps, making it closer to open-vocabulary, referring, and reasoning-based segmentation than to closed-set land-cover classification.

A key design of SocioSeg is its hierarchical task structure. The benchmark contains three progressively more abstract levels: *Socio-name*, *Socio-class*, and *Socio-function*. Socio-name is the most fine-grained setting, where the model is required to segment a specific named entity instance, such as a particular university or hospital, and contains over 5,000 fine-grained instances. Socio-class evaluates category-level understanding over more than 90 social-semantic categories, such as college, hospital, park, and shopping mall. Socio-function further abstracts these categories into more than 10 high-level functional concepts, such as education, healthcare, recreation, and commerce. This hierarchical design enables systematic evaluation from instance-level localization to category-level recognition and finally to high-level functional reasoning.

SocioSeg also introduces a practical multimodal geospatial representation by pairing satellite imagery with rendered digital maps. Instead of directly using heterogeneous raw geospatial data such as POIs and road networks, the dataset renders public map layers containing roads and points of interest into an image-like modality. This design has several advantages: the map layer is easier to access than restricted raw geospatial records, it is naturally aligned with satellite imagery in geographic space, and it provides rich social-semantic cues that are difficult to infer from visual appearance alone. The dataset contains over 13,000 samples, each consisting of a satellite image, a digital map, and the corresponding socio-semantic mask annotation. Following SocioReasoner Wang et al. (2026), the dataset is split into training, validation, and testing sets with a 6:1:3 ratio, and the task is evaluated using cIoU, gIoU, and F1 to measure both pixel-level overlap and instance-level segmentation quality.

B Baseline Introduction

We compare DISP with representative baselines from three groups. The first group contains standard semantic segmentation models. **UNet** (Ronneberger et al., 2015) is a classical encoder-decoder segmentation architecture originally proposed for biomedical image segmentation. Its contracting path captures high-level contextual information, while its expanding path recovers spatial resolution through upsampling. The skip connections between encoder and decoder layers fuse low-level spatial details with high-level semantic features, making UNet a strong and widely used baseline for dense prediction. **SegFormer** (Xie et al., 2021) is a Transformer-based segmentation framework that combines a hierarchical Transformer encoder with a lightweight MLP decoder. Its encoder produces multi-scale visual features without relying on fixed positional encodings, and the decoder aggregates representations from different stages to achieve efficient and robust semantic segmentation.

The second group includes recent reasoning-based segmentation methods for natural images. **VisionReasoner** (Liu et al., 2025b) is a unified reasoning-perception framework that uses reinforcement learning to enhance visual reasoning across detection, segmentation, and counting tasks. It generates structured reasoning processes before producing final predictions, enabling a single model to handle diverse perception queries. **Seg-R1** (You & Wu, 2025) explores pure reinforcement learning for pixel-level understanding. It trains a multimodal large model to generate point and box prompts in a next-token manner, which are then used to guide SAM2 for mask generation, showing strong zero-shot transfer to referring and reasoning segmentation. **SAM-R1** (Huang et al., 2025) further integrates fine-grained segmentation feedback into multimodal reasoning by using SAM as a reward provider. Through task-specific rewards and tailored optimization objectives, SAM-R1 improves the alignment between reasoning outputs and segmentation masks with a limited number of training samples.

Algorithm 1 Training and Inference Pipeline of DISP

Require: Satellite image I_s, digital map image I_m, text query T, ground-truth mask Y for training, selected ViT layers $\mathcal{S}$
Ensure: Predicted mask $\hat{Y}$
Encode (I_s, I_m) with the split-path dual-image encoder.
Insert **[SEG]** and **[IMG]** into the multimodal prompt.
Feed images and query T into the MLLM to obtain hidden states $\mathbf{h}_s$ and $\mathbf{h}_i$.
Project $\mathbf{h}_s$ and $\mathbf{h}_i$ into sparse prompts $\mathbf{p}_s$ and $\mathbf{p}_i$ using separate heads.
Decode the mask $\hat{Y}$ with the SAM2 decoder conditioned on $\mathbf{p}_s$ and $\mathbf{p}_i$.
if training **then**
    Compute segmentation loss $\mathcal{L}_{\text{seg}}$ with BCE and Dice losses.
    for each layer $\ell \in \mathcal{S}$ **do**
        Extract layer feature $\mathbf{F}_\ell$ from the image encoder.
        Use Y to perform masked attention pooling over $\mathbf{F}_\ell$.
        Match the pooled regional feature with the **[IMG]** representation.
        Compute first-order and second-order MS-MFM losses.
    end for
    Optimize $\mathcal{L}_{\text{LM}} + \lambda_1 \mathcal{L}_{\text{seg}} + \lambda_2 \mathcal{L}_{\text{MS-MFM}}$..
else
    Remove all mask-dependent MS-MFM operations.
    Use only **[SEG]**, **[IMG]**, projection heads, and the SAM2 decoder.
end if
return $\hat{Y}$

The third group consists of advanced remote sensing segmentation and reasoning methods. **SegEarth-OV** (Li et al., 2025a) is an open-vocabulary remote sensing segmentation framework that leverages pretrained vision-language models for annotation-free open-world segmentation. It introduces feature upsampling and global-bias alleviation strategies to recover high-resolution spatial details and improve local semantic fidelity. **RSRefSeg** (Chen et al., 2025) targets referring segmentation in remote sensing images, where models must localize and segment regions specified by free-form natural language descriptions. It addresses remote sensing challenges such as scale variation, complex scenes, and domain-specific terminology by aligning textual descriptions with image regions. **SegEarth-R1** (Li et al., 2025b) introduces geospatial pixel reasoning, where implicit language queries are used to infer target regions and generate segmentation masks. It combines a hierarchical visual encoder, a large language model for instruction parsing, and a customized mask generator adapted to high-resolution remote sensing imagery. **RemoteReasoner** (Yao et al., 2025) formulates remote sensing understanding as a unified geospatial reasoning workflow. It uses multimodal large language models and reinforcement learning to autonomously construct reasoning paths and support multi-granularity outputs across object-, region-, and pixel-level tasks. **SocioReasoner** (Wang et al., 2026) is the most relevant prior method for socio-semantic segmentation. It follows a two-stage render-and-refine strategy, first generating box prompts to obtain coarse masks with SAM and then refining boundaries with point prompts on rendered masks. Since its pipeline relies on non-differentiable prompting and refinement, it adopts GRPO-based reinforcement learning to train the reasoning process end-to-end.

C Algorithm Flow

Algorithm 1 summarizes the overall training and inference procedure of DISP. The key idea is to first obtain decoupled hidden states from the semantic **[SEG]** token and the image-oriented **[IMG]** token, then project them into independent sparse prompts for the SAM2 mask decoder. During training, MS-MFM additionally supervises the **[IMG]** token with masked multi-scale target-region features. We set $\lambda_1 = 0.5$ and $\lambda_2 = 0.1$ in all experiments to keep different loss terms at a comparable numerical scale.

D Prompt Settings

Table 3 summarizes the prompt templates used for different SocioSeg task levels. The templates are designed to explicitly specify the satellite-map dual input, the target social meaning, and the decou-

Table 3: Prompt settings for different SocioSeg task levels.

General Instruction Prompt Given a satellite image and its aligned digital map, identify the region referred to by the user query. Use the satellite image to capture visual appearance and spatial layout, and use the digital map to infer social-semantic context such as roads, points of interest, and functional surroundings. The semantic target is represented by **[SEG]**, while the image-grounded localization cue is represented by **[IMG]**. Predict the segmentation mask for the target region.
Socio-name Prompt The query refers to a specific named urban entity, such as *a particular school, hospital, park, or commercial area*. Locate this entity by jointly considering its name, nearby map labels, surrounding road structure, and visual evidence in the satellite image. Generate a mask for the complete region corresponding to this named entity.
Socio-class Prompt The query refers to a social-semantic category, such as *school*, *hospital*, *park*, or *shopping area*. Segment all regions that match the queried category by combining category-level semantic reasoning with visual-spatial evidence from the paired images.
Socio-function Prompt The query refers to a high-level social function, such as *education*, *healthcare*, *recreation*, or *commerce*. Infer the target region from functional context rather than appearance alone, and use the map information and satellite layout to determine the final segmentation boundary.
Training Response Prompt The response activates **[SEG]** for semantic understanding and **[IMG]** for visual-spatial grounding. These two signals are separately projected and then passed to the mask decoder to produce the final segmentation mask.

pled roles of **[SEG]** and **[IMG]**. Note that the table only shows one representative template used in our implementation; in practice, we adopt multiple semantically equivalent response templates to increase embedding diversity and reduce overfitting to a fixed textual pattern.

E LIMITATIONS

Although DISP achieves strong performance on SocioSeg, its robustness across broader geographic regions, map providers, and temporal remote sensing scenarios remains to be further explored. Since DISP uses digital maps as semantic context, noisy or outdated map cues may also affect localization in some cases, which we leave for future work.