OMNIEARTHSEG: UNIFIED INSTRUCTION-DRIVEN SEGMENTATION FOR HETEROGENEOUS REMOTE SENSING IMAGERY

Anonymous authors
Paper under double-blind review

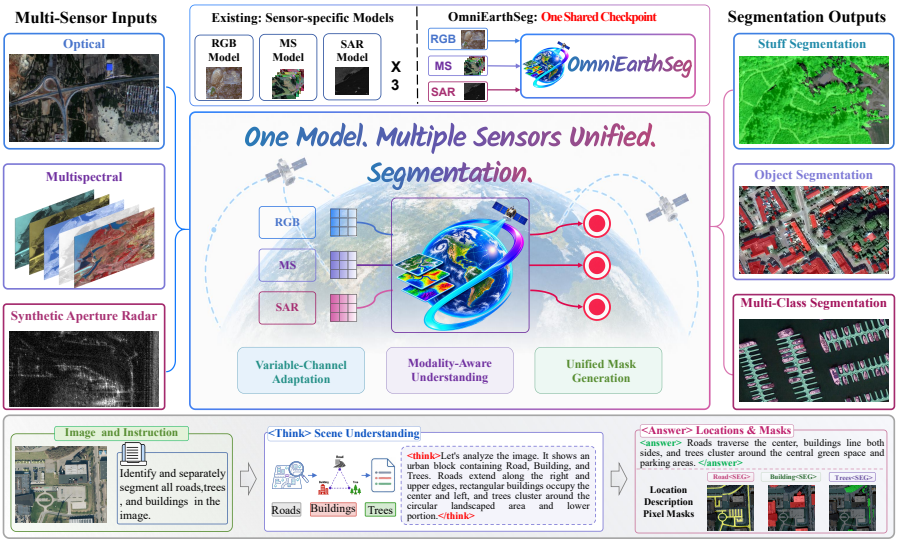


Figure 1: OmniEarthSeg enables unified and interpretable instruction-driven segmentation across heterogeneous remote sensing imagery.

ABSTRACT

Multimodal large language models (MLLMs) enable instruction-driven segmentation of remote sensing (RS) imagery, allowing users to specify targets through natural language. However, existing approaches often rely on modality-specific preprocessing or fixed input configurations, limiting their ability to accommodate variable-channel observations. Their mask decoders also depend primarily on high-level semantic tokens, which provide limited explicit spatial cues for target localization. To address these challenges, we curate OmniGlobal-1.9M, a large-scale dataset comprising 1.9 million instruction-segmentation samples from over 800K globally distributed RS images acquired by 17 sensors and platforms, spanning RGB, multispectral, and synthetic aperture radar modalities. It provides diverse annotations, including multi-granular segmentation masks spanning a broad range of categories and long-form textual supervision that supports interpretable reasoning and response generation. Building on this dataset, we propose OmniEarthSeg, an MLLM that integrates the OmniSensor Tokenizer for unifying variable-channel inputs, the Latent Modality-Aware Modulator for enhancing the model's awareness of modality-specific characteristics, and Prompt-Guided Visual Enhancement for localizing task-relevant visual evidence before segmentation decoding. Extensive experiments demonstrate that OmniEarthSeg achieves state-of-the-art performance on six benchmarks spanning different sensing modalities and generalizes effectively to unseen datasets under zero-shot evaluation. Codes are available at https://anonymous.4open.science/r/OmniEarthSeg-EFC7.

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.

1 INTRODUCTION

Instruction-driven segmentation generates pixel-level masks from natural language instructions, directly grounding linguistic intent in fine-grained spatial regions of visual data. In remote sensing (RS), this paradigm reduces the reliance on predefined task formulations by allowing users to flexibly specify targets of interest through natural language. This capability enables on-demand interpretation of complex geospatial scenes and is particularly valuable for applications requiring rapid target localization and accurate region extraction, including environmental monitoring (Savva et al., 2025; Bi et al., 2025), urban planning (Gao et al., 2024; Wang et al., 2025d), and disaster management (Wang et al., 2025a; Ahn et al., 2025).

Multimodal large language models (MLLMs) (Lai et al., 2024; Ren et al., 2024; Xia et al., 2024; Rasheed et al., 2024) provide a new technical pathway for remote sensing instruction-driven segmentation, enabling pixel-level scene interpretation and interaction through natural language instructions. Recent studies have progressively extended this paradigm from high-resolution pixel grounding to geospatial reasoning, unified instruction-driven segmentation, and broader sensing modalities (Shabbir et al., 2025; Li et al., 2025b; Shu et al., 2026; Ni et al., 2026; Si et al., 2026). For example, GeoPixel (Shabbir et al., 2025) enables pixel-level grounding in high-resolution RS imagery, UniGeoSeg (Ni et al., 2026) further incorporates reasoning segmentation, while SPEX (Si et al., 2026) extends instruction-driven segmentation to multispectral (MS) land-cover extraction.

Despite these advances, in our view, existing methods still face two major limitations. First, they often rely on modality-specific preprocessing or fixed input configurations, limiting their flexibility in handling heterogeneous imagery with varying channel configurations. Second, inspired by LISA (Lai et al., 2024), current instruction-driven segmentation paradigms usually combine semantic representations of `[SEG]` tokens generated by the language model with visual features for segmentation decoding. However, since `[SEG]` tokens primarily encode high-level semantics, they provide only limited fine-grained spatial evidence corresponding to the text instruction. As a result, the segmentation decoder must further identify and extract task-relevant information from global visual features, posing substantial challenges to precise target localization and boundary delineation.

To support unified instruction-driven segmentation across heterogeneous RS imagery, we first construct OmniGlobal-1.9M, a large-scale dataset comprising aligned images, instructions, responses, and segmentation masks. OmniGlobal-1.9M contains over 800K RS images from 17 sensor types across RGB, MS, and SAR modalities and approximately 1.9 million instruction-segmentation samples. Compared with existing RS vision-language datasets, OmniGlobal-1.9M provides more comprehensive support for unified instruction-driven segmentation by jointly encompassing heterogeneous modalities, multi-granular segmentation annotations covering both stuff and object classes under single- and multi-target settings, and long-form textual supervision for interpretable reasoning and response generation, as summarized in Table 1.

We then present OmniEarthSeg, a MLLM for unified instruction-driven segmentation on heterogeneous remote sensing imagery. To accommodate variable-channel inputs, we design an OmniSensor Tokenizer (OST), which maps imagery from different sensors into a unified visual representation. To preserve modality-specific information and explicitly enhance the model's modality awareness, we introduce a Latent Modality-Aware Modulator (LMAM), which extracts latent modality cues from visual representations to enrich the language features. Finally, we propose a Prompt-Guided Visual Enhancement (PGVE) module, which identifies visual evidence relevant to the prompt semantics before segmentation decoding, thereby strengthening semantic-spatial alignment for pixel-level prediction. Extensive experiments demonstrate that OmniEarthSeg achieves state-of-the-art performance across multiple public benchmarks spanning RGB, MS, and SAR imagery and generalizes effectively to unseen datasets under zero-shot evaluation.

Our primary contributions can be summarized as follows:

(1) We curate OmniGlobal-1.9M to provide comprehensive supervision for instruction-driven segmentation. Built from more than 800K RGB, multispectral, and SAR images, it comprises approximately 1.9 million samples that associate multi-granular mask annotations with single- and multi-target instructions covering both stuff and object classes, together with reasoning traces that improve the interpretability of the segmentation process.

Table 1: Comparison of different remote sensing vision-language datasets. The average number of words is computed on the training data.

Dataset	RGB	MS	SAR	Stuff		Object		CoT	Images	Samples	Image Size	Words (Avg.)
				STS	MTS	STS	MTS					
DisasterM3 (Wang et al., 2025b)	✓	✗	✓	✓	✗	✗	✗	✗	26,988	123,010	$[1024^2, 1024^2]$	44.85
EarthVQA (Wang et al., 2024)	✓	✗	✗	✓	✓	✗	✗	✗	6,000	208,593	$[1024^2, 1024^2]$	1.723
SPIE (Si et al., 2026)	✗	✓	✗	✓	✗	✗	✗	✗	403,432	403,432	$[128^2, 650^2]$	142.00
LaSeRS (Xin et al., 2026)	✓	✗	✗	✓	✓	✓	✓	✗	19,168	30,830	$[600^2, 1024^2]$	19.54
Terra-CoT (Shu et al., 2026)	✓	✗	✓	✓	✓	✗	✗	✓	346,758	1,015,223	$[120^2, 1024^2]$	111.07
GeoSeg-1M (Ni et al., 2026)	✓	✗	✗	✓	✓	✓	✗	✗	590,413	1,148,504	$[512^2, 512^2]$	12.18
EarthReason (Li et al., 2025b)	✓	✗	✗	✓	✗	✗	✗	✓	5,434	30,000+	$[170^2, 1075 \times 923]$	20.89
OmniGlobal-1.9M(Ours)	✓	✓	✓	✓	✓	✓	✓	✓	802,812	1,935,675	$[111^2, 1024^2]$	333.89

STS: Single Target Segmentation; *MTS*: Multi Target Segmentation; *VQA*: Visual Question Answering; *CoT*: Chain-of-Thought;

(2) We propose OmniEarthSeg, an MLLM integrating three innovations: the OmniSensor Tokenizer converts variable-channel imagery into unified visual representations; the Latent Modality-Aware Modulator injects latent modality cues into language features to facilitate modality-aware vision–language modeling; and the Prompt-Guided Visual Enhancement module emphasizes task-relevant visual evidence before segmentation decoding.

(3) Experiments demonstrate that OmniEarthSeg outperforms prior approaches on public benchmarks spanning RGB, multispectral, and SAR imagery and generalizes well to unseen datasets under zero-shot evaluation. Qualitative analyses further demonstrate that the proposed modules capture modality-specific information and localize task-relevant visual evidence, validating their effectiveness across diverse sensing modalities.

2 Related Work

2.1 Multimodal Large Language Models

General-purpose multimodal large language models (MLLMs) continue to advance in visual understanding and interaction. Models such as Flamingo (Alayrac et al., 2022), LLaVA (Li et al., 2023), MiniGPT-4 (Zhu et al., 2023), and InstructBLIP (Dai et al., 2023) connect visual encoders with LLMs through learned cross-modal interfaces, enabling general image-level vision–language understanding and instruction following. Subsequent models, including Ferret-v2 (Zhang et al., 2024a), RegionGPT (Guo et al., 2024), and InternGPT (Liu et al., 2023), further support fine-grained object-level visual grounding and interaction. Comparable advances have emerged in remote sensing. RSGPT (Hu et al., 2025) supports image captioning and visual question answering, while GeoChat (Kuckreja et al., 2024) introduces region-aware dialogue and grounding. Further studies enhance semantic and relational understanding (Li et al., 2025d; Luo et al., 2024; Bazi et al., 2024), multi-granular and visual-prompted interpretation (Zhan et al., 2025; Zhang et al., 2024c), spatial and temporal reasoning (Chu et al., 2024; Irvin et al., 2024; Feng et al., 2025), and multi-sensor image understanding (Zhang et al., 2024d; Soni et al., 2025). Nevertheless, they offer only limited support for instruction-driven segmentation, restricting their applicability to pixel-level prediction.

2.2 MLLMs for Instruction-Driven Segmentation

To support instruction-driven segmentation, a growing body of work equips MLLMs with dedicated mechanisms for pixel-level mask generation. A common design introduces specialized segmentation tokens whose hidden representations condition mask decoders. LISA (Lai et al., 2024), PixelLM (Ren et al., 2024), and GLaMM (Rasheed et al., 2024) follow this paradigm, mapping language representations to masks through dedicated decoders, including SAM-based designs (Kirillov et al., 2023). An alternative line casts segmentation as autoregressive generation: Text4Seg (Lan et al., 2024) represents masks as patch-aligned semantic descriptors, whereas SAMTok (Zhou et al., 2026) tokenizes each mask into two discrete tokens for standard next-token prediction. Within remote sensing, LISAt (Quenum et al., 2025), GeoPixel (Shabbir et al., 2025), SegEarth-R1 (Li et al., 2025b), and Think2Seg-RS (Zhang et al., 2026) adapt MLLM-based segmentation to geospatial imagery, enabling pixel-level mask prediction from both explicit target descriptions and reasoning-intensive language queries. SPEX (Si et al., 2026) focuses on multispectral land-cover extraction,

whereas TerraScope (Shu et al., 2026) extends pixel-grounded reasoning to SAR observations. Nevertheless, a unified instruction-driven segmentation framework for RS imagery with varying channel configurations remains underexplored.

3 THE OMNIGLOBAL-1.9M DATASET

OmniGlobal-1.9M is a comprehensive, large-scale dataset characterized by broad geographic coverage, diverse sensing modalities, extensive category coverage, multiple annotation granularities, and rich textual explanations. Its imagery spans geographically diverse regions across six continents and adjacent coastal waters, covering a wide range of latitudes and longitudes, as shown in Figure 2(a). Beyond its diversity in imaging modalities, OmniGlobal-1.9M also exhibits varying channel configurations within the same modality, as illustrated in Figure 2(b). In OmniGlobal-1.9M, a single sample can contain up to 10 categories or more than 1,000 instances (Figure 2(c,d)), with instance areas ranging from 10 pixels to one million pixels (Figure 2(e)). Together, this geographic, modality, and annotation diversity supports robust representation learning and model generalization. Notably, most samples contain more than 256 words of textual annotation (Figure 2(f)), supporting interpretable response generation. Overall, OmniGlobal-1.9M provides a strong data foundation for unified instruction-driven segmentation across heterogeneous remote sensing modalities. Further details of dataset composition and construction are provided in Appendix A.1.

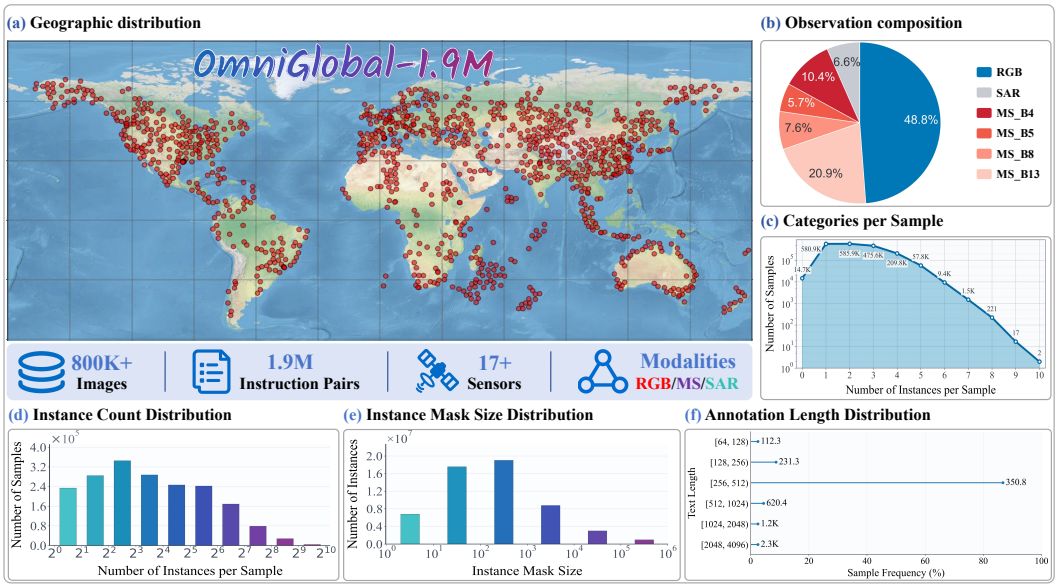


Figure 2: Overview of the diversity and annotation characteristics of OmniGlobal-1.9M. B in (b) denotes the number of channels. Annotation Length in (f) is calculated through number of words.

4 OMNIEARTHSEG

4.1 OVERVIEW

To enable unified instruction-driven segmentation across heterogeneous remote sensing imagery with varying channel configurations, we develop OmniEarthSeg. As illustrated in Figure 3, OmniEarthSeg adopts a mask-decoding paradigm conditioned on LLM hidden representations and comprises three main components: a visual encoder, an LLM, and a mask decoder. The visual encoder extracts hierarchical visual features from heterogeneous sensor observations. The LLM jointly processes high-level visual representations and language instructions to perform cross-modal understanding and generate textual responses that include both reasoning traces and answers. The mask decoder then combines the semantic hidden representations produced by the LLM with visual features to generate pixel-level segmentation masks.

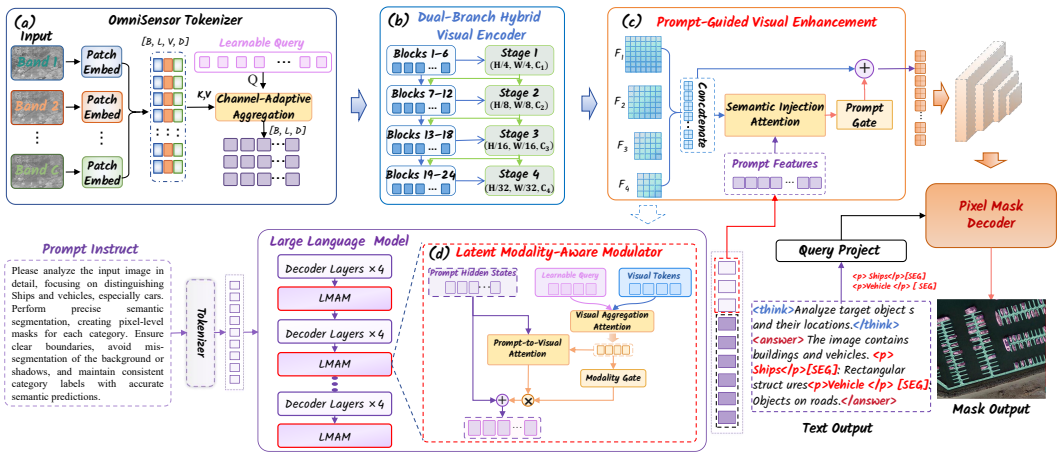


Figure 3: **Overall architecture of OmniEarthSeg.** (a) OST converts variable-channel remote sensing imagery into unified visual tokens. (b) DBHVE extracts hierarchical multi-scale visual features. (c) PGVE enhances task-relevant visual evidence under prompt guidance. (d) LMAM enables modality-aware vision–language interaction within the large language model.

The framework is distinguished by three core designs. First, the OmniSensor Tokenizer (OST) maps inputs with different numbers of channels into a unified token representation before visual encoding, enabling a shared processing pipeline across sensor types. Second, Latent Modality-Aware Modulator (LMAM) layers are inserted into the LLM to enrich the prompt hidden states with task-relevant modality cues inferred from the visual representations, thereby facilitating modality-aware language modeling. Third, the Prompt-Guided Visual Enhancement (PGVE) module uses prompt semantics to emphasize task-relevant regions in the visual features before mask decoding. For the visual backbone, inspired by previous works (Zhang et al., 2024d; Si et al., 2026; Xin et al., 2026), we adopt a CNN-ViT dual-branch architecture, termed the Dual-Branch Hybrid Visual Encoder (DBHVE), to extract multi-scale features that combine local details with global context. Further details of DBHVE are provided in Appendix A.2.1.

4.1.1 OmniSensor Tokenizer

To accommodate varying channel configurations within a shared visual backbone, we introduce the OmniSensor Tokenizer (OST). As illustrated in Figure 3(a), OST represents each spatial patch using a variable number of channel tokens and adaptively aggregates them into a unified patch representation. Given an input image $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, each channel is divided into spatially aligned patches of size $P \times P$. Each single-channel patch is then projected into a shared D-dimensional space using its corresponding channel-specific embedding function ϕ_c:

$$\begin{aligned} \mathbf{t}_{c,n} &= [\phi_c(\mathbf{X}_c)]_n \in \mathbb{R}^D, \\ \mathbf{T}_n &= [\mathbf{t}_{1,n}; \ldots; \mathbf{t}_{C,n}] \in \mathbb{R}^{C\times D}, \end{aligned} \tag{1}$$

where $\mathbf{X}_c$ denotes the cth channel, $n \in \{1, \ldots, N\}$ indexes the spatial patches, $N = (H/P)(W/P)$, and $\mathbf{T}_n$ stacks the tokens from all C channels for the n-th spatial patch.

OST then performs channel-adaptive aggregation (CAA) independently within each spatial patch. CAA comprises a learnable aggregation token and a multi-head attention (MHA) layer. The aggregation token serves as the query, while the tokens from all C channels provide the keys and values. Through attention-based weighting, CAA aggregates a variable number of channel tokens into a single D-dimensional patch representation. The representations from all N patches are stacked:

$$\begin{aligned} \mathbf{z}_n &= \mathrm{MHA}_{\mathrm{CAA}}(\mathbf{q}, \mathbf{T}_n, \mathbf{T}_n) \in \mathbb{R}^{1\times D}, \\ \mathbf{Z} &= [\mathbf{z}_1; \ldots; \mathbf{z}_N] \in \mathbb{R}^{N\times D}, \end{aligned} \tag{2}$$

where $\mathbf{q} \in \mathbb{R}^{1\times D}$ is the learnable aggregation query. The resulting sequence $\mathbf{Z}$ retains the original patch-wise spatial correspondence and is rearranged according to the patch grid before being fed into DBHVE. Details of the channel-specific embedding function ϕ_c are provided in Appendix A.2.2.

4.2 LATENT MODALITY-AWARE MODULATOR

Although OST provides a consistent visual embedding format for heterogeneous inputs, this unification may obscure modality-specific cues. Following Flamingo (Alayrac et al., 2022), we insert an LMAM layer after every four Transformer layers of the LLM. As illustrated in Figure 3(d), LMAM enriches prompt representations with latent modality information to enhance modality awareness without requiring explicit modality labels.

At each selected depth $\ell \in \mathcal{I}_{\mathrm{LMAM}}$, LMAM performs visual aggregation followed by prompt calibration. First, Visual Aggregation Attention (VAA) uses K learnable queries to aggregate modality-related cues from the visual tokens $\mathbf{V}$, producing observation-conditioned latent tokens:

$$\mathbf{M}^{\ell} = \mathrm{MHA}^{\ell}_{\mathrm{VAA}}\left(\mathbf{Q}^{\ell}, \mathrm{LN}^{\ell}(\mathbf{V}), \mathrm{LN}^{\ell}(\mathbf{V})\right) \in \mathbb{R}^{K\times d}, \tag{3}$$

where $\mathbf{Q}^{\ell} \in \mathbb{R}^{K\times d}$ contains the learnable queries shared across samples, LN^{ℓ} denotes layer normalization (LayerNorm), and d is the hidden dimension.

Next, the latent tokens $\mathbf{M}^{\ell}$ are used to calibrate the prompt representations. Specifically, Prompt-to-Visual Attention (P2VA) treats the prompt hidden states $\mathbf{H}^{\ell}_{p}$ as queries and $\mathbf{M}^{\ell}$ as keys and values to retrieve task-relevant modality information. The retrieved information is then incorporated through a gated residual update:

$$\begin{aligned} \mathbf{R}^{\ell}_{p} &= \mathrm{MHA}^{\ell}_{\mathrm{P2VA}}\left(\mathbf{H}^{\ell}_{p}, \mathbf{M}^{\ell}, \mathbf{M}^{\ell}\right), \\ \widehat{\mathbf{H}}^{\ell}_{p} &= \mathbf{H}^{\ell}_{p} + \alpha^{\ell}\mathbf{R}^{\ell}_{p}. \end{aligned} \tag{4}$$

Here, $\mathbf{R}^{\ell}_{p}$ denotes the retrieved modality information, and $\alpha^{\ell} \in (0,1)$ is a sample-adaptive scalar gate computed by applying a sigmoid-activated MLP to the mean-pooled latent tokens. The gate controls the contribution of the retrieved modality information to prompt representations.

4.3 PROMPT-GUIDED VISUAL ENHANCEMENT

As illustrated in Figure 3(c), PGVE uses LMAM-calibrated prompt representations to enhance task-relevant visual evidence in multi-scale features before mask decoding.

Given multi-scale feature maps $\{\mathbf{F}_s\}_{s=1}^{S}$, where $\mathbf{F}_s \in \mathbb{R}^{C\times H_s\times W_s}$, we flatten their spatial dimensions and concatenate the resulting tokens to obtain $\mathbf{V}_{\mathrm{ms}} \in \mathbb{R}^{N_v\times C}$, where $N_v = \sum_{s=1}^{S} H_s W_s$. The prompt representations $\mathbf{P} \in \mathbb{R}^{L_p\times d_p}$ are extracted from the final-layer LLM hidden states corresponding to the prompt tokens between `</img>` and `<think>`, where L_p is the number of prompt tokens and d_p is the hidden dimension.

PGVE uses the visual tokens as queries and the prompt representations as keys and values, followed by a gated residual update:

$$\begin{aligned} \Delta\mathbf{V}_{\mathrm{ms}} &= \mathrm{MHA}_{\mathrm{PGVE}}\left(\mathrm{LN}_v(\mathbf{V}_{\mathrm{ms}}), \mathrm{LN}_p(\mathbf{P}), \mathrm{LN}_p(\mathbf{P})\right), \\ \widehat{\mathbf{V}}_{\mathrm{ms}} &= \mathbf{V}_{\mathrm{ms}} + \mathbf{g}_p \odot \Delta\mathbf{V}_{\mathrm{ms}}. \end{aligned} \tag{5}$$

Here, LN_v and LN_p denote LayerNorm applied to visual and prompt features, respectively. The attention projections align their feature dimensions and produce a C-dimensional update for each visual token. The learnable channel-wise gate $\mathbf{g}_p \in \mathbb{R}^{C}$ controls the strength of this update and is initialized to 0.01 for stable optimization. The operator $\odot$ denotes element-wise multiplication, with the gate broadcast across visual tokens.

Finally, $\widehat{\mathbf{V}}_{\mathrm{ms}}$ is split by scale and reshaped into the enhanced feature maps $\{\widehat{\mathbf{F}}_s\}_{s=1}^{S}$, which are combined with the generated `[SEG]` representations in the mask decoder for pixel-level prediction.

5 EXPERIMENT

5.1 EXPERIMENT SETTINGS

Table 2: Quantitative results of different methods on RGB, MS and SAR datasets. Metric adopts mIoU and oIoU (%). **Bold** and underlined indicate the best and second-best scores, respectively.

Methods	RGB				MS				SAR				Avg	
	OEM		iSAID		FLAIR		BEN		OEM-SAR		HRSID			
Open-Vocabulary Segmentation Models														
RSKT-Seg (Li et al., 2026)	33.5	43.5	40.8	66.6	36.3	49.2	20.9	28.1	9.8	16.2	36.2	50.8	29.6	42.4
SegEarth-OV (Li et al., 2025a)	31.2	37.2	41.6	70.8	24.6	28.3	17.1	24.7	7.5	15.2	61.1	97.6	30.5	45.6
SegEarth-OV3 (Li et al., 2025c)	42.6	48.4	54.3	79.1	33.8	38.8	13.7	17.2	7.9	14.6	67.4	96.4	36.6	49.1
SkySense-O (Zhu et al., 2025)	42.1	52.3	65.0	94.1	44.4	55.2	26.2	42.4	9.4	19.8	60.1	99.2	41.2	60.5
Instruction-tuned Segmentation Models														
LISA (Lai et al., 2024)	45.6	52.7	52.1	82.3	44.8	57.6	34.4	61.5	15.5	36.4	78.1	97.9	45.1	64.7
PixelLM (Ren et al., 2024)	33.3	43.8	47.5	85.4	34.1	45.8	33.7	60.9	13.0	29.0	67.3	98.6	38.1	60.6
GeoPixel (Shabbir et al., 2025)	50.9	56.8	65.6	93.5	51.6	63.3	40.2	68.1	21.1	40.5	82.0	99.4	51.9	70.3
EarthMind (Shu et al., 2025)	50.2	57.0	66.1	94.4	49.0	60.6	38.3	66.9	15.1	35.7	82.5	99.3	50.2	69.0
SegEarth-R1 (Li et al., 2025b)	47.8	47.6	65.7	92.8	44.2	50.2	34.7	54.5	21.7	39.2	85.1	99.2	49.9	63.9
UniGeoSeg (Ni et al., 2026)	30.5	35.1	53.2	89.2	26.2	28.0	22.1	30.3	17.9	29.2	82.1	99.3	38.7	51.9
SegEarth-R2 (Xin et al., 2026)	47.6	54.5	55.8	92.6	50.3	54.7	36.6	50.1	13.1	27.5	76.6	**99.5**	46.7	63.2
OmniEarthSeg(Ours)	**53.7**	**61.6**	**66.2**	**94.7**	**58.6**	**69.0**	**50.2**	**77.6**	**28.1**	**47.1**	**85.9**	**99.5**	**57.1**	**74.9**

OEM: OpenEarthMap, *BEN:* BigEarthNet-S2.

Implementation Details. We use Qwen2.5-3B (Qwen et al., 2025) and SAM2 (Ravi et al., 2024) as the language backbone and mask decoder, respectively. The two branches of DBHVE are initialized with the pretrained weights of InternViT-300M-448px-V2_5 (Wang et al., 2025c) and ConvNeXt (Liu et al., 2022), respectively, while the mask decoder is initialized with pretrained SAM2 weights. All input images are resized to 512×512 pixels. We adopt a multi-stage strategy to train OmniEarthSeg, as detailed in Appendix A.3.1. Training is conducted on 8 NVIDIA A100 GPUs using bfloat16 precision and DeepSpeed ZeRO Stage 2 (Rajbhandari et al., 2020). We optimize the model using AdamW (Loshchilov & Hutter, 2019), with an initial learning rate of 1×10^{-4} and a cosine decay schedule. The batch size is set to 4 per device, without gradient accumulation.

Metrics. Following prior studies (Liu et al., 2024; Yuan et al., 2024), we use overall Intersection-over-Union (oIoU) and mean Intersection-over-Union (mIoU) as the evaluation metrics.

5.2 Main Results

5.2.1 Comparative Experiments

We compare OmniEarthSeg with representative open-vocabulary and instruction-tuned segmentation methods on six remote sensing benchmarks covering three modalities: RGB imagery from OpenEarthMap (Xia et al., 2023) and iSAID (Waqas Zamir et al., 2019), SAR imagery from OpenEarthMap-SAR (Xia et al., 2025) and HRSID (Wei et al., 2020), and multispectral (MS) imagery from FLAIR (Garioud et al., 2023) and BigEarthNet-S2 (Clasen et al., 2025). All evaluation benchmarks are kept strictly disjoint from OmniGlobal-1.9M. The open-vocabulary methods are evaluated directly using their officially released pretrained checkpoints, whereas the instruction-tuned methods are trained on OmniGlobal-1.9M under the same training setting. Because most approaches were originally designed for three-channel RGB imagery, they cannot directly process MS and SAR inputs. We therefore extract RGB channels from MS images and apply channel replication to SAR images to construct three-channel inputs. The comparison methods use these converted inputs, whereas OmniEarthSeg directly processes the original observations. Except for the input channel configuration, all other experimental settings are kept identical to ensure a fair comparison.

The results in Table 2 show that OmniEarthSeg achieves the best performance across all six benchmarks, outperforming the strongest open-vocabulary baseline, SkySense-O (Zhu et al., 2025), by an average of 15.9 mIoU and 14.4 oIoU points. This substantial margin highlights the limitations of adapting RGB-oriented models to heterogeneous sensor inputs through channel conversion alone. Notably, OmniEarthSeg also outperforms SegEarth-OV3, despite the latter being built on the advanced SAM3 architecture (Carion et al., 2026) and fine-tuned on remote sensing data. Even when trained on OmniGlobal-1.9M, the strongest instruction-tuned baseline still trails OmniEarthSeg by an average of 5.2 mIoU and 4.6 oIoU points. This result indicates that instruction tuning on multi-

Table 3: Zero-shot segmentation performance in terms of mIoU (%). **Bold** and underlined values indicate the best and second-best results, respectively.

Methods	DLRSD	ZS	APS
Open-Vocabulary Models			
RSKT-Seg (Li et al., 2026)	31.1	22.5	25.3
ConInfer (Chen et al., 2026)	36.7	32.3	12.5
SegEarth-OV3 (Li et al., 2025c)	44.1	34.2	20.5
SkySense-O (Zhu et al., 2025)	43.8	33.4	28.1
Instruction-Tuned Models			
LISA (Lai et al., 2024)	38.4	15.5	17.2
LISAt (Quenum et al., 2025)	38.2	18.7	19.7
PixelLM (Ren et al., 2024)	35.8	30.5	31.9
GeoPixel (Shabbir et al., 2025)	19.5	17.7	28.0
EarthMind (Shu et al., 2025)	32.5	28.5	27.7
SegEarth-R1 (Li et al., 2025b)	11.9	7.6	25.4
UniGeoSeg (Ni et al., 2026)	26.9	29.7	34.7
SegEarth-R2 (Xin et al., 2026)	29.1	21.8	37.7
OmniEarthSeg (Ours)	**46.7**	**34.3**	**43.1**

DLRSD (RGB); ZS: Zurich Summer (MS) ; APS: AIR-PolSAR (SAR).

Table 4: Ablation Study on OST, LMAM, and PGVE (mIoU, %). *: the MS and SAR inputs are converted into three-channel images.

OST	LMAM	PGVE	RGB	MS	SAR	Avg.
✗	✗	✗	47.7	29.1*	15.8*	30.9
✓	✗	✗	46.5	40.3	16.9	34.6
✓	✓	✗	58.7	46.4	22.6	42.6
✓	✗	✓	59.3	46.7	22.8	42.9
✓	✓	✓	59.6	46.7	23.1	43.1

RGB: iSAID; MS: BigEarthNet-S2; SAR: OpenEarthMap-SAR.

Table 5: Generalizability across MLLM Backbones (mIoU, %).

Methods	Backbone	Full	RGB	MS	SAR
Baseline	Qwen2.5-VL	✗	26.9	35.4	56.0
Ours	Qwen2.5-VL	✓	44.9	47.5	82.8
Baseline	InternVL2.5	✗	28.7	36.2	64.4
Ours	InternVL2.5	✓	47.2	55.4	82.2

RGB: OpenEarthMap; MS: FLAIR; SAR: HRSID; Full: all proposed modules enabled.

source data alone is insufficient to fully address sensor heterogeneity and achieve accurate mask decoding. By combining a unified visual embedding for variable-channel inputs, modality-aware learning, and prompt-guided localization of task-relevant visual evidence, OmniEarthSeg achieves more accurate pixel-level predictions across heterogeneous modalities.

5.2.2 Zero-shot results

We further evaluate zero-shot segmentation performance on three datasets unseen during training, covering RGB imagery from the Dense Labeling Remote Sensing Dataset (DLRSD) (Shao et al., 2018), MS imagery from Zurich Summer (ZS) (Volpi & Ferrari, 2015), and SAR imagery from AIR-PolSAR (APS) (Wang et al., 2025e). To ensure a fair comparison, we directly use the checkpoints obtained for each instruction-tuned segmentation method in the preceding experiments, without further training. As reported in Table 3, OmniEarthSeg continues to achieve the best performance on all three datasets, demonstrating strong generalization across unseen datasets and heterogeneous remote sensing modalities.

5.3 Ablation Analysis.

We conduct ablation experiments on representative RGB, MS, and SAR benchmarks to evaluate the effectiveness of the core components and their generalizability across MLLM backbones. For efficiency, all models in our ablation studies are trained on 25% of OmniGlobal-1.9M. Additional ablation studies are presented in Appendix A.4.

5.3.1 Effectiveness of Core Components

As shown in Table 4, introducing OST alone increases the average mIoU from 30.9% to 34.6%, with improvements on both MS and SAR imagery, particularly an 11.2%-point gain on MS imagery. These results indicate that OST not only provides a unified visual representation but also enables the model to exploit information from additional input channels. Its slight performance decrease on RGB imagery, however, suggests that enforcing a shared embedding across modalities may introduce interference when learning from heterogeneous inputs. Building on OST, adding LMAM or PGVE increases the average mIoU to 42.6% and 42.9%, respectively, with substantial improvements across all three modalities. LMAM facilitates modality-aware vision-language modeling, whereas PGVE localizes task-relevant spatial evidence through prompt semantics. Combining all three components achieves the highest average mIoU of 43.1% and the best or tied-best performance on every modality, confirming their complementary effects.

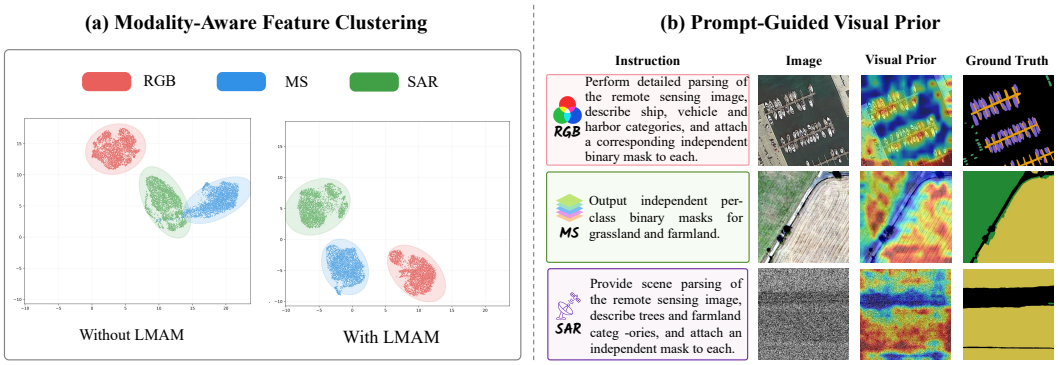


Figure 4: **Visualization of LMAM and PGVE.** (a) Representation spaces of LLMs with or without LMAM. (b) Prompt-guided visual priors generated by PGVE across RGB, MS, and SAR imagery.

5.3.2 Generalizability of the Proposed Modules

To evaluate whether the proposed modules generalize across MLLM backbones and benchmark datasets, we compare Qwen2.5-VL (Bai et al., 2025b) and InternVL2.5 (Chen et al., 2024) with and without the full set of modules under the same mask decoder configuration. Unlike the component analysis, this experiment uses the complementary benchmarks: OpenEarthMap, FLAIR, and HRSID, extending the evaluation to all six datasets introduced above. For the baseline configurations, MS and SAR inputs are converted into three-channel representations using the preprocessing described in the preceding comparison experiments. As shown in Table 5, the proposed modules consistently improve both backbones across RGB, MS, and SAR imagery, indicating that their effectiveness is not tied to a particular MLLM architecture or benchmark set.

5.4 Visualization

To visualize how LMAM and PGVE affect intermediate representations, we conduct two qualitative analyses. First, we pool the prompt features of different modalities extracted from the outputs of LLMs with and without LMAM, and project them into two dimensions using t-SNE. As shown in Figure 4(a), LMAM reduces inter-modality overlap and produces more compact and clearly separated clusters, indicating improved modality separability of prompt representations. Then, for PGVE, we construct visual-prior heatmaps by projecting $\widehat{\mathbf{V}}_{\mathrm{ms}}$ and the pooled prompt representation obtained from $\mathbf{P}$ into a common embedding space, computing their inner products, and mapping the resulting scores back to the original spatial layouts. As shown in Figure 4(b), the high-response regions align with the instruction-specified targets and ground-truth masks, indicating that PGVE effectively localizes and enhances task-relevant spatial evidence across RGB, MS, and SAR imagery.

6 Conclusion

In this work, we introduced OmniGlobal-1.9M, a large-scale dataset containing 1.9 million instruction-segmentation samples across RGB, multispectral, and SAR imagery, with multi-granular pixel-level annotations, broad category coverage, and fine-grained textual supervision. Building on this dataset, we developed OmniEarthSeg, which combines unified visual embedding of heterogeneous sensor inputs, modality-aware vision-language modeling, and prompt-guided enhancement of task-relevant visual evidence. Comparative experiments demonstrate that OmniEarthSeg achieves state-of-the-art performance across six public benchmarks spanning multiple sensing modalities and exhibits strong zero-shot generalization to three unseen datasets. Ablation studies confirm the complementary contributions of the proposed designs and their generalizability across MLLM backbones, while qualitative analyses further demonstrate their effectiveness in improving modality separability and aligning visual priors with target regions. These results provide a strong data and modeling foundation for unified language-guided segmentation across diverse Earth observation sensors.

AI USE STATEMENT

Generative AI tools were used to assist with dataset cleaning and reformatting, including mask filtering, the generation and quality assessment of textual annotations, as well as language polishing of the manuscript to improve clarity and readability. AI-assisted outputs were reviewed and verified by the authors, who take responsibility for the dataset, experimental results, and manuscript content.

ETHICS STATEMENT

The OmniGlobal-1.9M dataset does not intentionally contain personally identifiable or confidential information, and no human-subject data were collected for this work. The proposed dataset and models are intended for research on remote sensing image understanding and segmentation.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide the source code and representative data samples in an anonymous GitHub repository accompanying this submission. The complete OmniGlobal-1.9M dataset will be publicly released following the review process. Additional implementation details, model configurations, training settings, and experimental analyses are provided in the appendix.

REFERENCES

Kyeongjin Ahn, Sungwon Han, Sungwon Park, Jihee Kim, Sangyoon Park, and Meeyoung Cha. Generalizable disaster damage assessment via change detection with vision foundation model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 27784–27792, 2025.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Reza Bahmanyar, Daniela Espinoza-Molina, and Mihai Datcu. Multisensor earth observation image classification based on a multimodal latent dirichlet allocation model. *IEEE Geoscience and Remote Sensing Letters*, 15(3):459–463, 2018. doi: 10.1109/LGRS.2018.2794511.

Lei Bai, Zhongrui Cai, Maosong Cao, et al. Intern-S1: A scientific multimodal foundation model, 2025a. URL https://arxiv.org/abs/2508.15763.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

MR Bayanlou and M Khoshboresh-Masouleh. Multi-task learning from fixed-wing uav images for 2d/3d city modelling. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 44:1–5, 2021.

Yakoub Bazi, Laila Bashmal, Mohamad Mahmoud Al Rahhal, Riccardo Ricci, and Farid Melgani. Rs-llava: A large vision-language model for joint captioning and question answering in remote sensing imagery. *Remote Sensing*, 16(9):1477, 2024.

Hanbo Bi, Yingchao Feng, Yongqiang Mao, Jianning Pei, Wenhui Diao, Hongqi Wang, and Xian Sun. Agmtr: Agent mining transformer for few-shot segmentation in remote sensing. *International Journal of Computer Vision*, 133(4):1780–1807, 2025.

Adrian Boguszewski, Dominik Batorski, Natalia Ziemba-Jankowska, Tomasz Dziedzic, and Anna Zambrzycka. Landcover.ai: Dataset for automatic mapping of buildings, woodlands, water and

roads from aerial imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1102–1110, 2021. doi: 10.1109/CVPRW53098.2021.00121.

Marc Bosch, Kevin Foster, Gordon Christie, Sean Wang, Gregory D. Hager, and Myron Brown. Semantic stereo for incidental satellite images. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1524–1532, 2019. doi: 10.1109/WACV.2019.00167.

Wenxiao Cai, Ke Jin, Jinyan Hou, Cong Guo, Letian Wu, and Wankou Yang. Vdd: Varied drone dataset for semantic segmentation. *Journal of Visual Communication and Image Representation*, 109:104429, 2025.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, et al. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations*, pp. 138846–138923, 2026.

Wenyang Chen, Zhanxuan Hu, Yaping Zhang, Hailong Ning, and Yonghang Tai. Coninfer: Context-aware inference for training-free open-vocabulary remote sensing segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 7408–7418, June 2026.

Yu Chen, Yao Wang, Peng Lu, Yisong Chen, and Guoping Wang. Large-scale structure from motion with semantic constraints of aerial images. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 347–359. Springer, 2018.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

Meng Chu, Zhedong Zheng, Wei Ji, Tingyu Wang, and Tat-Seng Chua. Towards natural language-guided drones: Geotext-1652 benchmark with spatial relation matching. In *European Conference on Computer Vision*, pp. 213–231. Springer, 2024.

Kai Norman Clasen, Leonard Hackel, Tom Burgert, Gencer Sumbul, Begüm Demir, and Volker Markl. reben: Refined bigearthnet dataset for remote sensing image analysis. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 1264–1268, 2025. doi: 10.1109/IGARSS55030.2025.11242834.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

Jie Feng, Shengyuan Wang, Tianhui Liu, Yanxin Xi, and Yong Li. Urbanllava: A multi-modal large language model for urban intelligence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6209–6219, 2025.

Tianyi Gao, Wei Ao, Xing-ao Wang, Yuanhao Zhao, Ping Ma, Mengjie Xie, Hang Fu, Jinchang Ren, and Zhi Gao. Enrich distill and fuse: Generalized few-shot semantic segmentation in remote sensing leveraging foundation model's assistance. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2771–2780, 2024.

Anatol Garioud, Nicolas Gonthier, Loic Landrieu, Apolline De Wit, Marion Valette, Marc Poupée, Sébastien Giordano, et al. Flair: a country-scale land cover semantic segmentation dataset from multi-source optical imagery. *NeurIPS*, 36:16456–16482, 2023.

Qiushan Guo, Shalini De Mello, Hongxu Yin, Wonmin Byeon, Ka Chun Cheung, Yizhou Yu, Ping Luo, and Sifei Liu. Regiongpt: Towards region understanding vision language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13796–13806, 2024.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2018. doi: 10.1109/CVPR.2018.00745.

Yuan Hu, Jianlong Yuan, Congcong Wen, Xiaonan Lu, Yu Liu, and Xiang Li. Rsgpt: A remote sensing vision language model and benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 224:272–286, 2025.

Jeremy Andrew Irvin, Emily Ruoyu Liu, Joyce Chuyi Chen, Ines Dormoy, Jinyoung Kim, Samar Khanna, Zhuo Zheng, and Stefano Ermon. Teochat: A large vision-language assistant for temporal earth observation data. *arXiv preprint arXiv:2410.06234*, 2024.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 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*, 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*, pp. 9579–9589, 2024.

Mengcheng Lan, Chaofeng Chen, Yue Zhou, Jiaxing Xu, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Text4seg: Reimagining image segmentation as text generation. *arXiv preprint arXiv:2410.09855*, 2024.

Bingyu Li, Haocheng Dong, Da Zhang, Zhiyuan Zhao, Hao Sun, and Junyu Gao. Exploring efficient open-vocabulary segmentation in the remote sensing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(8):5982–5991, Mar. 2026. doi: 10.1609/aaai.v40i8.37521.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36: 28541–28564, 2023.

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.

Kaiyu Li, Shengqi Zhang, Yupeng Deng, Zhi Wang, Deyu Meng, and Xiangyong Cao. Segearth-ov3: Exploring sam 3 for open-vocabulary semantic segmentation in remote sensing images. *arXiv preprint arXiv:2512.08730*, 2025c.

Xue Li, Guo Zhang, Hao Cui, Shasha Hou, Shunyao Wang, Xin Li, Yujia Chen, Zhijiang Li, and Li Zhang. Mcanet: A joint semantic segmentation framework of optical and sar images for land use classification. *International Journal of Applied Earth Observation and Geoinformation*, 106: 102638, 2022. ISSN 1569-8432. doi: https://doi.org/10.1016/j.jag.2021.102638.

Zhenshi Li, Dilxat Muhtar, Feng Gu, Yanglangxing He, Xueliang Zhang, Pengfeng Xiao, Guangjun He, and Xiaoxiang Zhu. Lhrs-bot-nova: Improved multimodal large language model for remote sensing vision-language interpretation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 227:539–550, 2025d.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26648–26658, 2024. doi: 10.1109/CVPR52733.2024.02517.

Zhaoyang Liu, Yinan He, Wenhai Wang, Weiyun Wang, Yi Wang, Shoufa Chen, Qinglong Zhang, Zeqiang Lai, Yang Yang, Qingyun Li, et al. Interngpt: Solving vision-centric tasks by interacting with chatgpt beyond language. *arXiv preprint arXiv:2305.05662*, 2023.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Junwei Luo, Zhen Pang, Yongjun Zhang, Tingzhu Wang, Linlin Wang, Bo Dang, Jiangwei Lao, Jian Wang, Jingdong Chen, Yihua Tan, et al. Skysensegpt: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding. *arXiv preprint arXiv:2406.10100*, 2024.

Ye Lyu, George Vosselman, Gui-Song Xia, Alper Yilmaz, and Michael Ying Yang. Uavid: A semantic segmentation dataset for uav imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 165:108 – 119, 2020. ISSN 0924-2716. doi: https://doi.org/10.1016/j.isprsjprs.2020.05.009. URL http://www.sciencedirect.com/science/article/pii/S0924271620301295.

Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2017.

Yu Meng, Ligao Deng, Zhihao Xi, Jiansheng Chen, Jingbo Chen, Anzhi Yue, Diyou Liu, Kai Li, Chenhao Wang, Kaiyu Li, Yupeng Deng, and Xian Sun. Irsamap: Toward large-scale, high-resolution land cover map vectorization. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–19, 2025. doi: 10.1109/TGRS.2025.3600249.

Shuo Ni, Di Wang, He Chen, Haonan Guo, Ning Zhang, and Jing Zhang. Unigeoseg: Towards unified open-world segmentation for geospatial scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 34865–34876, June 2026.

OpenAI. GPT-4o system card, 2024.

Jerome Quenum, Wen-Han Hsieh, Tsung-Han Wu, Ritwik Gupta, Trevor Darrell, and David M Chan. Lisat: Language-instructed segmentation assistant for satellite imagery. *arXiv preprint arXiv:2505.02829*, 2025.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, et al. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models, 2020. URL https://arxiv.org/abs/1910.02054.

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*, 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. URL https://arxiv.org/abs/2408.00714.

Bo Ren, Shibin Ma, Biao Hou, Danfeng Hong, Jocelyn Chanussot, Jianlong Wang, and Licheng Jiao. A dual-stream high resolution network: Deep fusion of gf-2 and gf-3 data for land cover classification. *International Journal of Applied Earth Observation and Geoinformation*, 112:102896, 2022. ISSN 1569-8432. doi: 10.1016/j.jag.2022.102896. URL https://www.sciencedirect.com/science/article/pii/S156984322200098X.

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*, pp. 26374–26383, 2024.

Caleb Robinson, Le Hou, Kolya Malkin, Rachel Soobitsky, Jacob Czawlytko, Bistra Dilkina, and Nebojsa Jojic. Large scale high-resolution land cover mapping with multi-resolution data. In *CVPR*, pp. 12726–12735, 2019.

Franz Rottensteiner, Gunho Sohn, Jaewook Jung, Markus Gerke, Caroline Baillard, Sebastien Benitez, and Uwe Breitkopf. The isprs benchmark on urban object classification and 3d building reconstruction. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, I-3:293–298, 2012. doi: 10.5194/isprsannals-I-3-293-2012.

Antonis Savva, Christos Kyrkou, Panayiotis Kolios, and Theocharis Theocharides. Advances in remote sensing and artificial intelligence for vegetation monitoring in power line corridors: A review and future directions. *IEEE Geoscience and Remote Sensing Magazine*, 13(3):415–439, 2025. doi: 10.1109/MGRS.2025.3558741.

Akashah Shabbir, Mohammed Zumri, Mohammed Bennamoun, Fahad S Khan, and Salman Khan. Geopixel: Pixel grounding large multimodal model in remote sensing. *arXiv preprint arXiv:2501.13925*, 2025.

Zhenfeng Shao, Ke Yang, and Weixun Zhou. Performance evaluation of single-label and multi-label remote sensing image retrieval using a dense labeling dataset. *Remote Sensing*, 10(6), 2018. ISSN 2072-4292. doi: 10.3390/rs10060964.

Jacob Shermeyer, Daniel Hogan, Jason Brown, Adam Van Etten, Nicholas Weir, Fabio Pacifici, Ronny Hänsch, Alexei Bastidas, Scott Soenen, Todd Bacastow, and Ryan Lewis. Spacenet 6: Multi-sensor all weather mapping dataset. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 768–777, 2020. doi: 10.1109/CVPRW50498.2020.00106.

Qian Shi, Da He, Zhengyu Liu, Xiaoping Liu, and Jingqian Xue. Globe230k: A benchmark dense-pixel annotation dataset for global land cover mapping. *Journal of Remote Sensing*, 3:0078, 2023.

Yan Shu, Bin Ren, Zhitong Xiong, Danda Pani Paudel, Luc Van Gool, Begüm Demir, Nicu Sebe, and Paolo Rota. Earthmind: Leveraging cross-sensor data for advanced earth observation interpretation with a unified multimodal llm. *arXiv preprint arXiv:2506.01667*, 2025.

Yan Shu, Bin Ren, Zhitong Xiong, Xiao Xiang Zhu, Begüm Demir, Nicu Sebe, and Paolo Rota. Terrascope: Pixel-grounded visual reasoning for earth observation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16712–16722, 2026.

Dongchen Si, Di Wang, Erzhong Gao, Xiaolei Qin, Liu Zhao, Jing Zhang, Minqiang Xu, Jianbo Zhan, Jianshe Wang, Lin Liu, Bo Du, and Liangpei Zhang. Spex: A vision-language model for land cover extraction on spectral remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–1, 2026. doi: 10.1109/TGRS.2026.3670308.

Sagar Soni, Akshay Dudhane, Hiyam Debary, Mustansar Fiaz, Muhammad Akhtar Munir, Muhammad Sohail Danish, Paolo Fraccaro, Campbell D Watson, Levente J Klein, Fahad Shahbaz Khan, et al. Earthdial: Turning multi-sensory earth observations to interactive dialogues. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14303–14313, 2025.

Xin-Yi Tong, Gui-Song Xia, and Xiao Xiang Zhu. Enabling country-scale land cover mapping with meter-resolution satellite imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 196: 178–196, 2023.

Adam Van Etten, Dave Lindenbaum, and Todd M. Bacastow. Spacenet: A remote sensing dataset and challenge series, 2019. URL https://arxiv.org/abs/1807.01232.

Michele Volpi and Vittorio Ferrari. Semantic segmentation of urban scenes by learning local class interactions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1–9, 2015. doi: 10.1109/CVPRW.2015.7301377.

Fengxiang Wang, Hongzhen Wang, Zonghao Guo, Di Wang, Yulin Wang, Mingshuo Chen, Qiang Ma, Long Lan, Wenjing Yang, Jing Zhang, et al. Xlrs-bench: Could your multimodal llms understand extremely large ultra-high-resolution remote sensing imagery? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14325–14336, 2025a.

Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*. Curran Associates, Inc., 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/4e732ced3463d06de0ca9a15b6153677-Paper-round2.pdf.

Junjue Wang, Zhuo Zheng, Zihang Chen, Ailong Ma, and Yanfei Zhong. Earthvqa: Towards queryable earth via relational reasoning-based remote sensing visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, pp. 5481–5489, 2024.

Junjue Wang, Weihao Xuan, Heli Qi, Zhihao Liu, Kunyi Liu, Yuhan Wu, Hongruixuan Chen, Jian Song, Junshi Xia, Zhuo Zheng, and Naoto Yokoya. Disasterm3: A remote sensing vision-language dataset for disaster damage assessment and response. 2025b.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025c.

Yu Wang, Bo Dang, Wanchun Li, Wei Chen, and Yansheng Li. Holitracer: Holistic vectorization of geographic objects from large-size remote sensing imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8482–8491, 2025d.

Zhirui Wang, Liangjin Zhao, Yuelei Wang, Xuan Zeng, Jian Kang, Jian Yang, and Xian Sun. AIR-PolSAR-Seg-2.0: A polarimetric SAR ground terrain classification dataset for large-scale complex scenes. *Journal of Radars*, 14(2):353–365, 2025e.

Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. isaid: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 28–37, 2019.

Kan Wei, Jinkun Dai, Danfeng Hong, and Yuanxin Ye. Mgfnet: An mlp-dominated gated fusion network for semantic segmentation of high-resolution multi-modal remote sensing images. *International Journal of Applied Earth Observation and Geoinformation*, 135:104241, 2024. doi: 10.1016/j.jag.2024.104241.

Shunjun Wei, Xiangfeng Zeng, Qizhe Qu, Mou Wang, Hao Su, and Jun Shi. Hrsid: A high-resolution sar images dataset for ship detection and instance segmentation. *IEEE Access*, 8: 120234–120254, 2020. doi: 10.1109/ACCESS.2020.3005861.

Junshi Xia, Naoto Yokoya, Bruno Adriano, and Clifford Broni-Bediako. Openearthmap: A benchmark dataset for global high-resolution land cover mapping. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6243–6253, 2023. doi: 10.1109/WACV56688.2023.00619.

Junshi Xia, Hongruixuan Chen, Clifford Broni-Bediako, Yimin Wei, Jian Song, and Naoto Yokoya. Openearthmap-sar: A benchmark synthetic aperture radar dataset for global high-resolution land cover mapping [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 13(4): 476–487, 2025. doi: 10.1109/MGRS.2025.3599512.

Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3858–3869, 2024.

Zepeng Xin, Kaiyu Li, Luodi Chen, Wanchen Li, Xiao Yuchen, Hui Qiao, Weizhan Zhang, Deyu Meng, and Xiangyong Cao. Segearth-r2: Towards comprehensive language-guided segmentation for remote sensing images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13199–13210, 2026.

Zhenghang Yuan, Lichao Mou, Yuansheng Hua, and Xiao Xiang Zhu. Rrsis: Referring remote sensing image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

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. ISSN 0924-2716. doi: https://doi.org/10.1016/j.isprsjprs.2025.01.020. URL https://www.sciencedirect.com/science/article/pii/S0924271625000206.

Haotian Zhang, Haoxuan You, Philipp Dufter, Bowen Zhang, Chen Chen, Hong-You Chen, Tsu-Jui Fu, William Yang Wang, Shih-Fu Chang, Zhe Gan, et al. Ferret-v2: An improved baseline for referring and grounding with large language models. *arXiv preprint arXiv:2404.07973*, 2024a.

Pan Zhang, Baochai Peng, Chaoran Lu, Quanjin Huang, and Dongsheng Liu. Asanet: Asymmetric semantic aligning network for rgb and sar image land cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 218:574–587, 2024b. doi: 10.1016/j.isprsjprs.2024.09.025.

Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, Jun Li, and Xuerui Mao. Earthmarker: A visual prompting multimodal large language model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–19, 2024c.

Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, and Xuerui Mao. Earthgpt: A universal multimodal large language model for multisensor image comprehension in remote sensing domain. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–20, 2024d. doi: 10.1109/TGRS.2024.3409624.

Xu Zhang, Junyao Ge, Yang Zheng, Kaitai Guo, and Jimin Liang. Bridging semantics and geometry: A decoupled lvlm–sam framework for reasoning segmentation in optical remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, 237:217–235, 2026. ISSN 0924-2716. doi: https://doi.org/10.1016/j.isprsjprs.2026.04.036.

Yikang Zhou, Tao Zhang, Dengxian Gong, Yuanzheng Wu, Ye Tian, Haochen Wang, Haobo Yuan, Jiacong Wang, Lu Qi, Hao Fei, Shunping Ji, Anran Wang, Zhuochen Wang, Yujing Wang, Cheng Chen, and Xiangtai Li. Samtok: Representing any mask with two words. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 37852–37863, June 2026.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14733–14744, 2025.

A APPENDIX

- Section A.1. Details of OmniGlobal-1.9M Composition and Construction.
- Section A.2. Detailed Architecture of OmniEarthSeg.
- Section A.3. More Details in Experiments
- Section A.4. Additional Ablation Studies.
- Section A.5. Additional Qualitative Results of OmniEarthSeg.
- Section A.6. Datasheet for OmniGlobal-1.9M.

A.1 Details of OmniGlobal-1.9M Composition and Construction

A.1.1 Source datasets of OmniGlobal-1.9M

To construct OmniGlobal-1.9M, we aggregate multiple publicly available remote sensing datasets with pixel- or region-level annotations. We summarize the source datasets used to construct OmniGlobal-1.9M in Table 6. The RGB subset spans UAV, airborne, satellite, and multi-source imagery, with reported spatial resolutions ranging from $0.05\,\mathrm{m}$ to $4.5\,\mathrm{m}$. The multispectral (MS) subset includes Sentinel-2 MSI, WorldView-3, GF-2 PMS, and NAIP and IGN aerial imagery, covering different sensor configurations and spatial resolutions from $0.2\,\mathrm{m}$ to $60\,\mathrm{m}$. The SAR subset incorporates observations from Sentinel-1B, GF-3, TerraSAR-X, TanDEM-X, Umbra, and Capella Space, with reported spatial resolutions ranging from $0.15\,\mathrm{m}$ to $10\,\mathrm{m}$. Together, these data sources introduce substantial variation in channel configurations, spatial resolutions, spectral characteristics, and imaging mechanisms, providing a broad training distribution for learning representations across heterogeneous RGB, MS, and SAR observations. When platform/sensor information or spatial resolution is not reported by the corresponding source dataset, the entry is marked as unspecified or "–", respectively, rather than being estimated.

Next, we provide a brief introduction for each source dataset in sequence.

Chesapeake Land Cover (Robinson et al., 2019) is a high-resolution (1 m) land-cover classification dataset that covers the Chesapeake Bay watershed, including areas in Maryland, Virginia, West Virginia, Pennsylvania, Delaware, and Washington, D.C. The dataset contains raster products derived from NAIP imagery with Red, Green, Blue, and NIR channels, as well as Landsat-8 spectral bands. Its land-cover annotations include water, tree canopy/forest, low vegetation, barren land, other impervious surfaces, roads, and no-data regions. Developed through a collaboration between the Chesapeake Conservancy and the U.S. Geological Survey (USGS), the dataset provides multiple temporal snapshots from 2013/2014, 2017/2018, and 2021/2022, thereby facilitating detailed analyses of land-use and land-cover changes over time.

FLAIR-2: (Garioud et al., 2023) provides nationwide coverage and comprises more than 20 billion annotated pixels from very-high-resolution aerial imagery with a spatial resolution of 0.2 m. The imagery was acquired across different months over three years, forming diverse spatiotemporal domains. Each aerial image patch contains five channels (Red, Green, Blue, Near Infrared, and Elevation) and is paired with semantic annotations comprising 19 classes, of which 13 are used in the baseline experiments. To incorporate broader spatial context and temporal information, the dataset also provides one-year Sentinel-2 time series with 10 spectral bands. In total, more than 50,000 Sentinel-2 acquisitions with a spatial resolution of 10 m are included.

Globe230K (Shi et al., 2023) is a globally distributed land-cover dataset comprising approximately 232,819 image tiles, each with a size of 512×512 pixels. The tiles have a ground sampling distance of 1 m and are annotated with diverse land-use and land-cover categories across a wide range of geographic regions.

LoveDA (Wang et al., 2021) is a semantic segmentation dataset constructed from high-resolution imagery with a spatial resolution of 0.3 m obtained through Google Earth. The dataset covers both urban and rural areas and includes diverse scenes with varying spatial layouts and domain characteristics. It defines six foreground land-cover classes and provides pixel-level annotations for all categories.

Potsdam (Bayanlou & Khoshboresh-Masouleh, 2021) is a benchmark dataset for aerial image segmentation that comprises 38 orthorectified image tiles, each with a size of 6000×6000 pixels and

Table 6: Source datasets of OmniGlobal-1.9M, along with their platforms/sensors and spatial resolutions.

Modality	Source Dataset	Platform / Sensor	Spatial Resolution
RGB	VDD	DJI Mavic Air II UAV	–
	UDD6	DJI Phantom 4 UAV	–
	UAVid	DJI Phantom 3 Pro / Phantom 4	–
	iSAID	Google Earth / GF-2 / JL-1 / Aerial	0.1–4.5 m
	Sky-SA	Multi-source RS imagery	–
	SpaceNet-6	WorldView-2	0.5 m
	Potsdam	Airborne / Aerial Camera	0.05 m
	Vaihingen	Airborne / Intergraph ZI DMC	0.09 m
	LoveDA	Google Earth	0.3 m
	IRSAMap	Jilin-1 / Google Earth / Open Cities AI	0.5 m
	Globe230K	Google Earth	1 m
	LandCover.ai	Airborne / Aerial Camera	0.25–0.5 m
	OpenEarthMap	Multi-source Satellite / Airborne / UAV	0.25–0.5 m
	Inria Aerial Image Labeling	Airborne / Aerial Camera	0.3 m
MS	MSLCC	Sentinel-2A / MSI	10 m
	FLAIR-2	IGN Aerial Imagery	0.2 m
	DFC2019	WorldView-3 / VNIR	1.3 m
	SpaceNet-2	WorldView-3 / MS	1.24 m
	BigEarthNet v2.0	Sentinel-2 / MSI	10–60 m
	Five-Billion-Pixels	GF-2 / PMS	4 m
	Chesapeake Land Cover	NAIP Aerial Imagery	1 m
SAR	MSLCC	Sentinel-1B / C-band SAR	10 m
	HRSID	Sentinel-1B / TerraSAR-X / TanDEM-X	0.5 / 1 / 3 m
	DDHR Dataset	GF-3 / C-band SAR	1 m
	PIE-RGB-SAR	GF-3 / C-band SAR	3 m
	WHU-OPT-SAR	GF-3 / C-band SAR	5 m
	YESeg-OPT-SAR	SAR (sensor unspecified)	0.5 m
	OpenEarthMap-SAR	Umbra / Capella Space / X-band SAR	0.15–0.5 m

a ground sampling distance of 5 cm. The imagery contains four spectral bands (R, G, B, and NIR), and the associated ground-truth labels define six semantic classes: impervious surfaces, buildings, low vegetation, trees, cars, and background/clutter.

Vaihingen (Rottensteiner et al., 2012) is an aerial semantic segmentation dataset comprising 33 image patches with a ground sampling distance (GSD) of 9 cm. The dataset provides true orthophotos and the associated digital surface models (DSM), together with pixel-level annotations for six semantic classes: impervious surfaces, buildings, low vegetation, trees, cars, and clutter/background.

Five-Billion-Pixels (Tong et al., 2023) is a land-cover classification dataset derived from Gaofen-2 satellite imagery with a spatial resolution of 4 m. The dataset contains more than five billion labeled pixels distributed across 150 large-scale scenes and provides annotations for 24 land-cover classes.

Inria Aerial Image Labeling (Maggiori et al., 2017) covers a total area of $810\ \mathrm{km}^2$, evenly divided into $405\ \mathrm{km}^2$ for training and $405\ \mathrm{km}^2$ for testing. The dataset provides orthorectified aerial imagery with a spatial resolution of 0.3 m and binary annotations for building and non-building classes, although annotations are publicly available only for the training split. It encompasses diverse urban environments, ranging from densely populated metropolitan areas to small alpine towns.

BigEarthNet v2.0 (Clasen et al., 2025) is a large-scale Sentinel-2 dataset comprising 549,488 image patches derived from 115 tiles acquired across 10 European countries between June 2017 and May 2018. All images are atmospherically corrected using the Sentinel-2 Level-2A sen2cor v2.11 processor. Each patch is accompanied by a pixel-level reference map and multi-label land-cover annota-

tions derived from the CORINE Land Cover database (CLC2018 v2020_u1), supporting large-scale multi-label classification of remote sensing imagery.

DDHR Dataset (Ren et al., 2022) is a multimodal land-cover classification dataset constructed from paired GF-2 optical and GF-3 SAR imagery. The dataset supports the fusion of heterogeneous remote sensing information, particularly under adverse weather conditions.

DFC2019 (Bosch et al., 2019) is a remote sensing benchmark dataset derived from WorldView-3 imagery that provides panchromatic data and eight-band visible and near-infrared (VNIR) data. The dataset comprises 26 images acquired over Jacksonville, Florida, USA, between 2014 and 2016, and 43 images acquired over Omaha, Nebraska, USA, between 2014 and 2015. The ground sampling distance (GSD) is approximately 35 cm for the panchromatic images and 1.3 m for the pan-sharpened VNIR images. The data are organized into geographically non-overlapping tiles, with the airborne LiDAR data and semantic labels projected onto a common spatial reference system.

HRSID (Wei et al., 2020) is a high-resolution SAR dataset developed for ship detection, semantic segmentation, and instance segmentation. It contains 5,604 SAR images and 16,951 annotated ship instances, with annotations following the COCO format. The imagery encompasses diverse acquisition conditions, including different spatial resolutions, polarization modes, sea states, maritime regions, and coastal ports. As a benchmark for SAR-based object detection and segmentation, HRSID provides images with spatial resolutions of 0.5 m, 1 m, and 3 m.

IRSAMap (Meng et al., 2025) is a large-scale remote sensing dataset designed for high-resolution, multi-category vector mapping of land cover. By linking raster imagery with vector representations, the dataset supports applications in geospatial AI, semantic segmentation, and automated cartography. It provides fine-grained, geometry-aware annotations for multiple land-cover classes across diverse geographic regions, including urban areas, water bodies, forests, farmland, roads, and buildings. All annotations are available in GIS-ready vector formats, including Shapefile and GeoJSON, allowing direct integration into spatial analysis and mapping pipelines.

iSAID (Waqas Zamir et al., 2019) is a large-scale benchmark dataset for instance segmentation of aerial imagery, comprising 2,806 high-resolution images with 655,451 annotated instances across 15 semantic categories. The dataset features dense annotations, substantial variation in object scale, class imbalance, and diverse object orientations. It also contains numerous small and visually ambiguous objects that require contextual reasoning for accurate recognition. All annotations are provided at the instance level and carefully verified by expert annotators in accordance with strict annotation guidelines.

LandCover.ai (Boguszewski et al., 2021) is an aerial remote sensing dataset developed for the automatic mapping of buildings, woodlands, water bodies, and roads. It comprises orthorectified RGB imagery acquired over Poland at two spatial resolutions, including 33 images with a ground sampling distance (GSD) of 25 cm and eight images with a GSD of 50 cm. The imagery covers a total area of $216.27\ \mathrm{km}^2$. The dataset provides three-channel GeoTIFF images and the associated single-channel annotation masks, all represented in the EPSG:2180 coordinate reference system.

MSLCC (Bahmanyar et al., 2018) is a multi-sensor land-cover classification dataset comprising MS Sentinel-2A and SAR Sentinel-1B imagery acquired over Munich, Berlin, and the surrounding areas in Germany. The dataset provides annotations for four major land-cover classes: agricultural fields, forests, built-up areas, and water bodies.

OpenEarthMap (Xia et al., 2023) is a large-scale land-cover dataset comprising approximately 2.2 million annotated segments derived from 5,000 aerial and satellite images. The dataset covers 97 regions across 44 countries on six continents and provides manually annotated land-cover labels for eight classes, with ground sampling distances ranging from 0.25 to 0.5 m.

OpenEarthMap-SAR (Xia et al., 2025) is a benchmark SAR dataset developed for global high-resolution land-cover mapping. It comprises 1.5 million segments derived from 5,033 aerial and satellite images, each with a size of 1024×1024 pixels. The dataset covers 35 regions in Japan, France, and the United States and provides land-cover labels for eight classes through both manual annotation and pseudo-labeling. The ground sampling distances range from 0.15 to 0.5 m.

PIE-RGB-SAR (Zhang et al., 2024b) is a multimodal remote sensing dataset that provides paired optical (RGB) and SAR imagery for joint interpretation tasks. By exploiting the complementary

information provided by the two modalities, the dataset supports land-cover analysis under complex conditions, as well as research on cross-modal feature fusion, domain alignment, and multimodal semantic segmentation.

Sky-SA (Zhu et al., 2025) is a large-scale open-world remote sensing interpretation dataset comprising 35,000 images and 183,375 fully annotated segment–text pairs. The dataset covers 1,763 open-vocabulary semantic categories and provides dense pixel-level annotations, supporting research on open-world semantic segmentation and vision–language learning.

SpaceNet 2 and SpaceNet 6 (Van Etten et al., 2019; Shermeyer et al., 2020) are two large-scale benchmarks for building footprint extraction. SpaceNet 2 contains 24,586 scenes covering Las Vegas, Paris, Shanghai, and Khartoum, with each scene spanning 200×200 m and represented by a 650×650-pixel tile. Its WorldView-3 products include eight-band MS imagery at 1.24 m resolution and pan-sharpened imagery at 0.3 m resolution, together with 302,701 annotated building footprints. SpaceNet 6 covers approximately $120\ \mathrm{km}^2$ of Rotterdam and comprises 204 quad-polarized X-band SAR image strips and a WorldView-2 optical image strip. The data are processed at 0.5 m resolution and divided into 450×450 m (900×900-pixel) tiles, with annotations for approximately 48,000 unique building footprints.

UAVid (Lyu et al., 2020) is a high-resolution UAV-based semantic segmentation dataset developed for urban scene understanding. It comprises video sequences captured by UAVs, with pixel-level annotations for urban scene categories such as buildings, roads, vegetation, vehicles, and background. The dataset supports fine-grained semantic segmentation under complex urban conditions characterized by substantial variations in object scale and viewpoint.

UDD6 (Chen et al., 2018) is a UAV-based urban semantic segmentation dataset comprising high-resolution aerial images and pixel-level annotations for six land-cover classes: buildings, roads, vegetation, water, bare land, and background. The dataset is designed to support fine-grained urban scene understanding in UAV imagery with diverse viewpoints and substantial variations in object scale.

VDD (Cai et al., 2025) is a high-resolution UAV semantic segmentation dataset comprising 400 images with pixel-level annotations. The images were acquired across diverse scene types and under varying viewing angles, weather conditions, and illumination settings. The dataset defines seven semantic categories: wall, roof, road, water, vehicle, vegetation, and others. It encompasses a wide range of environments, including urban buildings, roads, public facilities, natural landscapes, and rural farmland.

YESeg-OPT-SAR (Wei et al., 2024) is a comprehensively annotated high-resolution remote sensing dataset for land-use classification. It integrates RGB optical imagery with SAR imagery at a spatial resolution of 0.5 m.

WHU-OPT-SAR (Li et al., 2022) comprises RGB and near-infrared (NIR) optical imagery paired with the corresponding SAR imagery. The dataset covers an area of $51,448.56\ \mathrm{km}^2$ and has a spatial resolution of 5 m. To the best of our knowledge, WHU-OPT-SAR is the first and largest comprehensively annotated land-use classification dataset to integrate high-resolution optical and SAR imagery.

A.1.2 Data Pre-processing

The source datasets used to construct OmniGlobal-1.9M differ substantially in spatial resolution, sensor characteristics, annotation formats, category taxonomies, and data types. In particular, segmentation masks may be stored as single-channel class indices, palette-indexed images, or RGB color values, making direct integration impractical. We therefore adopt a unified preprocessing pipeline comprising image cropping and semantic category harmonization before multimodal dataset generation.

Image Cropping. We extract fixed-size patches from each co-registered image–mask pair using a sliding-window strategy, with a stride of either 128 or 256 pixels depending on the size of the source image. By default, images and masks are cropped into 512×512 patches. For LoveDA, IRSAMap, UAVid, UDD6, and VDD, we instead adopt a crop size of 1024×1024 to preserve broader spatial context and fine-grained scene structures. This larger crop size is particularly important for the high-

resolution UAV imagery in UAVid, UDD6, and VDD, where objects are densely distributed across relatively large spatial regions. Windows extending beyond the image boundaries are zero-padded across all image channels. We record the unpadded region using a geometric validity mask and discard patches with a valid-area ratio below 80%. The same cropping coordinates and padding geometry are applied to the corresponding segmentation masks to ensure exact spatial alignment between images and masks.

Category Harmonization. We convert annotations in different formats into single-channel masks encoded with integer class IDs and merge semantically equivalent categories to construct a unified taxonomy of 50 semantic classes, thereby ensuring a consistent semantic space across datasets.

Prompt : You are an expert in quality assessment for remote sensing image segmentation annotations. Please determine whether the mask annotations represented by different colored contours are globally consistent with the actual ground objects in the remote sensing image.
Known correspondence between target categories and contour colors: {Green : Buildings , Blue : Roads Yellow : Water , Cyan : Trees , Orange : Farmland}
Instructions:
1. Each contour color represents the mask boundary of its corresponding category.
2. The contour colors are used only to indicate category identity and spatial location and are not part of the original image.
1. The black borders surrounding the colored contours are used only to enhance visibility.
2. Base your judgment on the visual content inside the contours, object boundaries, and surrounding context. Do not infer the target category solely from the contour color.



Evaluate the annotation quality of the entire image according to the following criteria:
(1) Target validity : Determine whether the contoured regions contain real and visually identifiable ground objects rather than background, shadows, noise, invalid regions, or imaging artifacts.
(2) Category consistency : Determine whether the dominant ground objects within each colored contour match the corresponding category. If many regions belong to other categories or cannot be semantically identified, the annotations should be considered category-inconsistent.
(3) Spatial consistency : Determine whether the contours are generally aligned with the corresponding target objects. If many contours are severely displaced, contain large amounts of irrelevant background, or omit most of the target objects, the annotations should be considered spatially inconsistent.
Decision rules:
"Yes": Most contoured regions contain the corresponding real targets, their categories and spatial locations are generally accurate, and the annotations exhibit no substantial overall deviation.
"No": A large proportion of the contours do not correspond to real targets, or the annotations contain obvious category errors, severe spatial misalignment, extensive invalid regions, or abnormal overall distributions.
Minor boundary errors, local omissions, or a small number of ambiguous regions are acceptable and should not directly result in a "No" decision.
Evaluate the image as a whole and provide only one overall decision. Do not evaluate each category or region separately.
Strictly follow the JSON format below. The value of decision must be either "Yes" or "No". The reason field must summarize the primary justification in one sentence. Do not output your analysis, Markdown formatting, or any additional content.

Figure 5: Prompt template for model based mask filtering.

A.1.3 Dataset Generation

To generate high-quality language annotations, we further develop a scalable data generation pipeline, as illustrated in Figure 6. The individual stages of the pipeline are detailed below.

Mask Filtering During mask filtering, we employ InternVL3-38B (Wang et al., 2025c) as a model-based evaluator to assess candidate mask annotations with respect to target validity, category consistency, and spatial alignment. Instead of overlaying filled masks on the original image, we visualize each candidate region using its boundary contour. This representation clearly delineates the spatial extent of the annotated target while preserving its internal visual appearance and surrounding context, thereby reducing the loss of visual information caused by filled-mask overlays. Given a contour-annotated image and the mapping between contour colors and target categories, InternVL3-38B produces a binary validity decision together with a concise justification. Samples classified as valid are retained for subsequent instruction generation, whereas those classified as invalid are discarded. The complete evaluation prompt and a representative contour visualization are shown in Figure 5.

Instruction Construction We employ the proprietary GPT-4o model (OpenAI, 2024) with task-specific prompt templates to generate four types of instructions:

(1) **Scene-analysis segmentation.** The model first interprets the overall scene and then segments the specified categories, providing both reasoning and segmentation supervision.

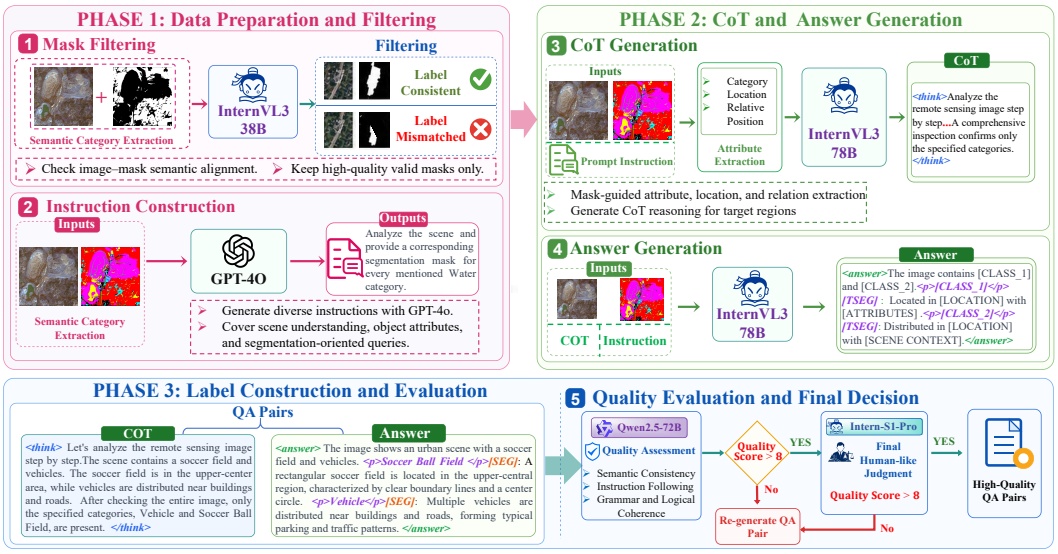


Figure 6: Overview of the OmniGlobal-1.9M dataset generation pipeline.

Prompt : You are a professional remote sensing annotation and instruction-template generation expert. Analyze the semantic meanings, typical visual characteristics, spatial patterns, and common scene contexts of the complete dataset category set, and then generate reusable SCENE instruction templates for remote sensing segmentation.
Dataset category set: {classes}
Task objective:
Analyze {classes} as a whole and generate 100 reusable SCENE instruction templates applicable to any combination of categories within the dataset. In each template, {categories} represents one or more target categories selected from {classes}.
Instruction-type definition:
SCENE templates combine holistic scene understanding with multi-category segmentation. Each template should ask the model to interpret the overall scene, identify all visible regions or instances corresponding to {categories}, and segment each target category separately.
Generation requirements:
1. Generate exactly 100 concise, natural, and linguistically diverse SCENE instruction templates.
2. Each template must contain {categories} exactly once and request a separate segmentation result for every target category it represents.
3. Keep the templates independent of any specific image and do not assume a particular scene type, geographic environment, spatial distribution, or surrounding object.
4. Use only {categories} as the placeholder; do not introduce any additional placeholders.
5. Output only the generated templates as a valid JSON array of strings without explanations or additional content.
Example template structures:
"Describe the overall scene, identify the target categories listed in {categories}, and generate a separate segmentation mask for each."
"Analyze the scene context and produce an individual segmentation mask for every category in {categories}."
"Identify and describe the spatial distribution of the categories in {categories}, then output a separate binary segmentation mask for each."

Figure 7: Prompt template for scene analysis segmentation instruction generation.

(2) **Direct semantic segmentation.** The model directly segments the specified categories without requiring preceding scene or attribute analysis, providing concise segmentation supervision.

(3) **Attribute-analysis segmentation.** The model analyzes the visual and spatial attributes of the target categories before generating their segmentation masks, providing attribute-aware reasoning and segmentation supervision.

(4) **Category-free scene description.** The model generates a holistic description of the image without specifying target categories or predicting segmentation masks, providing only scene-level language supervision.

The corresponding prompt templates are shown in Figure 7, 8, 9, and 10.

CoT and Answer Generation Given an image, its validated instruction, and the target categories derived from the corresponding segmentation mask, we employ InternVL3-78B (Wang et al., 2025c) to generate the CoT rationale and final answer in two successive steps.

Prompt : You are a professional remote sensing annotation and instruction-template generation expert. Analyze the semantic scope represented by the complete dataset category set, and then generate reusable SEMANTIC instruction templates for remote sensing segmentation.
Dataset category set: {classes}
Task objective:
Analyze {classes} as a whole and generate 100 general-purpose semantic segmentation templates applicable to any combination of categories within the dataset. In each template, {categories} represents one or more target categories selected from {classes}.
Instruction-type definition:
SEMANTIC templates should directly ask the model to identify and segment all visible regions or instances corresponding to the target categories. Each target category must be segmented separately.
These templates should focus exclusively on category-level semantic localization and must not rely on scene interpretation, visual attributes, spatial relationships, or other descriptive cues.
Generation requirements:
1. Generate exactly 100 concise, natural, and linguistically diverse SEMANTIC instruction templates.
2. Each template must contain {categories} exactly once and request a separate segmentation result for every target category it represents.
3. Keep the templates direct and focused on semantic segmentation without including scene-level analysis or visual-attribute descriptions.
4. Ensure that the templates remain broadly applicable across different categories, images, sensors, scenes, spatial resolutions, and acquisition conditions.
5. Use only {categories} as the placeholder and output the generated templates as a valid JSON array of strings without explanations or additional content.
Example template structures:
"Identify all visible regions corresponding to the categories in {categories} and segment each category separately."
"Locate every instance belonging to the target categories listed in {categories}, then generate an independent segmentation result for each."
"Perform semantic segmentation for the categories specified in {categories} and delineate the complete extent of each category independently."

Figure 8: Prompt template for direct semantic segmentation instruction generation.

Prompt : You are a professional remote sensing annotation and instruction-template generation expert. Analyze the visual and spatial characteristics represented by the complete dataset category set, and then generate reusable ATTRIBUTE instruction templates for remote sensing segmentation.
Dataset category set: {classes}
Task objective:
Analyze {classes} as a whole and generate 100 general-purpose, attribute-guided segmentation templates applicable to any combination of categories within the dataset. In each template, {categories} represents one or more target categories selected from {classes}.
Instruction-type definition:
ATTRIBUTE templates should guide the model to analyze the observable visual and spatial attributes of the target categories, use these attributes as evidence for locating the corresponding targets, and then segment each target category separately. Attribute analysis serves as an intermediate cue for segmentation rather than the final output.
For each template, select one or more broadly applicable attribute dimensions from [color, shape, texture, size or scale, boundary characteristics, structural arrangement, spatial location, distribution pattern, surrounding context, relationships with nearby objects].
The selected attribute dimensions must be meaningful across different target categories. Keep the descriptions general and do not assign fixed attribute values or assume specific colors, shapes, sizes, locations, surrounding environments, or scene conditions.
Generation requirements:
1. Generate exactly 100 concise, natural, and linguistically diverse ATTRIBUTE instruction templates.
2. Each template must contain {categories} exactly once and require the model to analyze relevant attributes before separately segmenting every target category represented by {categories}.
3. Each template should incorporate one or more relevant attribute dimensions without forcing all attributes into a single instruction.
4. Keep the templates broadly applicable across different categories, images, sensors, scenes, spatial resolutions, and acquisition conditions.
5. Use only {categories} as the placeholder and output the generated templates as a valid JSON array of strings without explanations or additional content.
Example template structures:
"Analyze the color, shape, texture, and spatial distribution of each category in {categories}, then generate a separate segmentation mask for each."
"Use the observable morphological features and spatial relationships to locate every category in {categories}, and delineate each one independently."
"Examine the visible attributes and distribution patterns of the categories listed in {categories}, then segment the complete extent of each target category separately."

Figure 9: Prompt template for attribute analysis segmentation instruction generation.

In the first step, the model generates a visually grounded CoT rationale using the template shown in Figure 11. The rationale first interprets the overall scene and subsequently analyzes the appearance, scale, location, distribution, and spatial relationships of the targets relevant to the instruction. To reduce hallucinations, the reasoning process is constrained by the target categories derived from the segmentation mask and the visual evidence observable in the image. The generated rationale is enclosed within `<think>...</think>`.

Prompt : You are a professional remote sensing annotation and instruction-template generation expert. Analyze the semantic scope, scene diversity, and land-cover composition represented by the complete dataset category set, and then generate reusable CAPTION instruction templates for remote sensing images.
Dataset category set: {classes}
Task objective:
Analyze {classes} as a whole to understand the range of scenes and geographic elements covered by the dataset, and generate 100 general-purpose CAPTION instruction templates. The templates must request holistic descriptions of remote sensing images rather than descriptions of specific categories or particular images.
Instruction-type definition:
CAPTION templates should request a comprehensive description of the remote sensing image, including its overall scene, major land-cover components, prominent objects, visual characteristics, spatial organization, and relationships among scene elements.
Generation requirements:
1. Generate exactly 100 concise, natural, and linguistically diverse CAPTION instruction templates.
2. Each template must request a holistic scene description without specifying any target category or requesting segmentation.
3. Keep the templates broadly applicable across different scenes, sensors, spatial resolutions, geographic regions, and acquisition conditions.
4. Do not use {classes}, category names, category placeholders, or any other placeholders in the generated templates.
5. Output only the generated templates as a valid JSON array of strings without explanations or additional content.
Example template structures:
"Describe the overall scene, its major land-cover components, and their spatial organization."
"Provide a holistic description of the prominent geographic elements and visual characteristics in the remote sensing image."
"Summarize the overall scene content and explain how its major elements are spatially arranged."

Figure 10: Prompt template for category free scene-description generation.

Prompt : Act as a senior remote sensing analyst. Based solely on visual observation of the provided remote sensing image, analyze the landscape and determine which target categories are present and how they are spatially distributed.
Candidate categories:{classes}
Instruction:{Instruction}
Requirements:
1. Analyze only the categories relevant to the provided instruction and visually supported by the image.
2. Describe the observable evidence for each potential category, including color, texture, shape, scale, spatial location, distribution pattern, and surrounding context.
3. Clearly distinguish confidently identified targets from uncertain or ambiguous regions.
4. Do not introduce categories or concepts outside the candidate category list.
5. Do not mention annotations, masks, labels, external instructions, or external information sources.
6. Do not use <class>...</class> or <answer>...</answer> tags.
7. Keep the reasoning objective, focused, and concise.
8. Output only the visual reasoning enclosed within <think>...</think> tags. Do not output any additional content.
Output format:
<think>Alright, let's dive into this remote sensing image! It covers a rural area bordering hills, with gently rolling terrain, lush green vegetation mixed with human-made areas, giving off a vibe of nature and agriculture coexisting nicely. The northern part is full of deep green, with dense, tightly packed tree canopies, suggesting a thick broadleaf forest. Right through the center runs a grayish-white line, pretty even in width, stretching north to south—definitely a main road connecting nearby villages and fields. The southern area has a neat grid-like pattern, light in color with some patches darker than others, likely farmland with different crops or growth stages. Over to the east, there are small, irregular patches of deep blue near the farmland, probably water bodies like irrigation channels or small lakes. I'll use these observations to describe the landscape's layout and how everything fits together.</think>

Figure 11: Prompt template for CoT generation.

In the second step, InternVL3-78B generates the final answer conditioned on the image, instruction, and preceding CoT rationale, following the template shown in Figure 12. The answer is formatted as a concise paragraph enclosed within `<answer>...</answer>`, with each target category mentioned only once and enclosed within a `<class>...</class>` tag. Categories that are not supported by the instruction or corresponding mask annotations are excluded from the answer. For category-free scene-description instructions, the model instead generates a holistic scene description without using class tags.

After completing the two-stage generation process, we post-process the class tags to construct the final segmentation-aware response used for training. Specifically, each target span formatted as `<class>category</class>` is converted into `<p>category</p>[SEG]`, where `<p>...</p>` identifies the corresponding textual target and `[SEG]` serves as its associated segmentation token. This transformation establishes an explicit correspondence between each target mention and the corresponding mask supervision. Category-free scene descriptions contain neither target spans nor `[SEG]` tokens.

Annotation Quality Assessment As shown in step (5) of Figure 6, we conduct a two-stage quality audit for each constituent dataset. Using a fixed random seed, we randomly sample 2,000 image–text

Prompt : Act as a senior remote sensing analyst. Generate the final answer based on the provided instruction and visual reasoning.
Candidate categories: {classes}
Instruction: {Instruction}
Visual reasoning: {Generated_Think}
Requirements:
1. Use the provided visual reasoning as the primary basis for the final answer.
2. Describe only categories that are confidently identified in the visual reasoning and relevant to the instruction.
3. Do not introduce categories outside the candidate category list.
4. Enclose each confirmed category name within <class>category name</class> tags.
5. Use the exact category names provided in the candidate category list.
6. Mention and tag each confirmed category only once.
7. Do not use <class>...</class> tags for uncertain, ambiguous, or unsupported categories.
8. Briefly describe the visible characteristics and spatial distribution of each confirmed category.
9. Write the final answer as a single continuous paragraph using concise, natural language.
10. Do not repeat the reasoning process or output <think>...</think> tags.
11. Output only the final answer enclosed within <answer>...</answer> tags. Do not output any additional content.
Output format:
<answer>The image reveals a rural landscape bordering hills, with gently rolling terrain, vibrant vegetation, and human-made areas blending together, creating a harmonious balance of nature and agriculture. <class>forest</class>: Spread across the northern area, it's a deep green with dense tree canopies, likely a broadleaf forest, covering about a third of the image and acting as a natural shield for the farmland below. <class>roads</class>: A smooth, grayish-white line runs north to south through the center, consistent in width, following the terrain's contours and linking farmland to villages, serving as a main traffic route. <class>farmland</class>: Found in the southern part, it forms a tidy grid pattern, with light green and brown patches showing different crops or growth stages, separated by hedgerows, reflecting traditional farming practices. <class>water bodies</class>: Located in the east, small, irregular deep blue patches, likely irrigation channels or small lakes, sit close to the farmland, supporting agricultural water needs. Spatially, farmland neighbors water bodies, roads tie villages to fields, and forests occupy higher ground, with the overall layout showcasing a beautiful balance between human activity and nature. </answer>

Figure 12: Prompt template for answer generation.

Prompt : You are an image description evaluation expert. Below is a description text containing both a Chain of Thought rationale and a final Answer. Please perform the following tasks:

1. Identify the CoT and Answer parts within the text.

2. Score the CoT and Answer separately, with a maximum of 10 points each:
- 0–3: Severely inconsistent with the image or logically incoherent
- 4–6: Partially reasonable, but with significant omissions or errors
- 7–9: Accurate and mostly complete, with only minor issues
- 10: Perfectly accurate, logically clear, and fully consistent with the image

3. Provide an overall evaluation reason summarizing the accuracy, completeness, and logical consistency of the entire description.

Please output strictly in the following JSON format:
{{"CoT_score": <0-10>, "Answer_score": <0-10>, "overall_reason": "<overall evaluation reason>"}}

Description text: {text}

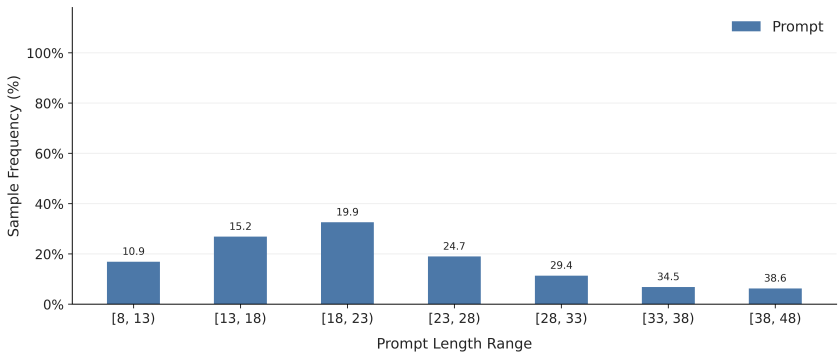
Figure 13: Prompt template for annotation quality assessment.

pairs from each dataset for evaluation. The corresponding evaluation prompt is shown in Figure 13.

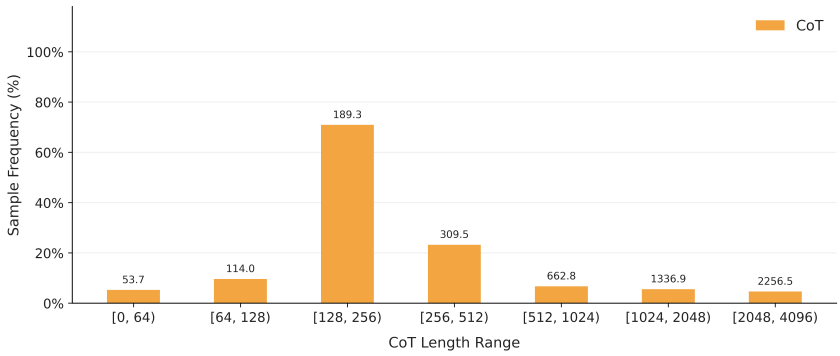
In the first stage, Qwen2.5-VL-72B (Bai et al., 2025b) independently assigns a score from 0 to 10 to the CoT rationale and the final answer. For the i-th sampled pair, the Qwen score q_i is defined as the average of these two scores. Samples with $q_i < 8$ are not forwarded to the second stage, and their Qwen scores are directly used as the corresponding sample-level audit scores. Samples with $q_i \geq 8$ are further evaluated by Intern-S1 (Bai et al., 2025a) using the same evaluation criteria. For each such sample, Intern-S1 independently assigns a score from 0 to 10 to the CoT rationale and the final answer, and the average of these two scores, denoted as s_i, is used as the corresponding sample-level audit score.

Accordingly, the audit score r_i of each sampled pair is defined as

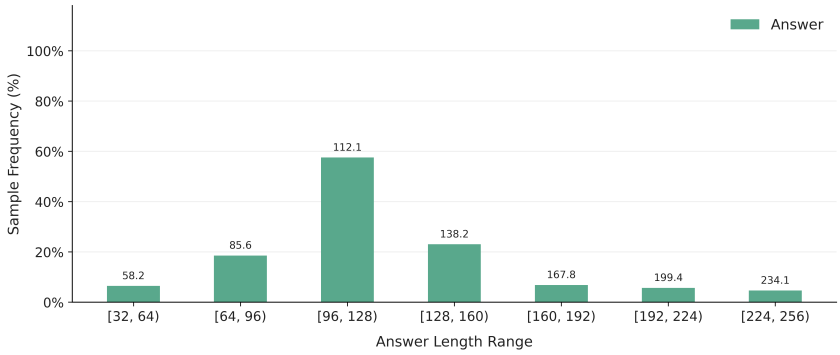
$$r_i = \begin{cases} q_i, & q_i < 8, \\ s_i, & q_i \geq 8. \end{cases} \tag{6}$$



(a) Prompt length statistics.



(b) CoT length statistics.



(c) Answer length statistics.

Figure 14: Word-count distributions of prompts, CoT rationales, and answers in OmniGlobal-1.9M across different length intervals.

The overall quality score of a constituent dataset $\mathcal{D}$ is then computed over all $N = 2{,}000$ sampled pairs:

$$Q_{\mathcal{D}} = \frac{1}{N} \sum_{i=1}^{N} r_i, \tag{7}$$

where $Q_{\mathcal{D}} \in [0, 10]$. These sampled pairs are used solely to estimate the overall annotation quality of each constituent dataset and do not determine the retention or removal of individual samples. If $Q_{\mathcal{D}} < 8$, the entire annotation batch associated with that constituent dataset is regenerated and subsequently subjected to the same quality audit.

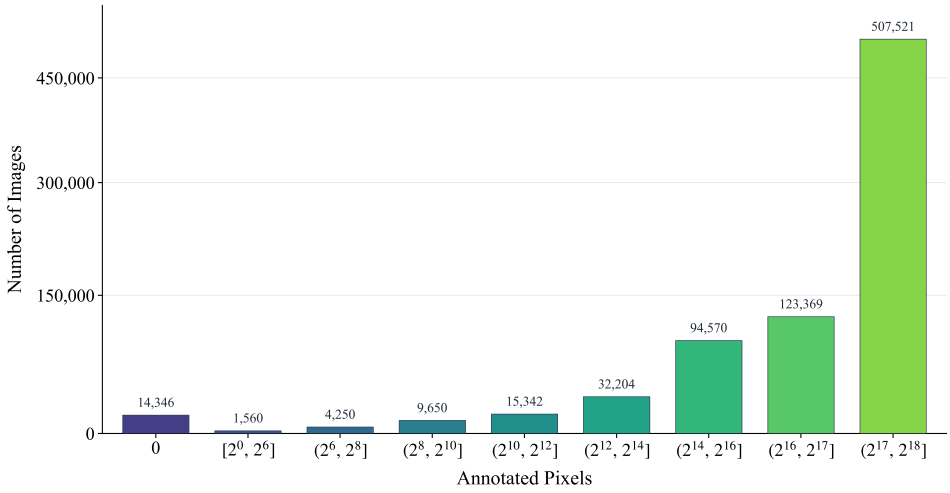


Figure 15: Distribution of annotated areas in OmniGlobal-1.9M.

A.1.4 Dataset Statistics

Text Length Distribution Figure 14 presents the word-count distributions of prompts, CoT rationales, and answers in OmniGlobal-1.9M. For each interval, we additionally report the average lengths of the prompts, CoT rationales, and answers falling within that interval, providing a more fine-grained characterization of their length variations. Prompt lengths range from 8 to 48 words, with the largest proportion falling within the interval of [18, 23] words, reflecting substantial structural diversity in the instructions. CoT rationales and answers exhibit broadly similar distributions but differ in scale. CoT rationales can exceed 2,048 words, with most containing between 128 and 256 words. Answers can exceed 224 words, with the majority falling within the interval of [96, 128] words. These statistics demonstrate that OmniGlobal-1.9M provides rich and interpretable textual supervision through detailed CoT rationales and answers.

Annotation Scale and Category Coverage Figure 15 presents the distribution of annotation areas in OmniGlobal-1.9M, measured by the number of annotated pixels per sample. Most samples contain pixel-level annotations, whose areas span a broad range: some cover fewer than 2^6 pixels, whereas others exceed 2^{17} pixels. Notably, the distribution is concentrated within $2^{17}, 2^{18}$, indicating that nearly the entire image is annotated in many samples. This concentration likely reflects the large proportion of dense land-cover annotations inherited from the source datasets, including categories such as grassland, cropland, and water, as summarized in Table 7. Therefore, in addition to the object-level annotations characterized in Figure 2(e), OmniGlobal-1.9M provides extensive dense land-cover annotations, resulting in broad coverage of both stuff and object categories.

A.2 Detailed Architecture of OmniEarthSeg

A.2.1 Dual-Branch Hybrid Visual Encoder

Remote sensing imagery exhibits substantial variations in spatial resolution and often contains targets with complex spatial structures and boundaries. Although the ViT-based visual encoders commonly adopted in RS MLLMs (Shu et al., 2025; Li et al., 2025b; Ni et al., 2026) effectively capture long-range dependencies and high-level semantics, their limited local inductive biases may hinder the representation of fine-scale spatial details. We therefore design the Dual-Branch Hybrid Visual Encoder (DBHVE), as illustrated in Figure 16(a). DBHVE combines a ViT (Dosovitskiy et al., 2021) branch for global context modeling with a hierarchical ConvNeXt (Liu et al., 2022) branch that progressively extracts multi-scale local features. The ConvNeXt branch produces local feature maps at four spatial scales, while the 24 Transformer blocks of the ViT branch are evenly divided into four stages, each comprising six consecutive blocks. Let $\mathbf{L}_i$ and $\mathbf{G}_i$ denote the local and global

Table 7: Semantic categories included in OmniGlobal-1.9M.

1. Trees	2. Farmland	3. Grassland
4. Water	5. Buildings	6. Barren Land
7. Roads	8. Shrubland	9. Snow
10. Wetland	11. Vineyard	12. Vehicle
13. Tennis Court	14. Parking Lot	15. Baseball Field
16. Bridge	17. Soccer Ball Field	18. Orchard
19. Ship	20. Harbor	21. Plane
22. Solar Panel	23. Greenhouse	24. Basketball Court
25. Highway	26. Ground Track Field	27. Storage Tank
28. Railway	29. Apron	30. Airport
31. Swimming Pool	32. Island	33. Factory
34. Baseball Diamond	35. Roundabout	36. Runway
37. Wind Turbine	38. Intersection	39. Tundra
40. Train Station	41. Pipeline	42. Dam
43. Reservoir	44. Athletic Track	45. Train
46. Container	47. Seat	48. Hangar
49. Badminton Court	50. Military Base	

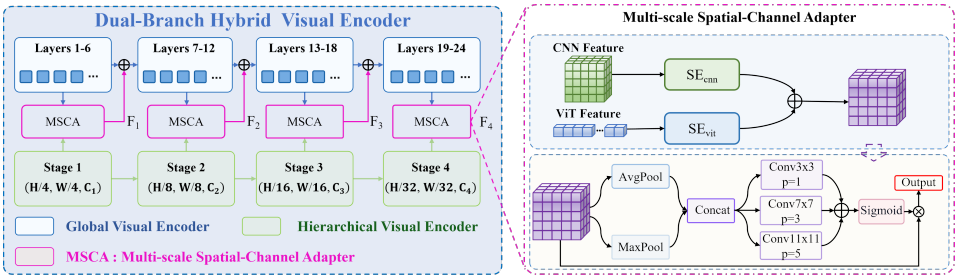


Figure 16: **Architecture of the Dual-Branch Hybrid Visual Encoder (DBHVE).** DBHVE integrates global semantic context and fine-grained spatial details through dual-branch encoding and progressive multi-scale feature fusion with MSCA.

representations produced at stage i, respectively:

$$\mathbf{L}_i = \mathcal{C}_i(\mathbf{L}_{i-1}) \in \mathbb{R}^{D_i \times \frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}}}, \qquad \mathbf{G}_i = \mathcal{T}_i(\mathbf{G}_{i-1}) \in \mathbb{R}^{N \times D}, \quad i = 1, \ldots, 4, \tag{8}$$

where $\mathcal{C}_i$ denotes the i-th ConvNeXt stage, and $\mathcal{T}_i$ denotes the corresponding ViT stage consisting of Transformer blocks $\{\mathcal{B}_j\}_{j=6(i-1)+1}^{6i}$. The resulting pairs $(\mathbf{L}_i, \mathbf{G}_i)_{i=1}^{4}$ provide complementary local and global representations at multiple semantic levels. Finally, ViT layers provide the multi-level backbone representations, and the third feature $\mathbf{F}_3$ (see Figure 3 (c)) is selected for interaction with the LLM. Table 8 further compares the architectural configuration of OmniEarthSeg with those of the other evaluated methods.

Multi-Scale Spatial Channel Adapter. Since the two branches produce representations with different spatial organizations and channel dimensions, we introduce the Multi-Scale Spatial Channel Adapter (MSCA) to align and fuse their features at each stage, as shown in Figure 16(b). At stage i, MSCA reshapes the ViT tokens $\mathbf{G}_i$ into a spatial feature map and projects the features from both branches ($\mathbf{L}_i$ for CNN) into a common spatial and channel space, yielding $\widetilde{\mathbf{X}}^i_{\text{vit}}$ and $\widetilde{\mathbf{X}}^i_{\text{conv}}$. Branch-specific channel attention first recalibrates the two representations through Squeeze-and-Excitation (Hu et al., 2018), which are then combined through element-wise addition to obtain $\mathbf{X}^i_{\text{fuse}}$. MSCA subsequently concatenates the channel-wise average- and max-pooled descriptors of $\mathbf{X}^i_{\text{fuse}}$ and processes them using parallel convolutions with kernel sizes of 3, 7, and 11. Their outputs are aggregated by element-wise addition followed by a sigmoid function to generate a multi-scale spatial attention map. This map reweights $\mathbf{X}^i_{\text{fuse}}$ to produce the fused representation $\mathbf{F}_i$, yielding multi-scale features that integrate global semantics with local spatial details.

A.2.2 CHANNEL-SPECIFIC EMBEDDING

In the OST module shown in Figure 3(a), each input channel is independently projected into a shared token space using a channel-specific patch embedding. Since the maximum number of input channels in OmniGlobal-1.9M is 13, corresponding to the full band configuration of Sentinel-2 imagery, we define a learnable bank of 13 channel-specific convolution kernels. With an embedding dimension of $D = 1024$ and a patch size of $P = 16$, the complete weight bank is denoted by

$$\mathbf{W} \in \mathbb{R}^{D\times 13\times P\times P}, \qquad P = 16, \tag{9}$$

where $\mathbf{W}_{:,m,:,:} \in \mathbb{R}^{D\times 1\times P\times P}$ contains the D two-dimensional projection kernels associated with the m-th channel position.

We select the required kernels along the channel dimension according to the number of channels in the input image. Under the 13-channel Sentinel-2 band ordering, the first channel position corresponds to the coastal-aerosol band. Therefore, a 13-channel input uses all kernel positions, whereas inputs with fewer channels, including RGB and SAR imagery, are assigned kernels starting from the second position. For an input with C channels, the kernel position associated with its c-th channel is defined as

$$s_C(c) = \begin{cases} c, & C = 13, \\ c+1, & C < 13, \end{cases} \qquad c = 1, \ldots, C. \tag{10}$$

Given the c-th input channel $\mathbf{X}_c \in \mathbb{R}^{1\times H\times W}$, its channel-specific embedding function ϕ_c applies the selected kernel with a stride of P:

$$\widetilde{\mathbf{E}}_c = \phi_c(\mathbf{X}_c) = \mathrm{Conv2D}\left(\mathbf{X}_c; \mathbf{W}_{:,s_C(c),:,:}\right) \in \mathbb{R}^{D\times \frac{H}{P}\times \frac{W}{P}}. \tag{11}$$

Because the kernel size and stride are both $P \times P$, this operation partitions each channel into non-overlapping spatial patches. The resulting feature map is flattened and transposed to obtain

$$\mathbf{E}_c = \mathrm{Flatten}\left(\widetilde{\mathbf{E}}_c\right)^\top = [\mathbf{t}_{c,1}; \ldots; \mathbf{t}_{c,N}] \in \mathbb{R}^{N\times D}, \qquad N = \frac{H}{P}\frac{W}{P}. \tag{12}$$

Here, $\mathbf{t}_{c,n}$ denotes the token generated from the n-th spatial patch of channel c. Tokens from all input channels at the same spatial location are subsequently stacked as

$$\mathbf{T}_n = [\mathbf{t}_{1,n}; \ldots; \mathbf{t}_{C,n}] \in \mathbb{R}^{C\times D}, \tag{13}$$

and passed to CAA for channel-adaptive aggregation. This design maintains independently learnable projections for different channel positions while providing a common embedding dimension for inputs with varying channel configurations.

A.3 MORE DETAILS IN EXPERIMENTS

A.3.1 MULTI-STAGE TRAINING STRATEGY

We train OmniEarthSeg progressively in three stages:

Stage 1: Visual Pretraining. We first pretrain the visual components using large-scale Earth-Dial (Soni et al., 2025) data and a subset of OmniGlobal-1.9M. The training data include Landsat, NAIP, Sentinel-1, and Sentinel-2 imagery spanning RGB, MS, and SAR modalities. Through image captioning, image classification, and visual question answering tasks, this stage learns unified visual representations for heterogeneous RS imagery. We freeze the ViT backbone and optimize only OST and DBHVE, enabling the visual branch to accommodate sensor observations with varying channel configurations.

Stage 2: Vision-Language Instruction Tuning. We then perform instruction tuning on OmniGlobal-1.9M. The LLM is adapted using LoRA, while LMAM is jointly optimized with the LoRA parameters. This stage incorporates latent modality cues into the prompt representations, facilitating modality-aware vision-language modeling and instruction following for heterogeneous RS observations.

Stage 3: Pixel-Level Segmentation Fine-Tuning. Finally, we introduce PGVE and jointly optimize it with the pixel mask decoder. The modality-aware prompt representations produced by LMAM

Table 8: Architectural configurations of the compared methods.

Method	Vision Encoder(s)	Encoder Config.	Language Module	Segmentation Head
Open-vocabulary segmentation methods				
RSKT-Seg	CLIP ViT-L/14 + DINO ViT-B/32 + RemoteCLIP ViT-B/16	Triple encoder	CLIP text encoder	RS-Fusion / RS-Transfer
SegEarth-OV	CLIP ViT-B/16	Single encoder	CLIP text encoder	SimFeatUp
SegEarth-OV3	SAM3 (PE-L+)	Single encoder	SAM3 text encoder	SAM3 mask decoder
SkySense-O	SwinV2-H-8	Single encoder	CLIP-style text encoder	Visual decoder
ConInfer	CLIP ViT-B/16 + DINOv3	Dual encoder	CLIP text encoder	Context-aware inference
Instruction-driven segmentation methods				
LISA	CLIP-L + SAM ViT-H	Dual encoder	Vicuna-7B	SAM
PixelLM	CLIP-L	Single encoder	Vicuna-7B	Pixel decoder
GeoPixel	CLIP ViT-L/14 + SAM2 Hiera-L	Dual encoder	InternLM2-7B	SAM2
EarthMind	InternViT + SAM2 Hiera	Dual encoder	Qwen2.5-3B	SAM2
SegEarth-R1	Swin-B	Single encoder	Phi-1.5 (1.3B)	Mask2Former
UniGeoSeg	Swin-B	Single encoder	Phi-1.5 (1.3B)	Mask2Former
SegEarth-R2	Swin-B + siglip-so400m-patch14-384	Dual encoder	Mipha-3B	Mask2Former
OmniEarthSeg	**DBHVE (InternViT + ConvNeXt)**	**Single coupled encoder**	**Qwen2.5-3B**	**SAM2**

guide PGVE in enhancing task-relevant visual evidence, which is combined with the generated `[SEG]` representations for mask prediction. The training objective combines focal loss, cross-entropy loss, and Dice loss.

The three stages are trained for 1, 2, and 2 epochs, respectively. Throughout training, gradient checkpointing is enabled to reduce memory consumption. We assign one data-loading worker to each GPU and employ lazy preprocessing to reduce data preparation overhead. The complete training configuration is provided in Table 9.

A.3.2 Details of Benchmark Splits

Table 10 summarizes the source splits underlying the comparative and zero-shot evaluations reported in Tables 2 and 3, together with the numbers of samples retained in OmniGlobal-1.9M and the corresponding testbeds. All counts refer to the actual image samples obtained after dataset-specific preprocessing, including image cropping for datasets marked with †; further details are provided in Appendix A.1.2. Subsequent data filtering and quality control reduce the sample counts in OmniGlobal-1.9M and the testbeds relative to the preprocessed source splits. Because only the training-set annotations of OpenEarthMap-SAR are publicly available, we manually divide their labeled data into strictly disjoint training and test subsets. DLRSD, Zurich Summer, and AIR-PolSAR are used exclusively for zero-shot evaluation, and none of their samples are included in OmniGlobal-1.9M.

A.4 Additional Ablation Studies

We conduct additional ablation studies to further examine dataset-related components in OmniGlobal-1.9M and architectural design choices in OmniEarthSeg. Unless otherwise specified, all models in these experiments are trained on 25% of OmniGlobal-1.9M for efficient evaluation.

Table 9: Training hyperparameters for the three-stage training procedure.

Shared Configuration	
Hyperparameter	**Value**
Numerical Precision	bfloat16
Computing Hardware	8× NVIDIA A100
Input Resolution	512×512
Maximum Sequence Length	8,192
Per-Accelerator Batch Size	4
Global Batch Size	32
Gradient Accumulation	None
Optimizer	AdamW
Learning Rate Scheduler	Cosine decay
Warmup Ratio	0.01
Weight Decay	0.0
Data-loader Worker Per GPU	1
ZeRO Optimization	Stage 2 (ZeRO-2)

Stage-wise Configuration				
Training Stage	**Epochs**	**Learning Rate**	**Grad. Checkpointing**	**LoRA Rank**
Visual Pretraining	1	1×10^{-4}	Enabled	–
LoRA Fine-tuning	2	1×10^{-4}	Enabled	32
Pixel Decoder Fine-tuning	2	1×10^{-4}	Enabled	–

LoRA denotes low-rank adaptation.

Table 10: Statistics of benchmark splits for the source datasets and their corresponding subsets in OmniGlobal-1.9M. † indicates the original images of the source dataset were cropped. ††: images are directly resized to 512×512.

Dataset	**Modality**	**Train**		**Validation**	
		Source	**OmniGlobal-1.9M**	**Source**	**Testbed**
iSAID†	RGB	36,182	35,973	11,793	11,663
OpenEarthMap†	RGB	25,884	24,598	4,314	4,272
FLAIR	MS	61,712	60,846	15,700	13,717
BigEarthNet-S2	MS	106,256	83,508	11,805	11,445
HRSID††	SAR	3,643	3,099	1,961	1,955
OpenEarthMap-SAR†	SAR	37,982	32,986	–	4,996
DLRSD	RGB	2,100	–	–	2,100
Zurich Summer†	MS	197	–	–	197
AIR-PolSAR	SAR	–	–	309	309

A.4.1 Ablation Study of CoT Supervision and Training Data Scale

We first isolate the effect of CoT supervision by training models on 25% of OmniGlobal-1.9M with and without CoT while keeping all other settings fixed. As shown in Table 11, CoT supervision increases the average mIoU from 60.8 to 61.6, with consistent gains across RGB, MS, and SAR imagery. This improvement suggests that explicit reasoning supervision may encourage the model to interpret visual scenes and task instructions more thoroughly, thereby benefiting segmentation prediction. We then examine the effect of data scale by training with 25%, 50%, 75%, and 100% of OmniGlobal-1.9M while enabling CoT supervision in all settings. The average mIoU steadily increases from 61.6 to 66.2 as the amount of training data grows, confirming the benefit of additional training samples. However, the marginal gains become smaller at larger data scales, suggesting

that performance begins to approach saturation under the current model and training configuration. Overall, both CoT supervision and increased training data contribute to the performance of OmniEarthSeg, although the benefit of further data scaling gradually diminishes.

Table 11: Effects of training data scale and CoT across three sensor modalities (mIoU, %).

Data Scale	CoT	RGB	MS	SAR	Avg.
25%	✗	46.0	54.5	81.9	60.8
25%	✓	47.2	55.4	82.2	61.6
50%	✓	49.9	57.1	84.0	63.7
75%	✓	52.0	58.6	85.1	65.2
100%	✓	53.7	59.0	85.9	66.2

RGB: OpenEarthMap; MS: FLAIR; SAR: HRSID.

A.4.2 Ablation Study of DBHVE and MSCA

We evaluate the effectiveness of DBHVE and MSCA. As shown in Table 12, the baseline retains only the ViT branch. Introducing DBHVE establishes a dual-branch visual encoder and substantially improves performance across all three modalities, demonstrating the benefit of jointly capturing global semantics and local spatial details. We then assess MSCA, which enables stage-wise interaction between the ViT and ConvNeXt branches. Without MSCA, features from corresponding stages are aligned through upsampling or downsampling and directly concatenated. Incorporating MSCA increases the average mIoU from 60.5 to 61.6, indicating that its attention-based spatial and channel fusion integrates complementary cross-branch features more effectively than simple feature alignment and concatenation. These results confirm the complementary contributions of DBHVE and MSCA to visual representation learning across heterogeneous RS modalities.

Table 12: Ablation of DBHVE and MSCA (mIoU %).

ViT	DBHVE	MSCA	RGB	MS	SAR	Avg.
✓	✗	✗	23.1	39.4	27.2	29.9
✓	✓	✗	45.0	54.2	82.2	60.5
✓	✓	✓	47.2	55.4	82.2	61.6

RGB: OpenEarthMap; MS: FLAIR; SAR: HRSID.

A.4.3 Ablation Study of Visual Feature Scale

We investigate which hierarchical visual feature level should be provided to the LLM. Specifically, we evaluate $\mathbf{F}_1$–$\mathbf{F}_4$ from DBHVE, as illustrated in Figure 3(c), with the results reported in Table 13. Using the high-resolution $\mathbf{F}_1$ feature incurs substantial computational overhead, requiring 154 training hours even with only 25% of OmniGlobal-1.9M, which makes full-scale training impractical. Lower-resolution features considerably reduce this cost. Among the individual feature levels, $\mathbf{F}_3$ achieves the highest average mIoU of 61.6 with a training time of 35 hours, outperforming both $\mathbf{F}_2$ and $\mathbf{F}_4$.

To isolate the effect of spatial resolution, we upsample $\mathbf{F}_4$ to the spatial dimensions of $\mathbf{F}_3$, denoting the resulting feature as $\mathbf{F}_4^*$. Despite having the same spatial dimensions as $\mathbf{F}_3$, $\mathbf{F}_4^*$ achieves a lower average mIoU of 60.6. This result suggests that performance depends not only on feature resolution but also on the level of semantic abstraction: the LLM benefits more from intermediate features that balance spatial detail and high-level semantics than from either shallow detail-dominant features or highly abstract deep features.

We further evaluate whether combining all hierarchical levels provides additional benefits. Specifically, $\mathbf{F}_1$, $\mathbf{F}_2$, and $\mathbf{F}_4$ are resampled to the spatial dimensions of $\mathbf{F}_3$, concatenated with $\mathbf{F}_3$ along the channel dimension, and projected using a 1×1 convolution. This multi-level combination does not improve upon $\mathbf{F}_3$ alone, indicating that introducing additional feature levels is not necessarily beneficial for vision-language interaction. Considering both segmentation accuracy and computational efficiency, we therefore select $\mathbf{F}_3$ as the visual input to the LLM.

Table 13: Ablation study on visual feature scale selection for LLM input.

Variant	Feat. Res.	F_1	F_2	F_3	F_4	RGB	MS	SAR	Avg.	Time (h)
F_1	128×128	✓	✗	✗	✗	-	-	-	-	154
F_2	64×64	✗	✓	✗	✗	44.4	54.0	81.5	60.0	60
F_4	16×16	✗	✗	✗	✓	45.7	54.6	82.2	60.8	32
$\mathbf{F}_4^*$	32×32	✗	✗	✗	✓	45.5	54.3	81.9	60.6	35
F_{1-4}	32×32	✓	✓	✓	✓	46.2	55.2	82.6	61.3	35
F_3	32×32	✗	✗	✓	✗	47.2	55.4	82.2	61.6	35

RGB: OpenEarthMap; MS: FLAIR; SAR: HRSID.

A.5 Additional Qualitative Results of OmniEarthSeg

A.5.1 Prompt-Guided Visual Prior

We provide additional qualitative results for prompt-guided visual priors generated by PGVE across RGB, multispectral, and SAR remote sensing imagery. As shown in Figure 17–19, the generated visual priors consistently highlight instruction-relevant regions across all three sensor modalities and show clear spatial correspondence with the ground-truth masks.

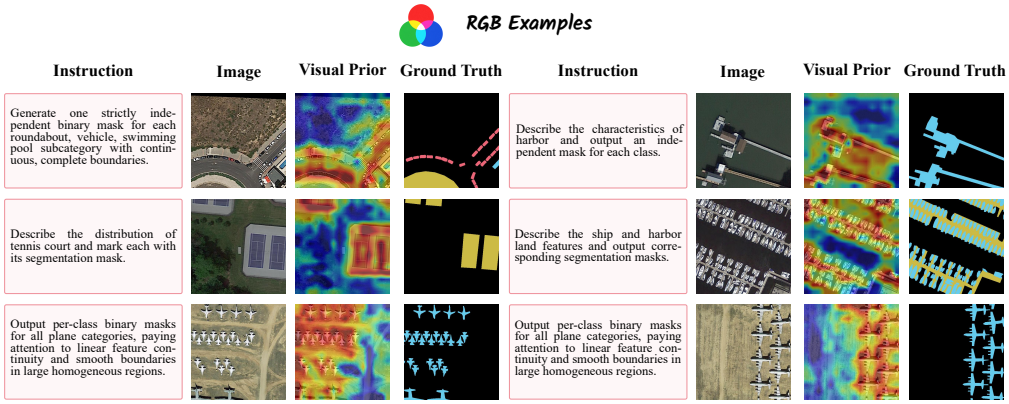


Figure 17: Additional qualitative results of visual priors on RGB remote sensing imagery.

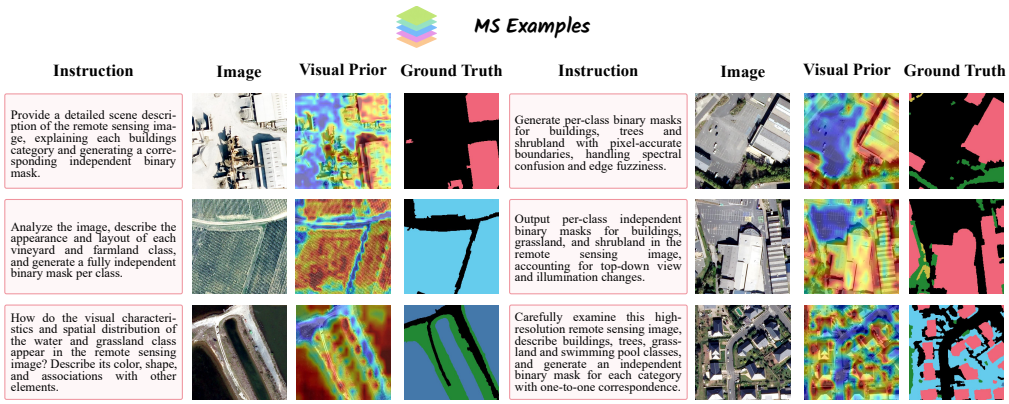


Figure 18: Additional qualitative results of visual priors on multispectral remote sensing imagery.

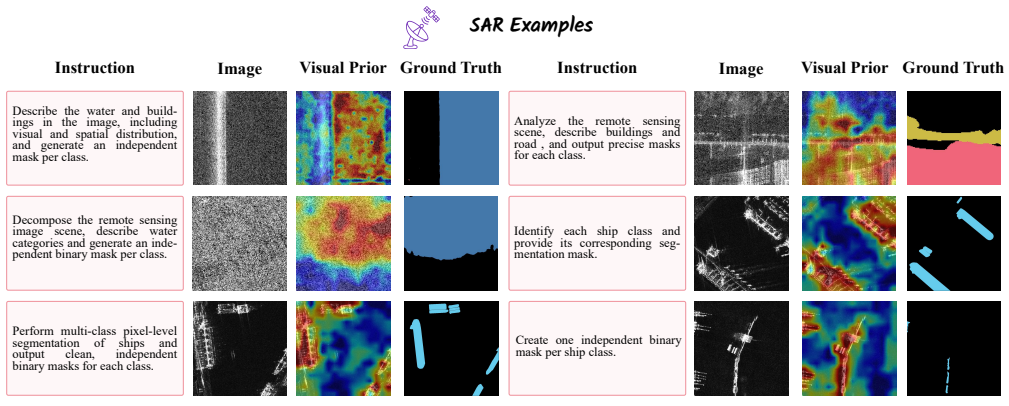


Figure 19: Additional qualitative results of visual priors on SAR remote sensing imagery.

A.5.2 Pixel-level Predictions

We provide additional pixel-level prediction results of OmniEarthSeg across RGB, multispectral, and SAR remote sensing imagery. As shown in Figure 20-25, OmniEarthSeg produces accurate segmentation masks across diverse scenes, target categories, and sensor modalities, covering both single-class and multi-class instructions as well as object- and stuff-level targets.

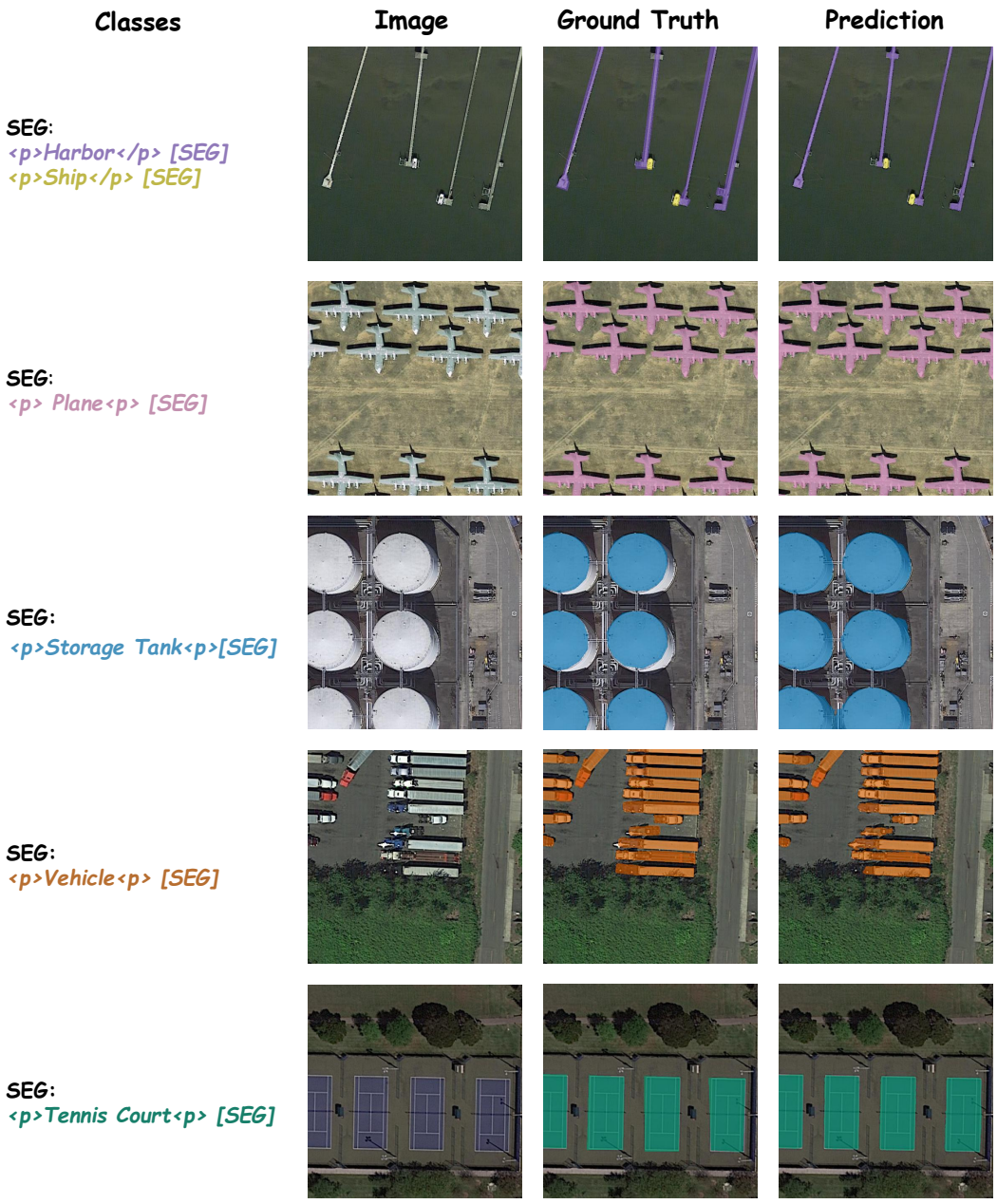


Figure 20: **Additional qualitative prediction results of OmniEarthSeg on RGB imagery.** The examples include both single-class and multi-class segmentation across diverse targets.

A.5.3 Instruction-Driven Segmentation Conversations

We present complete interaction examples of OmniEarthSeg across RGB, multispectral, and SAR imagery. As shown in Figure 26–30, these examples illustrate how OmniEarthSeg interprets user instructions, performs reasoning, and generates pixel-level masks for both semantic classes and individual instances.

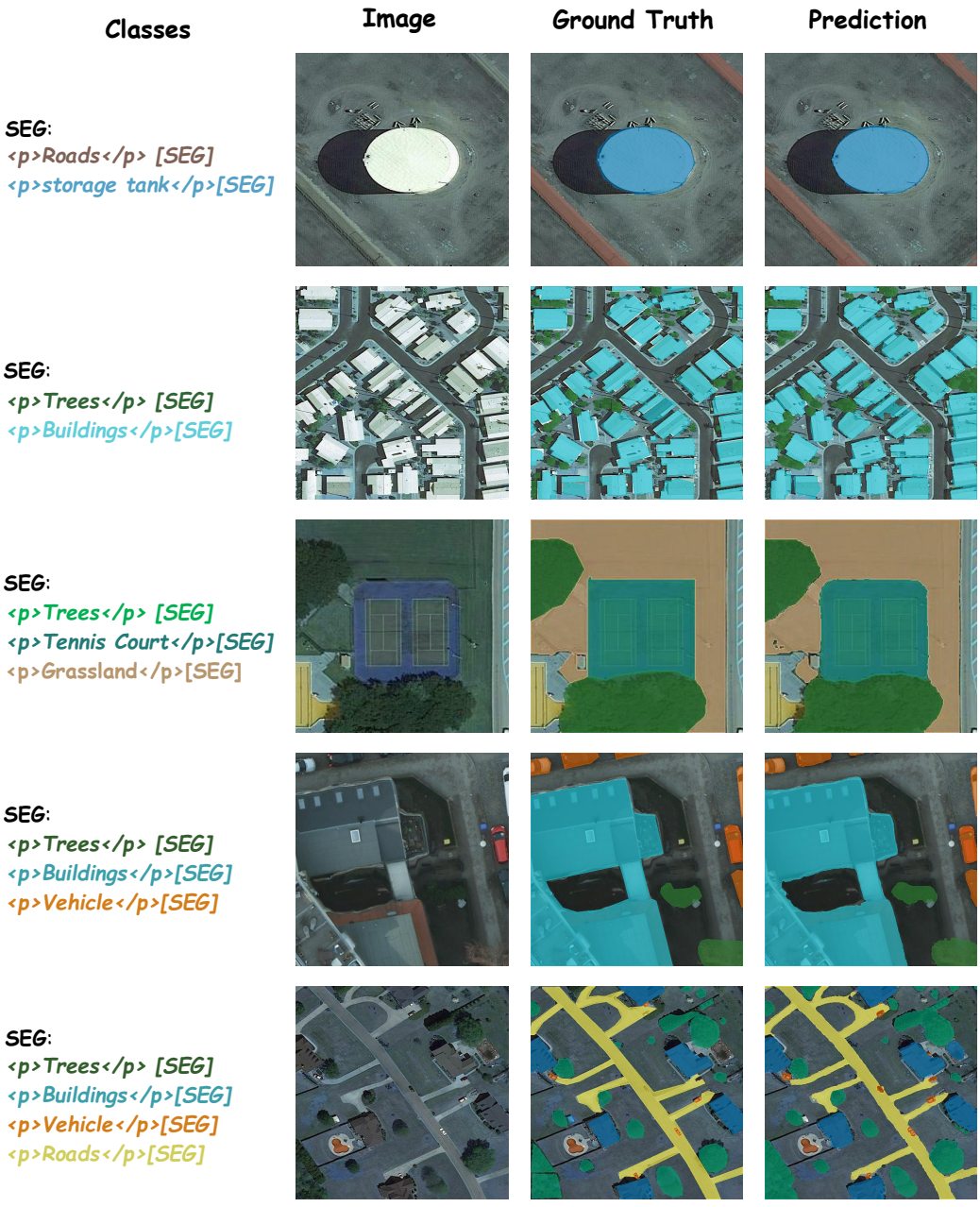


Figure 21: **Additional qualitative prediction results of OmniEarthSeg on RGB imagery.** The examples cover multi-class segmentation with both stuff and object classes.

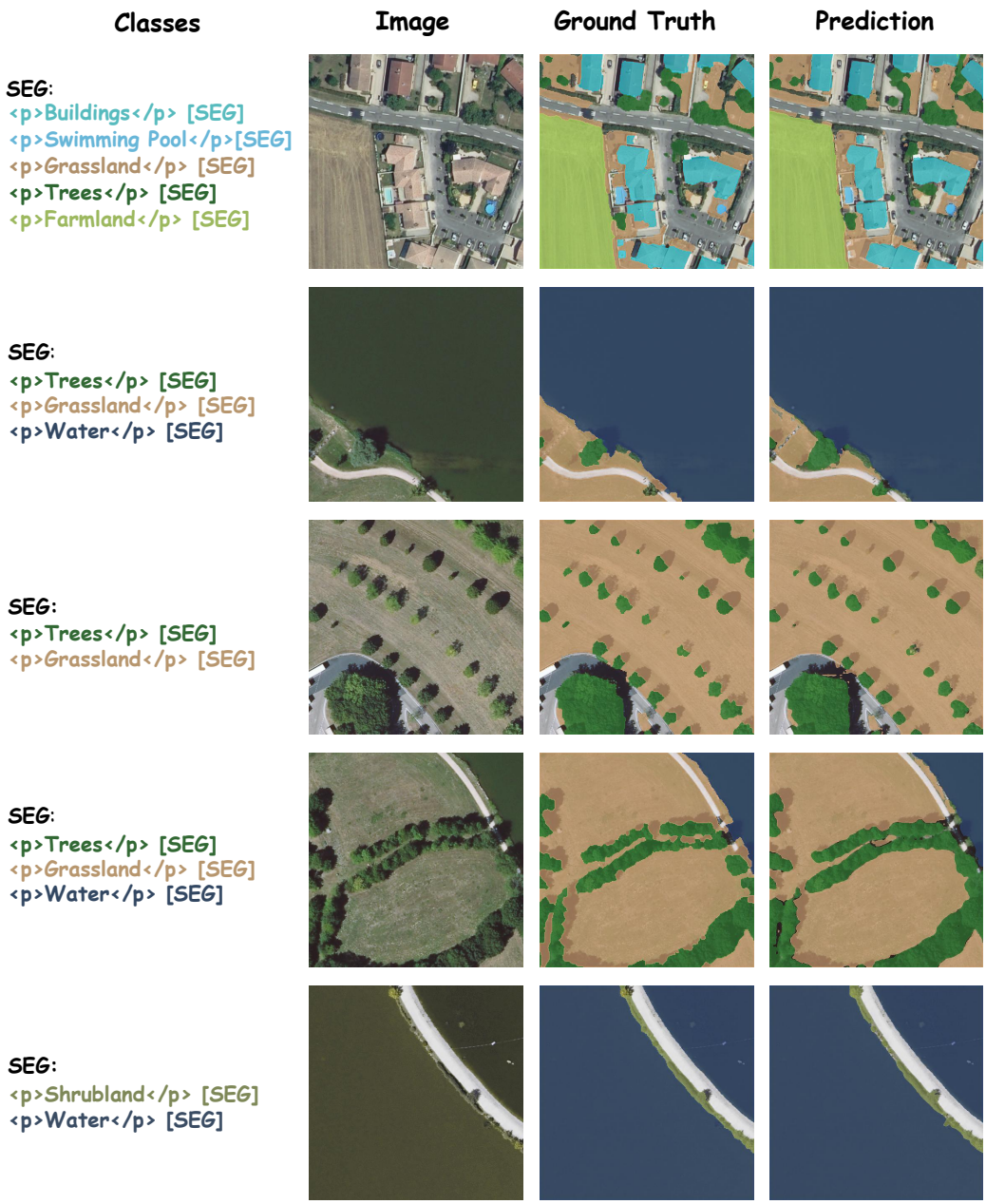


Figure 22: **Additional qualitative prediction results of OmniEarthSeg on multispectral imagery.** The examples show semantic segmentation of diverse land-cover regions, including categories with similar spectral characteristics.



Figure 23: **Additional qualitative prediction results of OmniEarthSeg on multispectral imagery.** The examples show semantic segmentation of diverse land-cover regions, including categories with similar spectral characteristics.

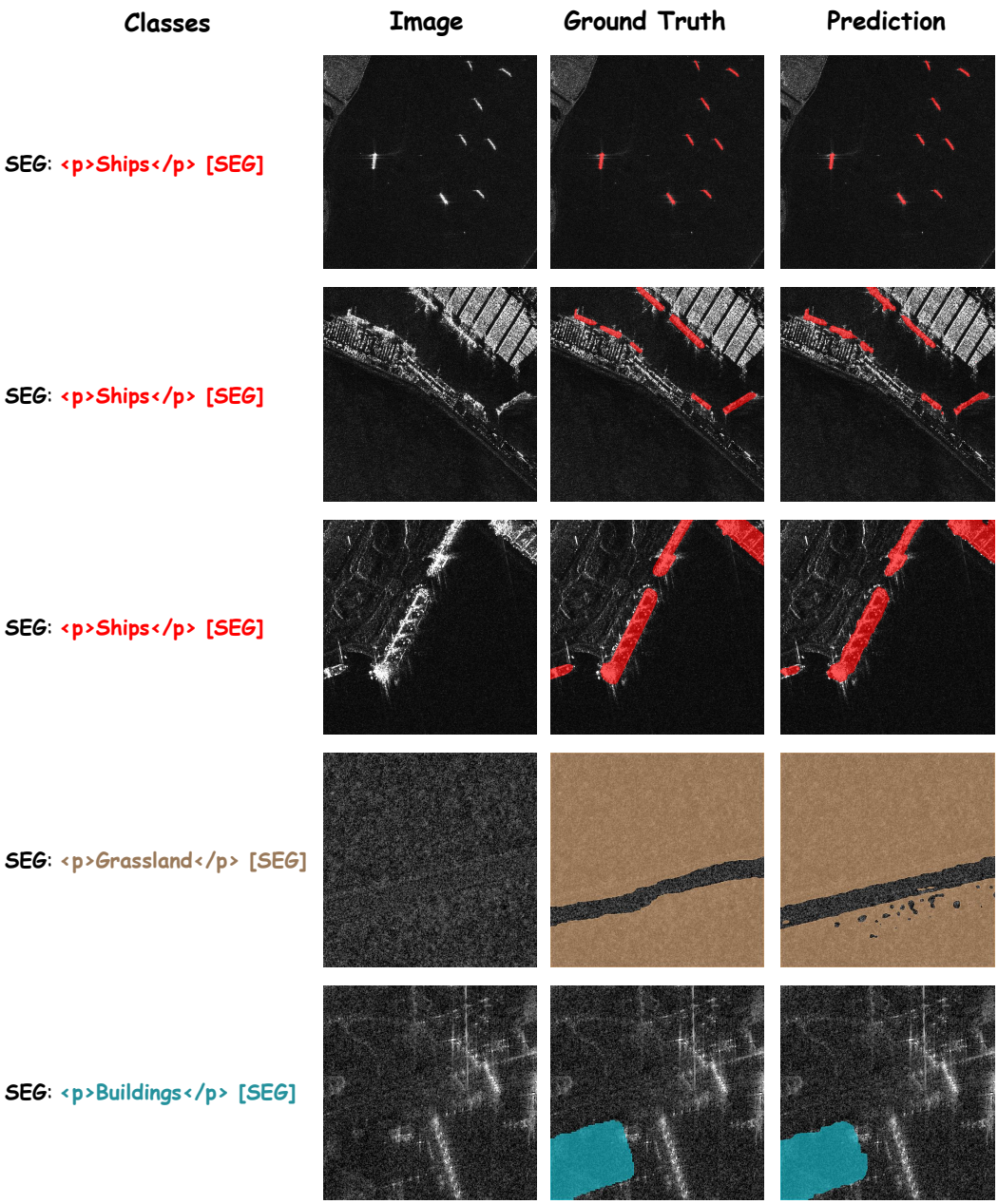


Figure 24: **Additional qualitative prediction results of OmniEarthSeg on SAR imagery.** The examples cover both object and stuff segmentation across diverse scenes, including challenging cases with complex scattering patterns and speckle noise.

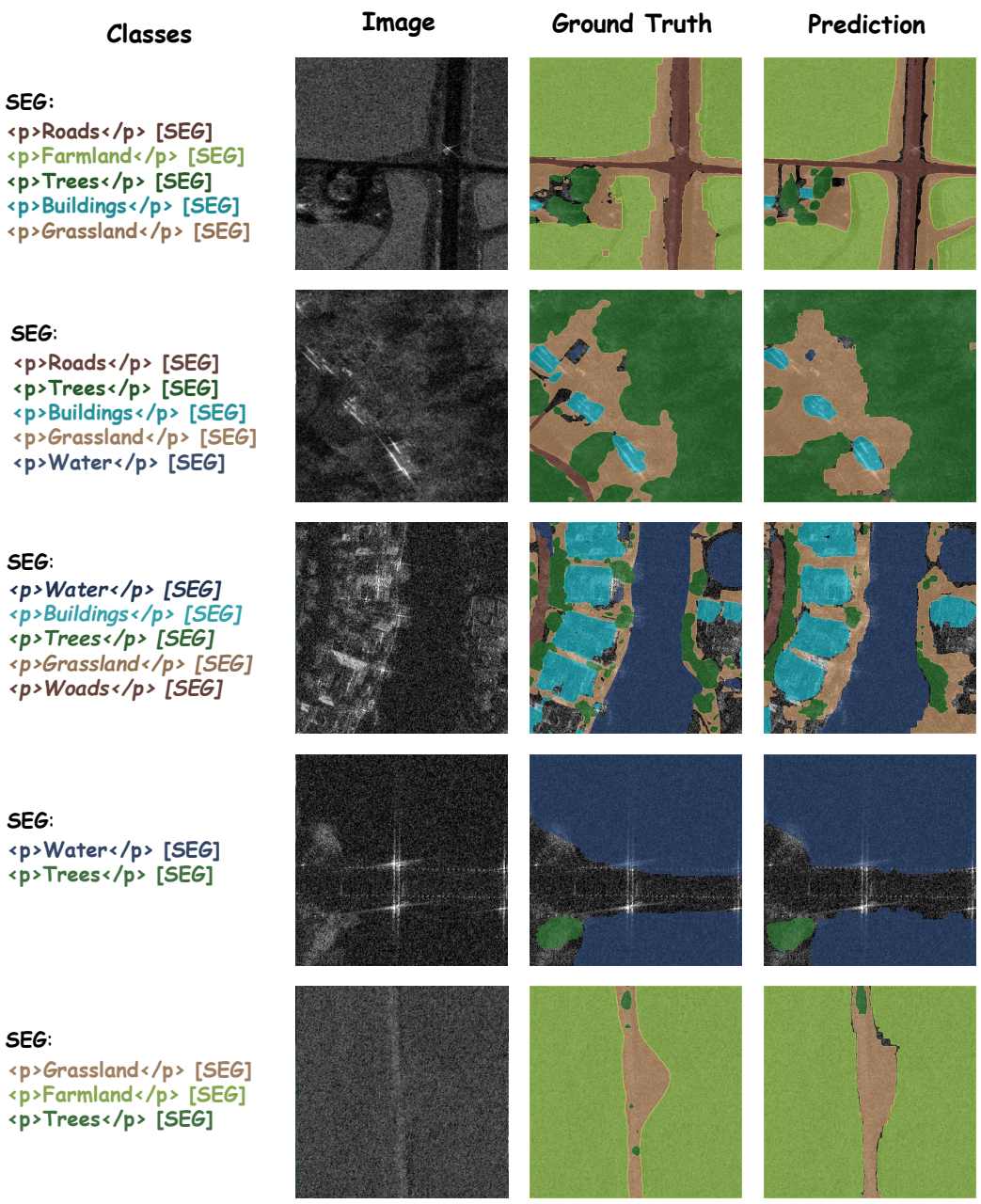


Figure 25: **Additional qualitative prediction results of OmniEarthSeg on SAR imagery.** The examples cover both object and stuff segmentation across diverse scenes, including challenging cases with complex scattering patterns and speckle noise.

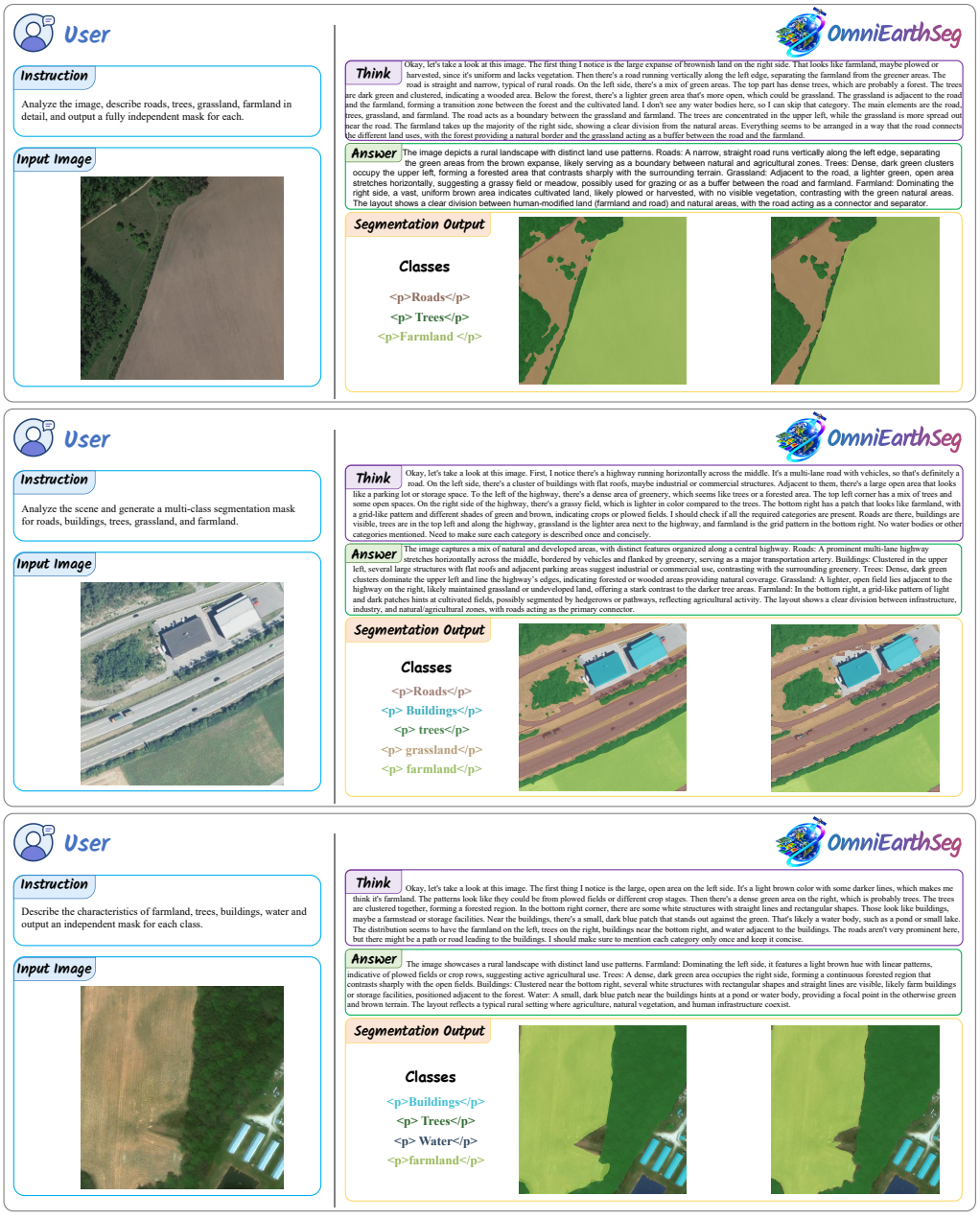


Figure 26: **Dialogue examples of OmniEarthSeg for semantic segmentation.** The dialogue demonstrates instruction-guided reasoning and semantic-level mask generation on RGB imagery.

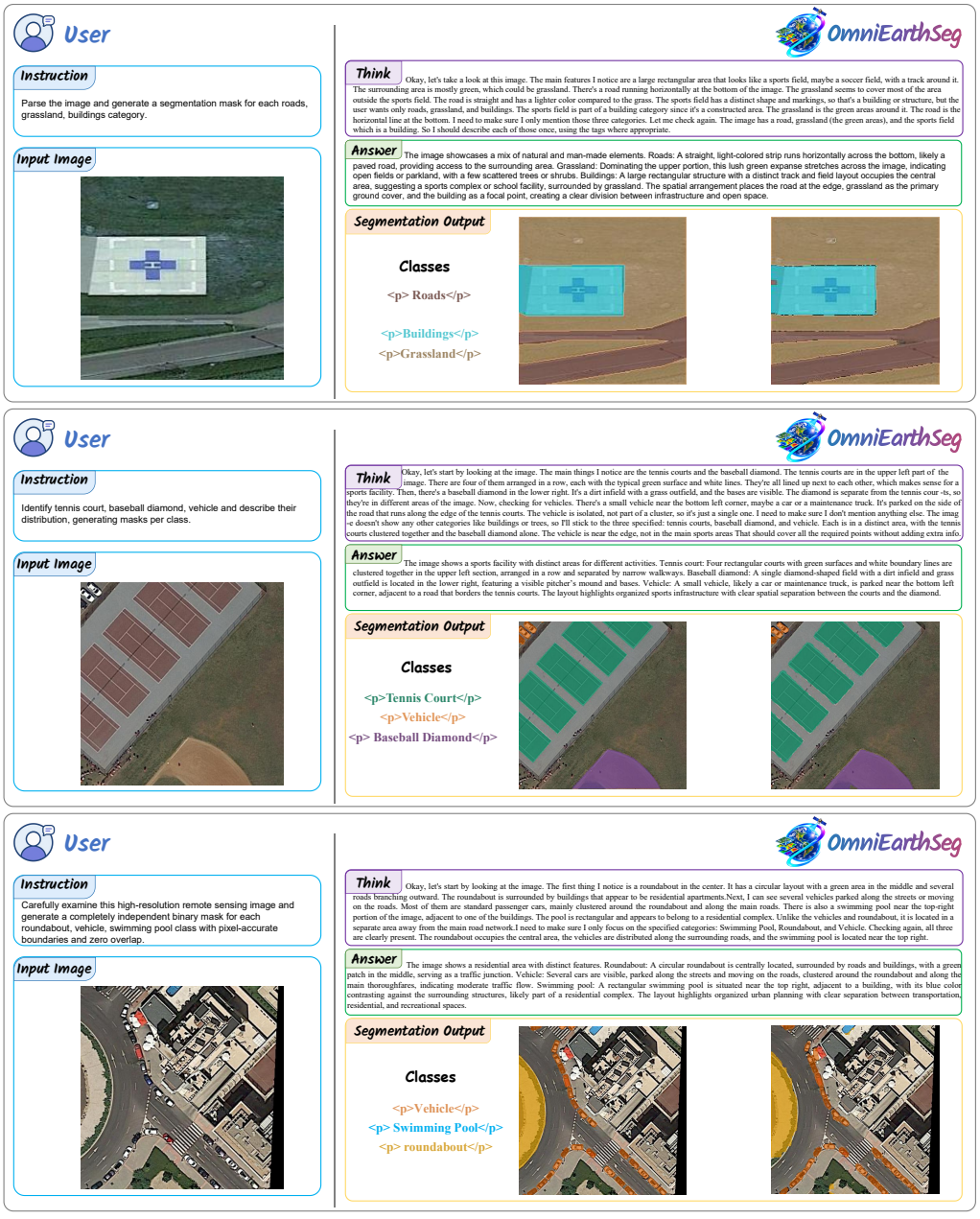


Figure 27: **Dialogue examples of OmniEarthSeg for instance segmentation.** The dialogue demonstrates instruction-guided reasoning and instance-level mask generation on RGB imagery.



Figure 28: **Dialogue examples of OmniEarthSeg on multispectral imagery.** The dialogue demonstrates instruction-guided reasoning and segmentation on multispectral observations.

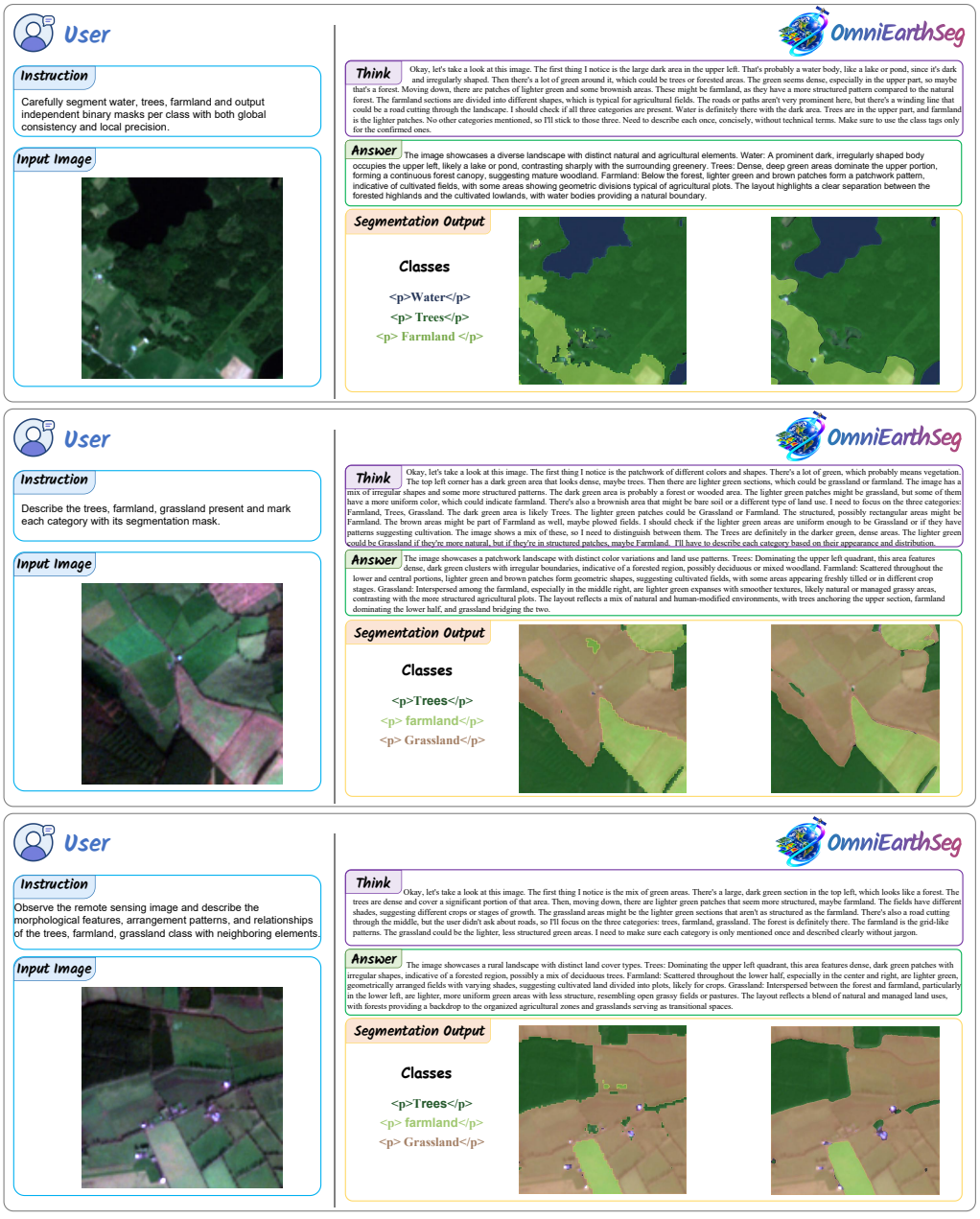


Figure 29: **Dialogue examples of OmniEarthSeg on multispectral imagery.** The dialogue demonstrates instruction-guided reasoning and segmentation on multispectral observations.

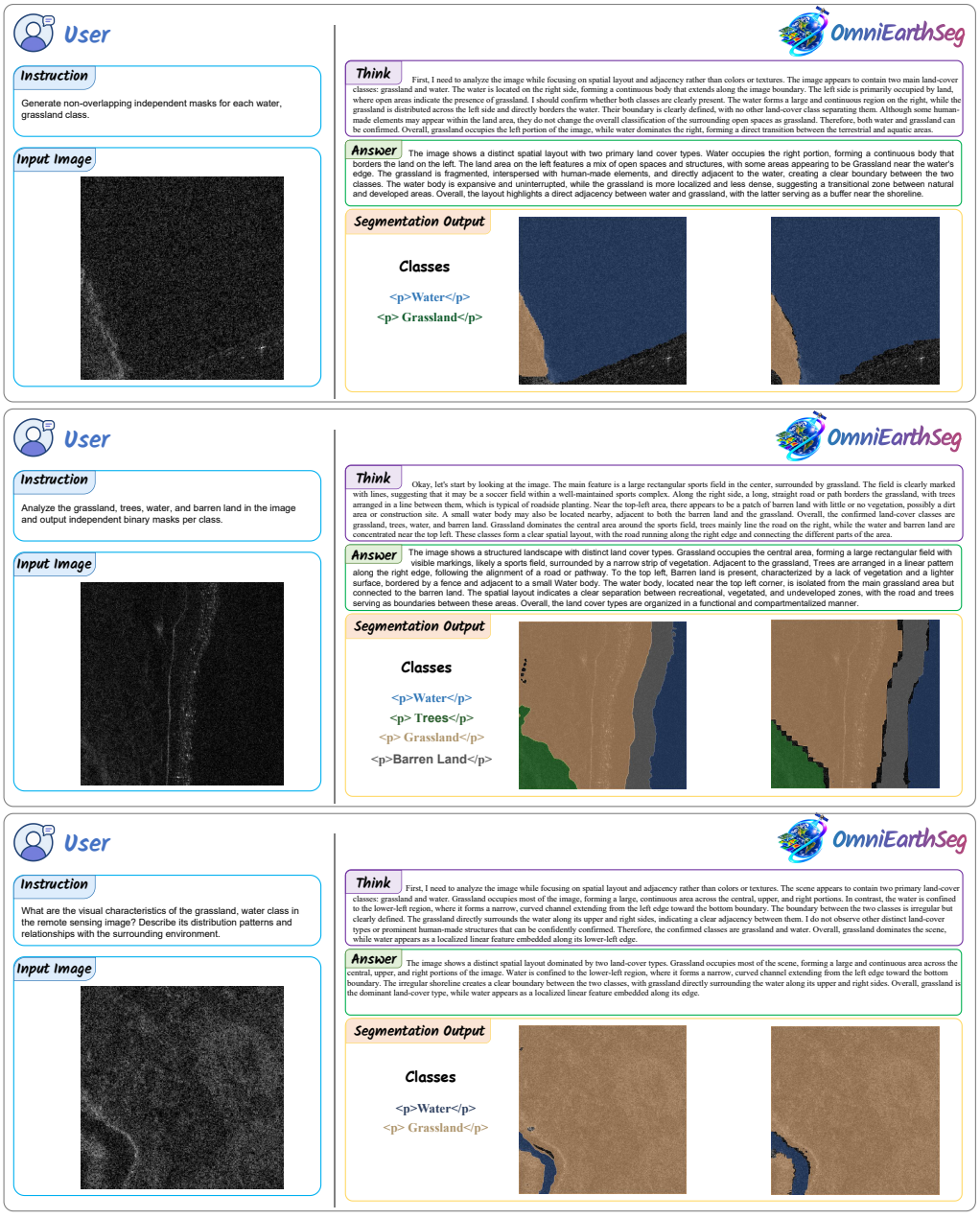


Figure 30: **Dialogue examples of OmniEarthSeg on SAR imagery.** The example illustrates the process from instruction interpretation and reasoning to pixel-level mask generation.

A.6 DATASHEET FOR OMNIGLOBAL-1.9M

A.6.1 MOTIVATION

1. *For what purpose was the dataset created?*

A1: OmniGlobal-1.9M was created to support the development of unified multimodal vision-language models for heterogeneous remote sensing imagery, with a particular focus on instruction-guided remote sensing understanding and segmentation. Existing remote sensing datasets are often limited to specific sensors, modalities, or task settings, making it difficult to develop models that generalize effectively across diverse observation conditions. To address this limitation, OmniGlobal-1.9M integrates large-scale RGB, multispectral, and SAR imagery acquired from multiple sensors and provides instruction-based supervision covering diverse scenes, target categories, and segmentation granularities. The dataset is designed to facilitate the development of models with strong cross-sensor and cross-modality generalization capabilities and to support research on unified remote sensing perception, instruction following, and fine-grained spatial understanding.

2. *Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?*

A2: The dataset was created by the authors of this paper.

3. *Who funded the creation of the dataset?*

A3: The dataset creation was funded by the affiliations of the authors involved in this work.

A.6.2 COMPOSITION

1. *What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances(e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

A1: OmniGlobal-1.9M is organized around instruction-response samples. Each sample primarily consists of a remote sensing image, an instruction-response pair, and the corresponding pixel-level segmentation mask. The imagery covers multiple sensor modalities, including RGB, multispectral (MS), and synthetic aperture radar (SAR). Depending on the task configuration, a sample may also include category labels, fine-grained target descriptions, and reasoning rationales. Overall, OmniGlobal-1.9M provides multiple complementary forms of supervision, including remote sensing imagery, natural-language instructions, textual responses, pixel-level segmentation annotations, and reasoning supervision.

2. *How many instances are there in total (of each type, if appropriate)?*

A2: OmniGlobal-1.9M contains 802,812 remote sensing images and 1,935,675 instruction-response samples.

3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).*

A3: OmniGlobal-1.9M is constructed by sampling and integrating multiple publicly available remote sensing datasets rather than attempting to exhaustively cover all possible remote sensing observations. The source datasets span diverse geographic regions, sensor platforms, spatial resolutions, scene types, and imaging modalities, including RGB, multispectral (MS), and synthetic aperture radar (SAR). They cover a broad range of representative platforms and sensors, such as DJI UAVs, Sentinel-2, WorldView-2/3, GF-2, NAIP aerial imagery, and IGN aerial imagery. During dataset construction, we preserve this diversity by sampling from heterogeneous data sources and scene categories. Therefore, although OmniGlobal-1.9M does not cover the full range of remote sensing imagery, it is designed to provide broad and representative coverage of sensor modalities, acquisition platforms, geographic environments, and target categories.

4. *What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)or features? In either case, please provide a description.*

A4: Each instance in OmniGlobal-1.9M consists of a remote sensing image or a cropped image patch, a corresponding natural-language instruction-response pair, and pixel-level segmentation annotations.

5. *Is there a label or target associated with each instance? If so, please provide a description.*

A5: Yes. Each instance is associated with one or more supervision targets. The primary target is a pixel-level segmentation mask corresponding to the object or region specified by the instruction. Depending on the task configuration, an instance may also include category labels, textual responses, fine-grained target descriptions, and reasoning rationales as additional supervision.

6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.*

A6: Some auxiliary metadata, such as acquisition time, precise geolocation, and sensor-specific information, may be unavailable for a subset of instances because these details were not provided by the original source datasets.

7. *Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are drawn.*

A7: No. Relationships between individual instances are not explicitly annotated in OmniGlobal-1.9M. Each instance is treated independently during dataset construction.

8. *Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

A8: Yes. OmniGlobal-1.9M includes training and evaluation splits. We do not introduce any additional data partitioning; instead, we directly follow the official splits provided by the original source datasets.

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

A9: No.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

A10: The released version of OmniGlobal-1.9M is self-contained and includes the processed remote sensing imagery, annotations, and instruction-response pairs required for its use. The dataset is constructed from publicly available remote sensing datasets obtained from their official sources. However, the original source datasets remain subject to their respective licenses and terms of use, and their availability may change over time.

11. *Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.*

A11: No.

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

A12: No.

A.6.3 COLLECTION PROCESS

1. *How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

A1: OmniGlobal-1.9M is built from publicly available RGB, multispectral (MS), and SAR datasets. Pixel-level masks and category labels are inherited from the original annotations, while instructions, responses, descriptions, and reasoning rationales are constructed from the imagery and annotations and checked for consistency.

2. *What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?*

A2: The data in OmniGlobal-1.9M were collected from the official releases of the source datasets and standardized using Python scripts to ensure consistency and reliability.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?*

A3: We did not apply any additional sampling strategy. OmniGlobal-1.9M was constructed from the official releases of the source datasets, following their original data availability and official splits.

4. *Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)*

A4: The data was collected and verified by the authors.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

A5: The data collection process took approximately 30 days. Mask filtering required about 15 days, while image cropping and category harmonization took approximately two weeks. The subsequent data annotation process lasted approximately four months.

A.6.4 PREPROCESSING/CLEANING/LABELING

1. *Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing 5 of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.*

A1: The remote sensing images were mainly processed into 512×512 patches, with some images retained at 1024×1024 resolution.

2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*

A2: No.

3. *Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.*

A3: We use Python for clipping.

A.6.5 USES

1. *Has the dataset been used for any tasks already? If so, please provide a description.*

A1: No.

2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*

A2: N/A.

3. *What (other) tasks could the dataset be used for?*

A3: OmniGlobal-1.9M can be used for multimodal model pre-training and instruction tuning. It can also support various remote sensing tasks, including image restoration, super-resolution, semantic segmentation, instance segmentation, object detection, scene classification, visual question answering, visual grounding, image-text retrieval, and open-vocabulary recognition and segmentation.

4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?*

A4: No.

5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*

A5: No.

A.6.6 DISTRIBUTION

1. *Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?*

A1: Yes. The dataset will be publicly available.

2. *How will the dataset be distributed (e.g., tarball on website, API, GitHub)?*

A2: The dataset will be publicly available on Hugging Face.

3. *When will the dataset be distributed?*

A3: The dataset will be distributed after the review process.

4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

A4: It will be distributed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 License.

5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

A5: No.

6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or*

otherwise reproduce, any supporting documentation.

A6: No.

A.6.7 MAINTENANCE

1. *Who will be supporting/hosting/maintaining the dataset?*

A1: The authors.

2. *How can the owner/curator/manager of the dataset be contacted (e.g., email address)?*

A2: They can be contacted via email available on the project website.

3. *Is there an erratum? If so, please provide a link or other access point.*

A3: No.

4. *Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?*

A4: No.

5. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.*

A5: N/A.

6. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.*

A6: N/A.