Unified Sequence Modeling for Remote Sensing: A Parameter-Efficient Foundation Model via Prompt-Driven Granularity Alignment

Yang Liu[1,2], **Weixing Luo**[1], **Huaizhou Qi**[1], **Suisui Jia**[1] and **Yongjing Guo**[1]
[1]The School of Telecommunications Engineering, Xidian University, Shaanxi 710071, China
[2]The PRC Ministry of Education Engineering Research Center of Intelligent Technology for Healthcare, Wuxi, Jiangsu 214122, China
yangl@xidian.edu.cn, {23011211117, 24011211181, 24011211204, 24011211154}@stu.xidian.edu.cn

Abstract

Current remote sensing (RS) perception systems suffer from task heterogeneity, necessitating distinct architectures for classification, localization, and reasoning. While vision–language models (VLMs) offer a route toward unification, their computational cost can hinder deployment. In this work, we propose RS-Florence, a compact unified model that addresses these tasks through a Prompt-Driven Sequence-to-Sequence framework. Unlike traditional approaches that segregate semantic understanding and geometric localization, RS-Florence maps images and task-specific prompts into a unified sequence of natural language and discrete geometric tokens. This formulation helps bridge high-level semantics and low-level pixel perception. Experiments across 4 task families and 8 benchmarks show that our 0.23B model remains competitive with task-specific specialists. These results also suggest that multitask joint training improves performance on several benchmarks over single-task fine-tuning, indicating that a shared prompt-driven interface can serve both language and geometry tasks.

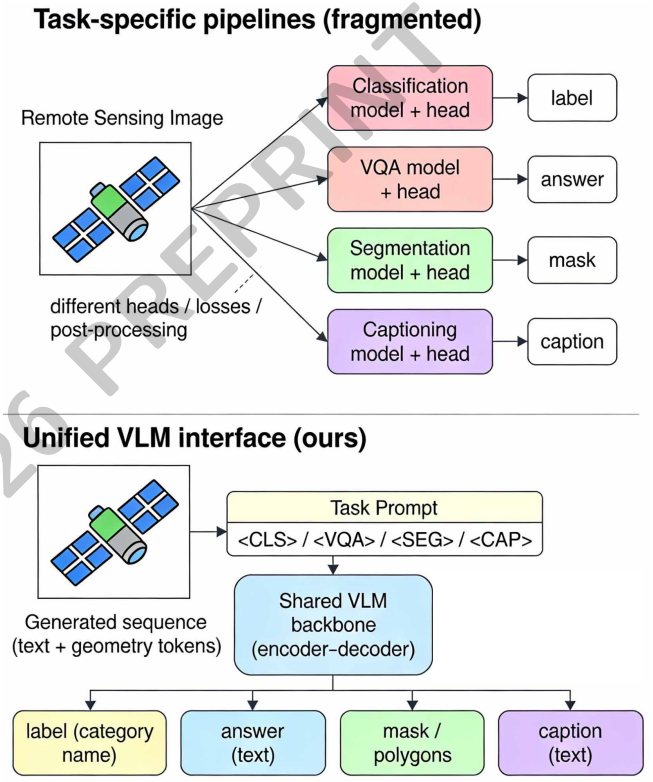

Figure 1: Fragmented pipelines decouple semantic reasoning from geometric localization, while our unified model uses a single prompted backbone with generate-then-parse decoding.

1 Introduction

Comprehensive interpretation of Earth observation imagery demands a duality of capabilities. A practical system must provide high-level semantic understanding and low-level geometric localization with fine spatial precision. However, the remote sensing community has long developed these capabilities along separate trajectories. High-level recognition [Yu *et al.*, 2020; Zhang *et al.*, 2022; Zheng *et al.*, 2023] and reasoning [Lobry *et al.*, 2020; Chappuis *et al.*, 2022; Wang *et al.*, 2024; Wang and Ghamisi, 2024] are typically addressed by task-specific classification [Guo *et al.*, 2022; Duan *et al.*, 2025; Liu *et al.*, 2025] and multimodal understanding models, while dense localization relies on dedicated segmentation backbones and losses [Wang *et al.*, 2022a; Zhang *et al.*, 2024; Sun *et al.*, 2024]. This task heterogeneity structurally decouples semantic understanding and geometric precision, leading to fragmented pipelines with task-dependent heads, objectives, and post-processing rules.

Vision–language models have shown strong generalization through language supervision and instruction-style prompting [Radford *et al.*, 2021; Li *et al.*, 2023; Dai *et al.*, 2023; Liu *et al.*, 2023a; Zhu *et al.*, 2023]. In remote sensing, large-scale VLMs improve multimodal understanding and offer a promising route toward unified perception [Kuckreja *et al.*, 2024; Hu *et al.*, 2025; Zhan *et al.*, 2025]. Yet current progress often follows scaling-law driven expansion to billions of parameters, while localization is commonly treated as continuous regression that mismatches discrete token generation in language modeling. These trends motivate a more efficient, sequence-centric formulation for studying semantic–

geometric synergy in unified remote sensing perception. Figure 1 contrasts fragmented pipelines with a unified interface.

This paper studies a unified formulation for remote sensing perception and reasoning under a single generate-then-parse workflow. We focus on three aspects. First, we examine whether heterogeneous objectives can be reduced to one unified decoding interface without task-specific heads. Second, we evaluate whether a compact model with 0.23B parameters can remain competitive on language-centric benchmarks while still supporting dense geometric prediction. Third, we study whether multitask joint training provides positive transfer across task families under matched budgets. This matters in remote sensing because classification, captioning, VQA, and localization are often built and maintained as separate pipelines, which increases implementation overhead. A single prompt interface can reduce that fragmentation by sharing one backbone and one decoding stack across multiple outputs. More broadly, the goal is not to replace every dedicated specialist, but to provide a compact shared representation in which language and geometry can be learned together and compared under the same interface.

To address these aspects, we propose **RS-Florence**, a prompt-driven framework that reformulates heterogeneous tasks into a single sequence-to-sequence language modeling problem. The key idea is to introduce discrete geometric tokens that quantize spatial coordinates into a location vocabulary. The model generates a single interleaved sequence in which semantic tokens and geometric tokens share the same decoding space, and task outputs are obtained by deterministic parsing. We instantiate RS-Florence with a compact vision–language encoder–decoder backbone that combines a hierarchical vision tower based on DaViT with a BART-style encoder–decoder [Ding *et al.*, 2022; Lewis *et al.*, 2020].

We emphasize compact modeling and empirical cross-task effects. Our evaluation reports task metrics together with parameter counts, enabling transparent accuracy–efficiency comparisons against lightweight task-specific models and large remote sensing VLM baselines [Yu *et al.*, 2020; Wang *et al.*, 2022a; Kuckreja *et al.*, 2024; Hu *et al.*, 2025; Zhan *et al.*, 2025]. Despite its compact size, RS-Florence achieves competitive performance across multiple task families. Our experiments suggest a positive effect from multitask training. Joint training improves over single-task fine-tuning under matched budgets, indicating that semantic objectives can provide useful signals for geometric localization.

Our main contributions are:

- We propose **RS-Florence**, which reformulates heterogeneous remote sensing tasks into a single prompt-driven sequence-to-sequence language modeling problem via discrete geometric tokens, bridging semantic reasoning and geometric localization within a shared decoding space.
- We evaluate a compact 0.23B model against task-specific systems and larger remote sensing VLMs, providing an accuracy–efficiency comparison for unified remote sensing modeling.
- We analyze multitask joint training in remote sensing, showing improvements over single-task baselines under matched budgets and indicating useful interaction between semantic and geometric objectives.

2 Related Work

2.1 From Task-Specific Specialists to Unified Generalists

Remote sensing vision has been dominated by task-specific specialist models that achieve strong performance within narrow problem definitions. Compact CNN and Transformer architectures remain highly competitive for scene classification on established benchmarks [Xia *et al.*, 2016; Cheng *et al.*, 2017; Yang and Newsam, 2010], and modern segmentation backbones deliver accurate dense prediction under fully supervised protocols [Wang *et al.*, 2021; Wang *et al.*, 2022a; Zhang *et al.*, 2024; Sun *et al.*, 2024]. These specialist designs have advanced efficiency and accuracy for their respective tasks, and they form the backbone of many operational RS pipelines [Yu *et al.*, 2020; Zhang *et al.*, 2022; Zheng *et al.*, 2023; Guo *et al.*, 2022; Duan *et al.*, 2025].

Despite their success, specialist models commonly suffer from architectural silos. Their interfaces are tightly coupled to task-specific heads, losses, and output representations, which hinders knowledge transfer between semantic reasoning and geometric localization. As a result, semantic understanding learned for recognition is not naturally reusable for dense delineation, and geometric precision optimized for segmentation does not translate into language-grounded interaction. This structural decoupling limits holistic understanding of remote sensing scenes, where a system is expected to jointly explain scene semantics and localize objects or regions with fine geometry.

In general computer vision, unified perception frameworks have begun to reduce these silos by exposing a shared interface across heterogeneous tasks [Wang *et al.*, 2022b]. However, remote sensing scenes introduce unique challenges such as extreme scale variation, dense instance packing, and strong orientation effects, which make direct transfer of unified CV systems non-trivial. Our work follows the unified perception direction while emphasizing a sequence-based unification tailored to remote sensing, so that semantic and geometric predictions can be produced within one coherent decoding space. Related RS efforts toward language-aligned dense prediction have explored open-vocabulary or training-free settings [Li *et al.*, 2025; Ye *et al.*, 2025], but they typically remain focused on specific regimes or rely on heavy pretrained components rather than a compact unified interface.

2.2 RS-VLMs: Scaling Versus Efficiency and Modality Alignment

Vision–language models (VLMs) have demonstrated that language supervision and instruction-style prompting can unify multiple vision capabilities behind a single interface [Radford *et al.*, 2021; Li *et al.*, 2023; Dai *et al.*, 2023; Liu *et al.*, 2023a; Zhu *et al.*, 2023]. Motivated by this progress, recent remote sensing vision–language models adapt general VLMs or LLMs to remote sensing imagery and show strong gains on captioning and VQA [Kuckreja *et al.*, 2024; Hu *et al.*, 2025;

Zhan *et al.*, 2025; Chappuis *et al.*, 2022; Wang *et al.*, 2024]. These systems largely follow a scaling-law trajectory, where broader capability is pursued by increasing model capacity and relying on large backbones to absorb domain shifts.

A key limitation of this trend is that parameter scaling alone does not resolve the modality mismatch between language generation and geometric localization. Many VLMs treat localization through continuous regression or add external detection components, which results in an implicit two-stream design where discrete token generation governs language while geometry is handled by a separate prediction mechanism. This decoupling weakens modality alignment and complicates the goal of a truly unified inference interface. In addition, remote sensing applications often face strict latency and compute constraints, which makes compact modeling practically important.

RS-Florencetargets these issues by prioritizing modality alignment and compact modeling. By introducing discrete geometric tokens, we align spatial signals with language tokens in a shared discrete decoding space, enabling unified generation and deterministic parsing for both semantic and geometric outputs. This design complements parameter-efficient adaptation techniques that reduce training cost [Wang and Ghamisi, 2024], while differing in that our primary objective is a unified sequence-centric interface rather than improving a single downstream task.

2.3 Vision as Sequence Modeling

Our formulation is rooted in the vision-as-sequence paradigm, which casts visual perception as conditional language modeling and unifies structured prediction through sequence generation [Chen *et al.*, 2022; Wang *et al.*, 2022b; Xiao *et al.*, 2024]. Pix2Seq demonstrates that object detection can be modeled as token prediction with structured decoding, while later unified systems extend sequence modeling to broader task families by generating text and structured outputs in a shared vocabulary [Chen *et al.*, 2022; Wang *et al.*, 2022b; Xiao *et al.*, 2024]. These advances provide a principled foundation for unifying heterogeneous tasks under one generate-then-parse workflow.

Adapting this paradigm to remote sensing is not straightforward. Remote sensing scenes exhibit extreme scale variation, dense small instances, and pronounced orientation effects, which amplify sensitivity to coordinate representation, sequence length, and parsing validity. A direct application of natural-image sequence modeling may therefore degrade geometric fidelity or lead to brittle decoding under RS distributions. RS-Florence addresses these challenges by coupling prompt-driven task conditioning with a coordinate quantization strategy, so that geometric localization can be expressed as discrete tokens and optimized jointly with semantic generation. This makes unified sequence modeling practical for remote sensing without reverting to task-specific heads.

RS-Florence extends a lightweight, sequence-based unified paradigm to remote sensing by adapting the Florence-2 sequence-modeling interface to this domain.

3 Method

3.1 Overview

As illustrated in Fig. 2, RS-Florence is a unified prompt-conditioned sequence modeling framework for remote sensing perception. It contains a DaViT visual tower and a BART-style encoder–decoder language tower [Ding *et al.*, 2022; Lewis *et al.*, 2020]. The visual tower extracts hierarchical features $V = E_{\text{img}}(x)$ with multiple spatial resolutions, which is essential under large scale variance. The language tower projects visual tokens into its hidden space, concatenates them with prompt tokens, and decodes an output sequence conditioned on the multimodal context.

RS-Florence follows unified general-vision models such as Florence-2 [Xiao *et al.*, 2024], but remote sensing introduces heterogeneous supervision and severe task imbalance. We address these issues by unifying the decoding space for language and geometry, and by adopting multitask optimization with temperature-controlled task sampling. Under the same interface, it supports captioning, instruction-style visual question answering, localization, and segmentation. We further instantiate two remote sensing tasks within this interface. Scene classification is implemented by generating category names with `<CLS>`, and instance segmentation is implemented by composing `<OVD>` with `<R2S>`.

3.2 Problem Formulation: Perception as Language Modeling

We formulate heterogeneous remote sensing perception tasks as conditional sequence generation. Given an input image x and a prompt p, RS-Florence produces an output sequence

$$y = \mathcal{F}_\theta(x, p), \tag{1}$$

where $\mathcal{F}_\theta$ denotes the decoded sequence produced from the model distribution under a fixed decoding strategy. The output sequence $y = \{y_1, \ldots, y_L\}$ is generated from a unified vocabulary $\mathcal{V}$ that includes natural language tokens and discrete geometry tokens. We model the posterior probability autoregressively:

$$P_\theta(y \mid x, p) = \prod_{i=1}^{|y|} P_\theta\left(y_i \mid y_{<i}, x, p\right). \tag{2}$$

We write the prompt as a pair

$$p = \langle t, s \rangle, \tag{3}$$

where t is a task token and s is an optional text condition. When a task does not require extra text input, s is empty. The final prediction is obtained by a task-dependent parser

$$\hat{o} = \mathcal{G}_t(y), \tag{4}$$

which maps the generated sequence to either a plain-text output or a structured output. This generate-then-parse interface follows Florence-2 and enables a single backbone to serve multiple tasks through prompts [Xiao *et al.*, 2024; Chen *et al.*, 2022].

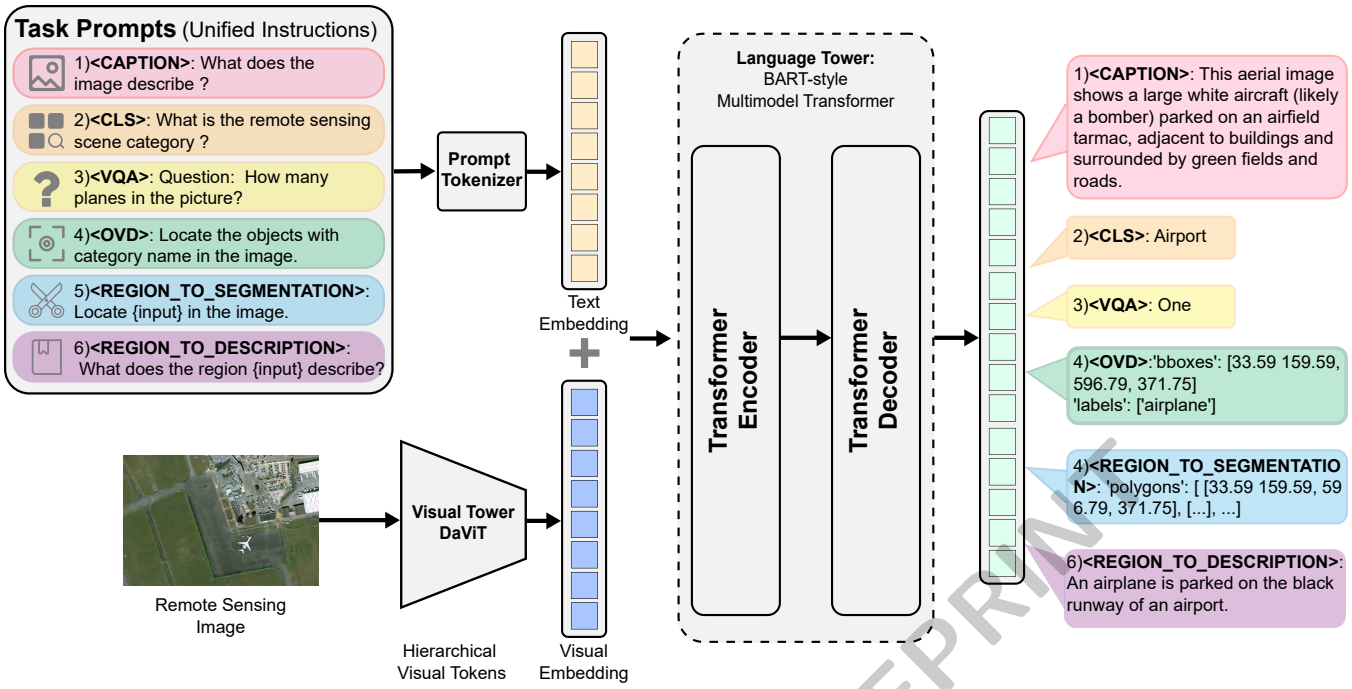

Figure 2: Architecture of RS-Florence. The model takes an image and a task prompt, generates a sequence with a shared backbone, and applies task-specific parsing to obtain plain-text or structured outputs.

3.3 Discrete Geometric Quantization

Unifying semantic generation and geometric prediction is hindered by the incompatibility between continuous coordinate regression and discrete token prediction. RS-Florence resolves this mismatch by quantizing spatial coordinates into a finite set of location tokens, so that localization is learned as token classification under the same autoregressive decoding space.

Following Florence-2 [Xiao *et al.*, 2024], we use 1000 discrete location tokens indexed by $k \in \{0, \ldots, 999\}$ as a balance between spatial granularity and vocabulary size. For a normalized coordinate $u \in [0, 1]$, the quantization operator is defined as

$$Q(u) = \lfloor \mathrm{clip}(u, 0, 1) \times 999 \rfloor , \tag{5}$$

where $\mathrm{clip}(\cdot, 0, 1)$ truncates a value into $[0, 1]$. Given an image with width W and height H, a pixel coordinate (x, y) is normalized as $u_x = x/W$ and $u_y = y/H$, and then mapped to a location index by applying Eq. 5. We denote the k-th location token by τ_k.

This quantization yields a unified serialization for different geometric structures. A bounding box $b = (x_1, y_1, x_2, y_2)$ is serialized as a four-token location sequence

$$\mathbf{z}(b) = \big[\tau_{Q(x_1/W)},\ \tau_{Q(y_1/H)},\ \tau_{Q(x_2/W)},\ \tau_{Q(y_2/H)}\big], \tag{6}$$

followed by its category token. For polygon-based segmentation, each vertex is represented by a pair of location tokens, and the polygon is serialized by concatenating all vertex pairs in order. During inference, the task-specific parser $\mathcal{G}_t$ deterministically converts the generated location-token stream into boxes or polygons in the image coordinate system.

3.4 Multi-Task Optimization

RS-Florence is optimized with a unified autoregressive objective over the output tokens. For each training triple (x, p, y), we minimize the negative log-likelihood

$$\mathcal{L}_{\mathrm{gen}}(x, p, y) = -\sum_{i=1}^{|y|} \log P_\theta \left(y_i \mid y_{<i}, x, p\right). \tag{7}$$

For multi-task training over a task set $\mathcal{T}$, the overall objective is

$$\min_\theta \sum_{t \in \mathcal{T}} \lambda_t \, \mathbb{E}_{(x,p,y)\sim\mathcal{D}_t} \left[\mathcal{L}_{\mathrm{gen}}(x, p, y)\right], \tag{8}$$

where λ_t controls the relative importance of each task. We adopt a temperature-controlled task sampling scheme based on dataset sizes:

$$P(t) = \frac{N_t^\alpha}{\sum_{j \in \mathcal{T}} N_j^\alpha}, \tag{9}$$

which reduces the dominance of large datasets when mixing heterogeneous tasks.

3.5 Task Instantiations

Remote-Sensing Scene Classification

Classification is instantiated as category-name generation with the task token `<CLS>`:

$$y_{\mathrm{cls}} = \mathcal{F}_\theta\big(x, \langle \texttt{<CLS>}, \varnothing \rangle\big). \tag{10}$$

The output y_{cls} contains a single class label, and the final prediction is $\hat{o}_{\mathrm{cls}} = \mathcal{G}_{\texttt{<CLS>}}(y_{\mathrm{cls}})$ through plain-text parsing. To keep the behavior consistent with prompt-based activation [Xiao *et al.*, 2024], we map `<CLS>` to the instruction: "What is the remote sensing scene category?"

Algorithm 1 Inference for classification and instance segmentation in RS-Florence.

Require: Image x; task type τ; optional query text q
if $\tau = \text{CLASSIFICATION}$ **then**
  $y \leftarrow \text{GENERATE}(\mathcal{F}_\theta, x, \langle \texttt{<CLS>}, \varnothing \rangle)$
  return $\text{TRIMTEXT}(y)$
else if $\tau = \text{INSTANCESEG}$ **then**
  $y_{\text{det}} \leftarrow \text{GENERATE}(\mathcal{F}_\theta, x, \langle \texttt{<OVD>}, q \rangle)$
  $B \leftarrow \text{PARSEBOXES}(y_{\text{det}})$
  for each $b \in B$ **do**
    $p_b \leftarrow \langle \texttt{<R2S>}, \text{ENCODEBOX}(b) \rangle$
    $y_b \leftarrow \text{GENERATE}(\mathcal{F}_\theta, x, p_b)$
    $m_b \leftarrow \text{PARSEMASK}(y_b)$
  end for
  return $\{(b, m_b)\}_{b \in B}$
end if

Instance Segmentation via Open-Vocabulary Detection and Region Segmentation

Instance segmentation is instantiated by composing two prompted skills under the same interface. We first predict instance boxes using open-vocabulary detection with <OVD>:

$$y_{\text{det}} = \mathcal{F}_\theta\big(x, \langle \texttt{<OVD>}, q \rangle\big), \tag{11}$$

$$B = \mathcal{G}_{\texttt{<OVD>}}(y_{\text{det}}), \tag{12}$$

where q is the query text and B is the parsed set of bounding boxes. For each $b \in B$, we then generate a region-conditioned mask using <R2S>:

$$p_b = \langle \texttt{<R2S>}, \text{ENCODEBOX}(b) \rangle, \tag{13}$$

$$y_b = \mathcal{F}_\theta(x, p_b), \tag{14}$$

$$m_b = \mathcal{G}_{\texttt{<R2S>}}(y_b). \tag{15}$$

The final instance set is obtained by aggregating $\{(b, m_b)\}$ over all predicted regions. This box-then-mask decomposition follows region-based instance segmentation designs and benefits from open-vocabulary detection for open-set generalization [He *et al.*, 2017; Li *et al.*, 2022; Minderer *et al.*, 2022; Liu *et al.*, 2023b].

Algorithm 1 summarizes inference for the two newly introduced tasks under the unified prompt-to-sequence-to-structure interface. Here, GENERATE is equivalent to calling $\mathcal{F}_\theta(x, \langle t, s \rangle)$ with the corresponding task token t and text condition s.

4 Experiments and Empirical Analysis

4.1 Experimental Setup

We evaluate **RS-Florence** as a unified remote-sensing model on four task families: visual question answering, image captioning, instance segmentation, and scene classification. All tasks are triggered by textual prompts and solved by sequence generation followed by deterministic parsing, so one decoding pipeline is shared across heterogeneous objectives.

Benchmarks and protocols. Our main evaluation suite covers seven benchmarks spanning semantic abstraction and geometric precision. For RS-VQA, we use RSVQA-LR and RSIVQA and report average accuracy in percentage, following common practice [Lobry *et al.*, 2020]. For captioning, we use RSICD and NWPU-Captions and report BLEU-4, METEOR, ROUGE-L, and CIDEr. For instance segmentation, we evaluate on iSAID and report mAP and Recall@100 under the GZSRI protocol [Zheng *et al.*, 2021; He *et al.*, 2023; Huang *et al.*, 2024]. For scene classification, we report Top-1 accuracy on AID, NWPU-RESISC45, and UC Merced [Xia *et al.*, 2016; Cheng *et al.*, 2017; Yang and Newsam, 2010]. We additionally include WHU-RS19 only for representation analysis in Fig. 3, and we exclude it from quantitative tables.

Implementation details. We initialize **RS-Florence** from Florence-2-base with 0.23B parameters and perform instruction tuning with Discrete Geometric Tokens. Unless otherwise stated, all tasks are trained jointly to evaluate multitask synergy under a unified interface. We train for 10 epochs with batch size 10 on 8× A100 GPUs, using AdamW with learning rate $1e{-}6$ and a constant schedule. We decode with beam size 3 and a maximum generation length of 1024 tokens. Structured outputs are validated and converted to predictions by deterministic parsers with strict checks, and ill-formed sequences are discarded. This standardization keeps training and inference consistent across task families and avoids task-specific heuristics.

4.2 Comparison with State of the Art

Tables 1–4 summarize the main results on captioning, RS-VQA, instance segmentation, and scene classification, together with parameter counts to reflect the accuracy–efficiency trade-off.

Performance against task-specific pipelines. Across task families, **RS-Florence** remains competitive with specialist systems optimized for a single output format. For captioning, it achieves CIDEr 2.850 on RSICD and 2.013 on NWPU-Captions with balanced BLEU-4, METEOR, and ROUGE-L in Table 1. For RS-VQA, it reaches 88.60 on RSVQA-LR and 83.20 on RSIVQA in Table 2, which is close to strong RS-VQA baselines under matched protocols. For iSAID, it obtains 10.65 mAP and 45.22 Recall@100 under GZSRI in Table 3, improving mAP over FC-CLIP while keeping recall in a comparable range. For scene classification, it achieves 96.92 on AID, 94.20 on NWPU-RESISC45, and 98.52 on UC Merced in Table 4, showing that the same generate-then-parse interface supports recognition without a dedicated classifier head.

Efficiency against large RS foundation models. RS-Florence uses 0.23B parameters, which corresponds to about 1.77% of a 13B model and about 3.29% of a 7B model. Despite the scale gap, it sustains competitive performance on language-conditioned tasks and supports dense prediction through the same sequence representation. These comparisons suggest that discrete geometric vocabularies and prompt-driven decoding can provide strong domain alignment without relying on parameter scaling.

	RSICD				NWPU-Captions				
Model	BLEU-4	METEOR	ROUGE-L	CIDEr	BLEU-4	METEOR	ROUGE-L	CIDEr	Params
MLCA-Net	0.461	0.351	0.646	2.356	0.478	0.337	0.601	1.164	25M
PCSFTr	0.544	0.391	0.702	3.012	0.690	0.443	0.780	2.036	55M
BITA	0.504	0.420	0.717	3.045	0.676	0.453	0.785	1.970	2.7B
RS-CapRet	0.455	0.376	0.649	2.605	0.650	0.439	0.775	1.919	∼7B
RSGPT	0.368	0.301	0.533	1.029	–	–	–	–	∼13B
SkyEyeGPT	0.600	0.354	0.626	0.837	–	–	–	–	∼7B
RS-Florence (Ours)	0.555	0.412	0.720	2.850	0.640	0.425	0.792	2.013	0.23B

Table 1: Captioning results on RSICD and NWPU-Captions. Higher is better for BLEU-4, METEOR, ROUGE-L, and CIDEr. "–" means not reported.

Model	RSVQA-LR	RSIVQA	Params
RSVQA (baseline)	81.49	–	85.69M
EasyToHard	84.76	–	148.83M
SHRNet	87.34	80.04	∼95M
RSAdapter	87.78	84.34	120M
RS-LLaVA	88.13	–	∼7B
Co-LLaVA	81.06	–	∼7B
RS-Florence (Ours)	88.60	83.20	0.23B

Table 2: RS-VQA results on RSVQA-LR and RSIVQA. Higher is better for average accuracy in percent. "–" means not reported.

Model	mAP	Recall@100	Params
ZSI	1.26	26.19	∼44M
D^2Zero	0.77	17.52	∼44M
FC-CLIP	8.83	47.10	∼259M
ZoRI	15.53	48.76	∼259M
RS-Florence (Ours)	10.65	45.22	0.23B

Table 3: iSAID instance segmentation results under the GZSRI protocol. Higher is better for mAP and Recall@100.

Model	AID	RESISC45	UC Merced	Params
BiMobileNet	96.87	94.08	99.03	∼8M
PCNet	96.76	94.59	99.25	∼26M
LDBST	96.84	93.56	99.52	∼29M
CSAT	95.44	93.06	97.86	∼26M
STMSF	97.51	94.95	99.58	∼30M
RS-Florence (Ours)	96.92	94.20	98.52	0.23B

Table 4: Scene classification results on AID, RESISC45, and UC Merced. Higher is better for Top-1 accuracy in percent.

Task	Single-task	Multitask	Δ	Metric
VQA	87.90	88.60	+0.70	Avg Acc ↑
Cap	2.561	2.850	+0.289	CIDEr ↑
Seg	9.75	10.65	+0.90	mAP ↑
Cls	95.40	96.92	+1.52	Top-1 Acc ↑

Table 5: Single-task fine-tuning versus multitask joint training under matched budgets. Δ is multitask minus single-task. Empty cells indicate not applicable. For segmentation, we report iSAID instance segmentation mAP under the GZSRI protocol.

4.3 Analysis of Multitask Synergy

We study whether heterogeneous objectives interfere or cooperate under a shared backbone through a controlled comparison between single-task fine-tuning and multitask joint training under matched parameter count and training budget. The generate-then-parse interface is unchanged, and only the task token and minimal text condition vary across tasks. Optimization follows Eq. 8 with temperature-controlled task sampling in Eq. 9, which balances gradients from imbalanced datasets and stabilizes cross-task updates. For segmentation, we use the same iSAID protocol as the main results and report GZSRI mAP, consistent with Table 3.

Table 5 shows consistent gains from joint training across all task families. VQA Avg Acc increases from 87.90 to 88.60, captioning CIDEr increases from 2.561 to 2.850, iSAID mAP increases from 9.75 to 10.65, and classification Top-1 accuracy increases from 95.40 to 96.92. This pattern indicates that, in our setting, multitask supervision provides complementary signals rather than pervasive negative transfer, and it improves both semantic discrimination and coordinate-sensitive decoding under the same backbone capacity.

We further interpret the cross-task effect through a semantic-to-geometric lens as an empirical association. Under joint training, language-conditioned objectives co-exist with improved geometry-centric decoding, evidenced by the iSAID mAP gain in Table 5. Fig. 4 localizes typical false positives and misses, and over-detections concentrate in crowded layouts and texture-confusing backgrounds, while omissions often occur on small instances and thin structures. Together, Table 5 and Fig. 4 support the view that multitask supervision regularizes shared representations and can improve geometric precision without changing the decoding interface.

4.4 Qualitative Visualization and Error Analysis

We provide qualitative visualization to characterize representation changes induced by joint training and the stability of token-based dense decoding. This qualitative evidence sup-

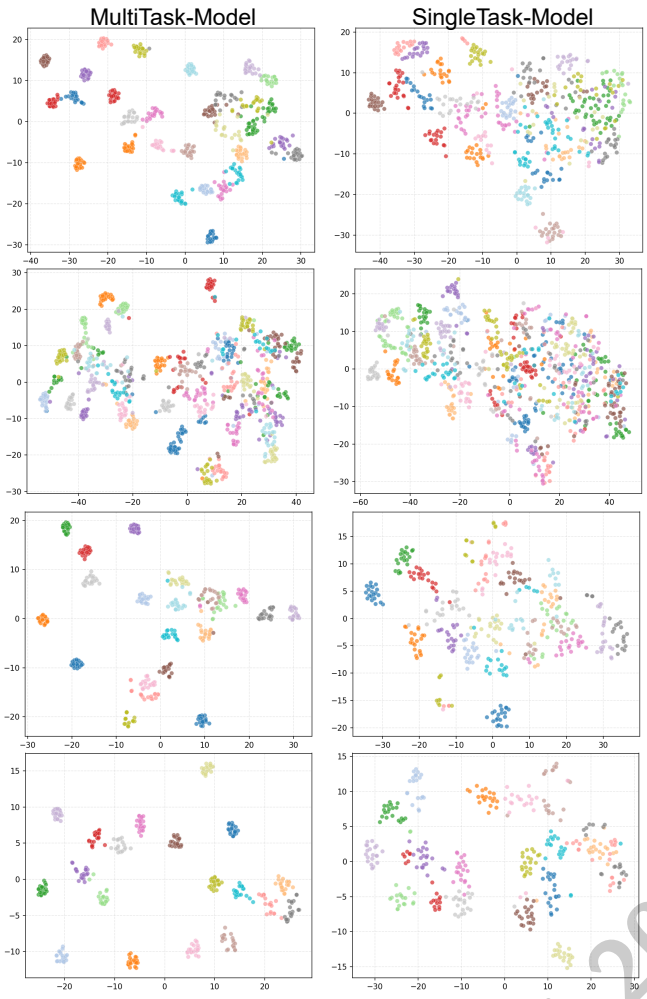

Figure 3: **t-SNE of scene embeddings.** Left shows multitask RS-Florence and right shows the single-task counterpart. Rows from top to bottom are AID, NWPU-RESISC45, UC Merced, and WHU-RS19. Points are colored by the ground-truth category.

ports that joint training improves cross-task consistency and reduces brittle decoding under domain-specific clutter in real-world imagery. These analyses complement the quantitative comparisons by exposing how the unified generate-then-parse interface behaves in challenging remote-sensing scenes.

t-SNE representation analysis. Fig. 3 visualizes scene-level embeddings from the Florence visual tower for the multitask model and its single-task counterpart under the same backbone setting. We extract one embedding per image and apply t-SNE on AID, NWPU-RESISC45, UC Merced, and WHU-RS19. Across datasets, the multitask model forms tighter clusters with less category mixing, suggesting improved feature robustness and sharper class boundaries. Notably, confusion is reduced among fine-grained categories with similar textures, which is consistent with the stronger recognition results observed under joint training.

Instance segmentation error decomposition and stability. Fig. 4 shows qualitative iSAID results with explicit error decomposition. Over-detection concentrates in crowded layouts and texture-similar backgrounds, while missed regions are more common for small instances, thin structures, and boundary-adjacent pixels. These patterns indicate that density and scale variation remain the primary sources of ambiguity. Meanwhile, the decoded outputs remain structured and spatially coherent in dense scenarios, and the deterministic mapping from discrete tokens to boundaries supports stable delineation and faithful error localization. The difference maps make it easier to distinguish boundary drift from true semantic confusion, which guides future improvements.

Figure 4: **Qualitative instance segmentation with error decomposition.** For each example, the top panel shows prediction overlays. The bottom panel shows a difference map where green is correct, red is over-detected, and blue is missed.

5 Conclusion

In this work, we introduce **RS-Florence**, a compact unified model that casts heterogeneous remote sensing tasks as prompt-driven sequence generation with Discrete Geometric Tokens. By sharing one decoding space for language and geometry, it supports classification, visual question answering, captioning, and instance segmentation without task-specific heads. Experiments across multiple benchmarks show the practicality of this formulation and suggest benefits from multitask joint training under matched capacity. Limitations remain on crowded scenes, small instances, prompt design, and coordinate discretization, leaving clear directions for future refinement within the same unified interface.

Acknowledgments

This research was funded by the National Natural Science Foundation of China under Grant 62376207; in part by the Xidian University Specially Funded Project for Interdisciplinary Exploration, China (No. TZJH2024045), in part by the Innovation Fund of Xidian University (No. YJSJ26022), and in part by the Fundamental Research Funds for the Central Universities, China.

References

[Chappuis *et al.*, 2022] Christel Chappuis, Valérie Zermatten, Sylvain Lobry, Bertrand Le Saux, and Devis Tuia. Prompt-rsvqa: Prompting visual context to a language model for remote sensing vqa. In *CVPRW*, 2022.

[Chen *et al.*, 2022] Ting Chen, Saurabh Saxena, Lala Li, David J. Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. In *ICLR*, 2022.

[Cheng *et al.*, 2017] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *arXiv preprint arXiv:1703.00121*, 2017.

[Dai *et al.*, 2023] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023.

[Ding *et al.*, 2022] Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. Davit: Dual attention vision transformers. In *ECCV*, 2022.

[Duan *et al.*, 2025] Yingtao Duan, Chao Song, Yifan Zhang, Puyu Cheng, and Shaohui Mei. Stmsf: Swin transformer with multi-scale fusion for remote sensing scene classification. *Remote Sensing*, 17(4):668, 2025.

[Guo *et al.*, 2022] Jingxia Guo, Nan Jia, and Jinniu Bai. Transformer based on channel-spatial attention for accurate classification of scenes in remote sensing image. *Scientific Reports*, 12:15473, 2022.

[He *et al.*, 2017] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017.

[He *et al.*, 2023] Shuting He, Henghui Ding, and Wei Jiang. Semantic-promoted debiasing and background disambiguation for zero-shot instance segmentation. In *CVPR*, 2023.

[Hu *et al.*, 2025] Yuan Hu, Jianlong Yuan, Congcong Wen, Xiaonan Lu, Yu Liu, and Xiang Li. Rsgpt: A remote sensing vision language model and benchmark. *ISPRS*, 224:272–286, 2025.

[Huang *et al.*, 2024] Shiqi Huang, Shuting He, and Bihan Wen. Zori: Towards discriminative zero-shot remote sensing instance segmentation, 2024. AAAI 2025.

[Kuckreja *et al.*, 2024] 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 *CVPR*, 2024.

[Lewis *et al.*, 2020] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, 2020.

[Li *et al.*, 2022] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, 2022.

[Li *et al.*, 2023] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.

[Li *et al.*, 2025] 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 *CVPR*, 2025.

[Liu *et al.*, 2023a] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

[Liu *et al.*, 2023b] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[Liu *et al.*, 2025] Yang Liu, Weixing Luo, Xinbo Gao, Jungong Han, and Ling Shao. Robust and discriminative visual-semantic alignment for zero-shot remote sensing image scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–14, 2025.

[Lobry *et al.*, 2020] Sylvain Lobry, Jesse Murray, Diego Marcos, and Devis Tuia. Rsvqa: Visual question answering for remote sensing data. *arXiv preprint arXiv:2003.07333*, 2020.

[Minderer *et al.*, 2022] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection with vision transformers. In *ECCV*, 2022.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[Sun *et al.*, 2024] Hongpeng Sun, Zhiqi Li, Jiawei Sun, Zhi Qiao, Zhiqiang Hu, Wei Li, Xin Wang, Kunlun Qi, Jiayi Ma, and Wenzhi Liao. Decouplenet: A lightweight backbone network with efficient feature decoupling for remote sensing image segmentation. *IEEE TGRS*, 62, 2024.

[Wang and Ghamisi, 2024] Yuduo Wang and Pedram Ghamisi. Rsadapter: Parameter-efficient adaptation of vision-language models for remote sensing visual question answering. *IEEE TGRS*, 2024.

[Wang *et al.*, 2021] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*, 2021.

[Wang *et al.*, 2022a] Libo Wang, Rui Li, Ce Zhang, Shenghui Fang, Chenxi Duan, Xiaoliang Meng, and Peter M. Atkinson. Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS*, 190:196–214, 2022.

[Wang *et al.*, 2022b] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*, 2022.

[Wang *et al.*, 2024] 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 *AAAI*, 2024.

[Xia *et al.*, 2016] Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, and Liangpei Zhang. Aid: A benchmark dataset for performance evaluation of aerial scene classification. *arXiv preprint arXiv:1608.05167*, 2016.

[Xiao *et al.*, 2024] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *CVPR*, 2024.

[Yang and Newsam, 2010] Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *ACM GIS*, 2010.

[Ye *et al.*, 2025] Chengyang Ye, Yunzhi Zhuge, and Pingping Zhang. Towards open-vocabulary remote sensing image semantic segmentation. In *AAAI*, 2025.

[Yu *et al.*, 2020] Donghang Yu, Qing Xu, Haitao Guo, Chuan Zhao, Yuzhun Lin, and Daoji Li. An efficient and lightweight convolutional neural network for remote sensing image scene classification. *Sensors*, 20(7):1999, 2020.

[Zhan *et al.*, 2025] Yang Zhan, Zhitong Xiong, and Yuan Yuan. Skyeyegpt: A large vision-language model for remote sensing. *ISPRS*, 221:64–77, 2025.

[Zhang *et al.*, 2022] Yue Zhang, Xiangtao Zheng, and Xiaoqiang Lu. Pcnet: Pairwise comparison network for remote-sensing scene classification. *IEEE Geoscience and Remote Sensing Letters*, 2022.

[Zhang *et al.*, 2024] Renhe Zhang, Qian Zhang, and Guixu Zhang. Lsrformer: Efficient transformer supply convolutional neural networks with global information for aerial image segmentation. *IEEE TGRS*, 62, 2024.

[Zheng *et al.*, 2021] Ye Zheng, Jiahong Wu, Yongqiang Qin, Faen Zhang, and Li Cui. Zero-shot instance segmentation. In *CVPR*, 2021.

[Zheng *et al.*, 2023] Feng Zheng, Shuting Lin, Wei Zhou, and Huilin Huang. A lightweight dual-branch swin transformer for remote sensing scene classification. *Remote Sensing*, 15(11):2865, 2023.

[Zhu *et al.*, 2023] 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.