SV-RoPE: Token Profiles of Spatial Variation for Remote Sensing Foundation Models

Anonymous authors
Paper under double-blind review

Abstract

Remote-sensing foundation models (RSFMs) acquire transferable spatial priors from large-scale Earth observation (EO) imagery, which are commonly inherited during downstream parameter-efficient fine-tuning (PEFT). Yet, for Transformer-based RSFMs, the spatial meaning of the same index-space token relation can vary substantially across observations, leaving inherited priors poorly matched to target-specific spatial variation. This raises a central question: *how can inherited priors be made responsive to target-specific spatial variation within a PEFT adaptation framework?* We therefore introduce a spatial variation-aware view of RSFM adaptation and propose SV-RoPE (Spatial-Variation-Aware Rotary Position Embedding), a PEFT method that associates each visual token with a spatial-variation profile (SV-Profile). Each profile summarizes how the token's feature agreement with neighboring tokens changes across increasing spatial extents. Leveraging the coupled displacement–frequency structure of RoPE, SV-RoPE encodes spatial variation directly into a unified rotary phase, jointly parameterized by Relation Coordinate Calibration (RCC) for initial ground-extent-calibrated displacement and Relation Frequency Calibration (RFC) for SV-Profile-conditioned frequency modulation. SV-RoPE is realized through a lightweight attention-parallel branch whose output is added to the attention of the frozen backbone. Across diverse RSFMs, sensors, resolutions, and downstream tasks, SV-RoPE consistently improves adaptation performance, demonstrating the value of explicitly modeling spatial variation in downstream EO adaptation. *Our code will be publicly available.*

1 Introduction

Remote-sensing foundation models (RSFMs) pretrained on large-scale Earth observation (EO) imagery acquire rich spatial priors that support transfer across diverse downstream tasks (Astruc et al., 2024; Do et al., 2025; Astruc et al., 2025). To adapt these pretrained models to downstream settings, parameter-efficient fine-tuning (PEFT) has become a common strategy (Scheibenreif et al., 2024; Dong et al., 2024; Thoreau et al., 2025; Luo et al., 2026). For widely used Transformer-based RSFMs, visual representations are organized on token grids, with spatial relations defined over grid-index offsets. These offsets alone do not characterize the spatial structure across the tokens: the same indexed relation can arise under different ground extents across pretraining and downstream data, leaving the inherited spatial prior poorly matched to the target imagery.

In EO imagery, an indexed relation can exhibit substantial *spatial variation* across observations. As the ground sampling distance (GSD) changes with varied sensing scale, the same indexed relation can span different ground distances (Reed et al., 2023; Revankar et al., 2025). Yet even at a fixed GSD, its surrounding spatial organization can vary markedly across different scene structures. As shown in Figure 1, this variation ranges from smooth homogeneous areas to densely structured regions and sharp structural transitions. However, common PEFT methods largely retain the pretrained spatial prior, including how token relations are modeled, limiting the adaptation to target-specific spatial variation. This raises a central research question for RSFM adaptation:

Research Question. How can inherited priors be made responsive to target-specific spatial variation within a PEFT adaptation framework?

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.

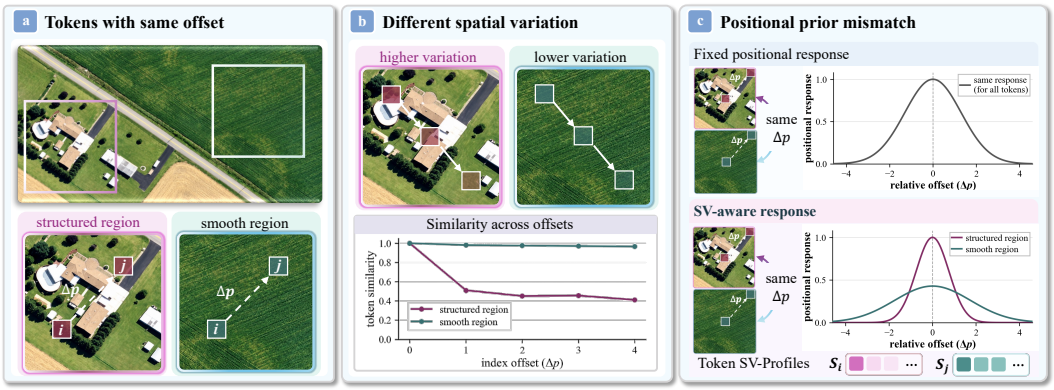


Figure 1: **(a)** The same index offset can correspond to markedly different local spatial structures, **(b)** with distinct spatial-variation profiles across neighborhood distances. **(c)** A shared positional response cannot explicitly account for such token-wise variation and may therefore mismatch the spatial context required by different regional sensitivity.

More broadly, studies beyond RSFM adaptation show that spatial modeling can benefit from cues specific to each input. Scale-aware methods condition positional modeling on explicit physical metadata such as GSD (Reed et al., 2023), while others directly use input features to modulate positional interactions (Wu et al., 2021; Zheng et al., 2024). However, spatial variation remains largely implicit in these formulations. In downstream adaptation, leaving such variation implicit can make positional sensitivity poorly matched to the target, *e.g.*, under-responsive in densely structured regions or overly responsive in homogeneous ones. We therefore introduce a spatial-variation profile (SV-Profile) for each visual token, constructed from neighborhood similarity across increasing spatial extents to summarize how center-to-neighbor feature agreement changes around the token.

Building on SV-Profile, we propose **SV-RoPE** (**S**patial-**V**ariation-Aware **Ro**tary **P**osition **E**mbedding), a PEFT method that adapts pretrained priors to target-specific spatial variation. SV-RoPE operates on RoPE (Su et al., 2024), whose phase couples relative displacement with a frequency basis, providing a natural adapting interface. Specifically, *Relation Coordinate Calibration (RCC)* adjusts the displacement according to the ground extent represented by each token step, while *Relation Frequency Calibration (RFC)* maps the SV-Profile to relation-specific frequency modulation. Their joint effect produces a spatial-variation-aware rotary phase, allowing the same indexed relation to exhibit different positional sensitivities across observations. SV-RoPE is injected through a lightweight attention-parallel adapter while keeping the pretrained backbone frozen. Experiments across multiple RSFMs, sensors, resolutions, and downstream tasks show consistent improvements, with ablations further validating the SV-Profile and its role in positional adaptation.

The key contributions of this work are summarized as follows:

- We introduce a spatial variation-aware view of downstream adaptation for RSFMs, treating spatial variation as an explicit signal for adapting pretrained spatial priors.
- We propose SV-RoPE, which characterizes each token with a spatial-variation profile derived from multi-extent neighborhood similarity to guide unified rotary phase calibration in PEFT.
- Extensive experiments across multiple RSFMs, sensors, resolutions, and tasks show consistent gains across EO settings, demonstrating the broad applicability of SV-RoPE.

2 Related Work

Scale-Aware Representation in RSFMs. EO observations exhibit substantial variation in spatial scale across imaging conditions (e.g., GSD). To tackle this diversity, foundational approaches have developed diverse pretraining objectives and architectures to learn scale-transferable representations. OmniSat (Astruc et al., 2024) aligns multimodal observations of different native resolutions over shared ground regions with contrastive learning, while SkySense (Guo et al., 2024) further explores multi-granularity learning. Resolution-adaptive architecture designs (Xiong et al., 2024b) also extend

this flexibility: AnySat (Astruc et al., 2025) combines joint-embedding predictive learning with resolution-adaptive spatial encoders, while AOM (Li et al., 2026) introduces multi-scale adaptive patch embedding through a multi-kernel convolution bank. A complementary direction incorporates explicit spatial metadata into modeling. PIPE (Li et al., 2025a) incorporates latitude, longitude and day-of-year information for spatio-temporal forecasting, while Copernicus-FM (Wang et al., 2025) and Scale-MAE (Reed et al., 2023) explicitly inject physical metadata and GSDs into the model with positional encodings during pretraining. Together, these advances highlight growing efforts to incorporate scalability into RSFM pretraining. Yet spatial variation extends beyond scale alone, and remains largely implicit during downstream adaptation.

Adaptive Positional Encoding. Positional encoding determines how spatial displacement influences attention interactions in Transformer (Vaswani et al., 2017). RoPE (Su et al., 2024) encodes relative positions through rotations of query and key features. A line of subsequent work extends RoPE itself by modifying frequency schedules, position indices, or axis-wise frequency allocation for extrapolation and multidimensional modeling (Peng et al., 2023; Ding et al., 2024; Zhu et al., 2023; Heo et al., 2024; Wei et al., 2025; Huang et al., 2026). Other formulations make the positional component itself responsive to input content. iRPE (Wu et al., 2021) models interactions between queries and relative position embeddings, while DAPE (Zheng et al., 2024) adapts positional encoding according to input context. Positional modeling can also be conditioned on explicit observation metadata, including camera parameters (Li et al., 2025b; Zhang et al., 2026a) and geospatial or temporal information in EO imagery (Wang et al., 2025). Nonetheless, spatial variation remains implicit in these formulations, limiting their ability to adapt positional sensitivity to heterogeneous structure. Our approach makes this variation explicit and uses it to modulate positional responses.

Parameter-Efficient Fine-Tuning in RSFMs. PEFT adapts large backbones by updating a minimal set of parameters via adapters (Houlsby et al., 2019), bias/norm tuning (Zaken et al., 2022; Zhao et al., 2023), or low-rank updates (LoRA) (Hu et al., 2022). In EO imagery, these mechanisms have been extended to heterogeneous data, such as Scaled Low-Rank (SLR) adaptation (Scheibenreif et al., 2024). More recent methods introduce specialized adaptation for spectral and spatial information: DEFLECT (Thoreau et al., 2025) disentangles spatial and spectral updates and SpectralX (Zhang et al., 2026b) decomposes spatial–spectral attributes and dynamically routes them through mixtures of adapters. These methods address EO heterogeneity, but leave spatial variation largely implicit, making variation-aware spatial adaptation under PEFT still underexplored.

3 Methodology

3.1 Problem Formulation

We formalize the downstream adaptation of a pretrained RSFM across heterogeneous EO observations. Let $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$ denote the spatial input presented to the encoder, corresponding to imagery acquired at GSD G, where G is the ground length represented by one image pixel. Here, H and W denote the spatial dimensions, and C denotes the channel dimension. For temporal inputs, $\mathbf{X}$ denotes the spatial representation obtained after backbone-specific temporal aggregation.

At stage l, the backbone produces an intermediate spatial token map $\mathbf{Z}^{(l)} \in \mathbb{R}^{N_x^{(l)}\times N_y^{(l)}\times D^{(l)}}$, where $N_x^{(l)}$ and $N_y^{(l)}$ are the token-grid dimensions and $D^{(l)}$ is the feature dimension. Each token is indexed by a 2D grid coordinate $\mathbf{u} = (u_x, u_y) \in \mathcal{N}^{(l)}$, with

$$\mathcal{N}^{(l)} = \{0, \dots, N_x^{(l)} - 1\} \times \{0, \dots, N_y^{(l)} - 1\}. \tag{1}$$

For a token pair $\mathbf{u}, \mathbf{v} \in \mathcal{N}^{(l)}$, we define the relative grid displacement as

$$\mathbf{\Delta}_{\mathbf{u},\mathbf{v}} = \mathbf{v} - \mathbf{u} = (\Delta^x_{\mathbf{u},\mathbf{v}}, \Delta^y_{\mathbf{u},\mathbf{v}}) = (v_x - u_x,\ v_y - u_y). \tag{2}$$

For brevity, we omit the stage index l when no ambiguity arises.

Motivation. The indexed displacement $\mathbf{\Delta}_{\mathbf{u},\mathbf{v}}$ specifies the relative relation in token-grid space, but does not characterize the spatial conditions under which that relation occurs. Let τ denote the effective spatial stride, in input pixels, between adjacent tokens at stage l. One token step therefore corresponds to a ground distance:

$$\ell = \tau G, \tag{3}$$

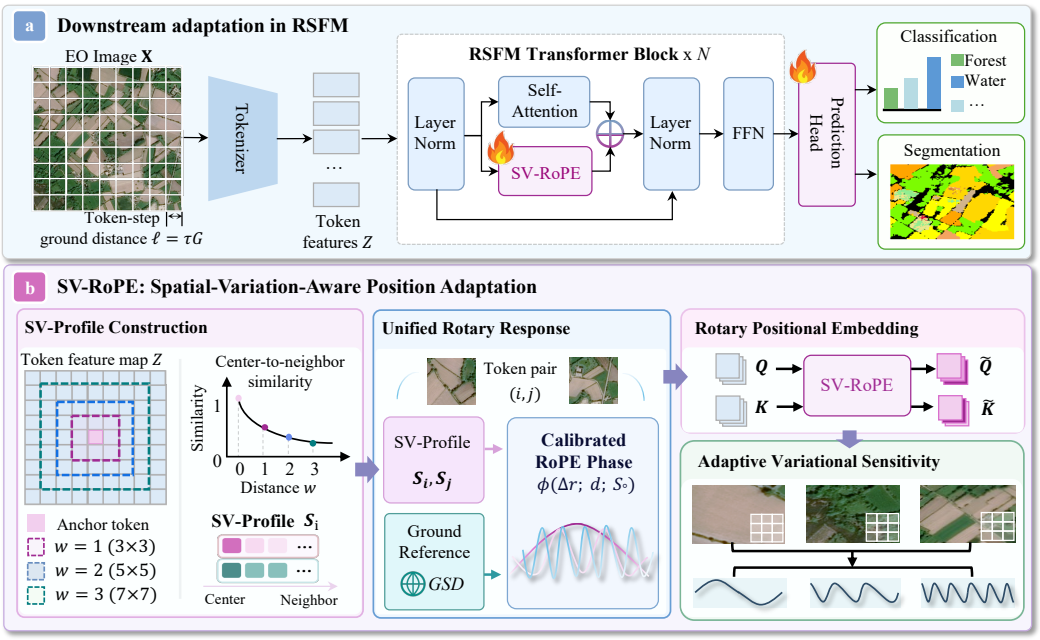


Figure 2: Overview of SV-RoPE. SV-RoPE augments the frozen RSFM with a lightweight attention-parallel branch. Token-wise SV-Profiles characterize spatial variation, while RCC and RFC calibrate displacement and frequency within a unified rotary phase. The calibrated rotations produce an additive attention update for downstream adaptation.

and the corresponding ground displacement is $\mathbf{d}_{\mathbf{u},\mathbf{v}} = \ell \boldsymbol{\Delta}_{\mathbf{u},\mathbf{v}}$.

Across observations, changes in ℓ alter the ground extent represented by the same indexed displacement. This scale dependence captures only part of the variation associated with an indexed relation: even at a fixed ℓ, the same relation may occur in homogeneous areas, densely structured regions, or sharp structural transitions. Hence, identical index-space relations can correspond to substantially different spatial conditions in the target imagery. We refer to this variability as *spatial variation* and seek to make it explicit during downstream adaptation.

Therefore, we construct a spatial-variation profile (SV-Profile) $\mathbf{s}_\mathbf{u}$ for each token $\mathbf{u}$, summarizing how its feature affinity with surrounding tokens changes across spatial extents. Together with the token-step ground extent ℓ, the profile $\mathbf{s}_\mathbf{u}$ provides the target-specific context needed to reinterpret an indexed displacement $\boldsymbol{\Delta}_{\mathbf{u},\mathbf{v}}$. SV-RoPE incorporates this information into a unified rotary phase for downstream adaptation.

3.2 Rotary Positional Prior

We next revisit Rotary Position Embedding (RoPE) (Su et al., 2024), which encodes relative positions through rotations of query and key features. For spatial axis $p \in \{x, y\}$ and frequency index m, the relative rotary phase between tokens $\mathbf{u}$ and $\mathbf{v}$ is

$$\phi^p_{\mathbf{u},\mathbf{v},m} = \theta_m \Delta^p_{\mathbf{u},\mathbf{v}}, \qquad \theta_m = \text{base}^{-4m/d_r}, \tag{4}$$

where d_r is the number of rotary channels, $m = 0, \ldots, M-1$, and $M = d_r/4$.

Each phase rotates a two-dimensional channel pair through $\mathbf{r}(\phi) = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}$. The full 2D relative rotation is obtained by applying these rotations independently across spatial axes and frequency bands (Heo et al., 2024):

$$\mathbf{R}_{\boldsymbol{\Delta}_{\mathbf{u},\mathbf{v}}} = \text{blockdiag}\left(\left\{\mathbf{r}(\phi^x_{\mathbf{u},\mathbf{v},m}), \mathbf{r}(\phi^y_{\mathbf{u},\mathbf{v},m})\right\}_{m=0}^{M-1}\right). \tag{5}$$

Since $\mathbf{r}(\alpha)^\top \mathbf{r}(\beta) = \mathbf{r}(\beta - \alpha)$, the position-wise rotations of queries and keys reduce to this relative rotation in their inner product:

$$a_{\mathbf{u},\mathbf{v}} = (\mathbf{Q}_\mathbf{u})^\top \mathbf{R}_{\mathbf{\Delta}_{\mathbf{u},\mathbf{v}}} \mathbf{K}_\mathbf{v}. \tag{6}$$

Standard RoPE determines its positional transformation solely from the indexed displacement $\mathbf{\Delta}_{\mathbf{u},\mathbf{v}}$ and the fixed frequency schedule $\{\theta_m\}$. Consequently, identical indexed displacements receive the same positional transformation regardless of their surrounding spatial variation. SV-RoPE instead makes this positional transformation responsive to the spatial variation associated with tokens.

3.3 SV-RoPE: Spatial-Variation-Aware Rotary Adaptation

3.3.1 Overview

We propose SV-RoPE, a spatial-variation-aware rotary adaptation mechanism for pretrained RSFMs. SV-RoPE characterizes each token with an *SV-Profile*, which summarizes how its feature affinity with surrounding tokens changes across spatial extents. These token-wise profiles are then used to make the rotary response sensitive to spatial variation in the target observation.

SV-RoPE calibrates the standard rotary frequency with a relation-specific coefficient $\kappa_{\mathbf{u},m}$:

$$\theta^{\mathrm{SV}}_{\mathbf{u},m} = \theta_m \kappa_{\mathbf{u},m}, \qquad \kappa_{\mathbf{u},m} = \rho_{\ell,m}\, \alpha_{\mathbf{u},m}. \tag{7}$$

Here, $\kappa_{\mathbf{u},m}$ jointly calibrates how the pretrained rotary frequency is expressed at token $\mathbf{u}$. Its factorization combines the token-step ground extent through ρ_ℓ with SV-Profile-conditioned sensitivity through $\alpha_{\mathbf{u},m}$. We parameterize these two terms using *Relation Coordinate Calibration (RCC, Section 3.3.2)* and *Relation Frequency Calibration (RFC, Section 3.3.3)*, respectively.

The calibrated frequency defines the SV-RoPE rotary phase as $\psi^{\mathrm{SV}}_{\mathbf{u},m} = p_\mathbf{u}\theta_m\kappa_{\mathbf{u},m}$.. The relative phase between tokens $\mathbf{u}$ and $\mathbf{v}$ is then induced by the difference between their token-wise phases:

$$\phi^{\mathrm{SV}}_{\mathbf{u},\mathbf{v},m} = \psi^{\mathrm{SV}}_{\mathbf{u},m} - \psi^{\mathrm{SV}}_{\mathbf{v},m} = \theta_m \left(p_\mathbf{u}\kappa_{\mathbf{u},m} - p_\mathbf{v}\kappa_{\mathbf{v},m}\right). \tag{8}$$

The calibrated phases define the pairwise SV-RoPE rotation matrix:

$$\mathbf{R}^{\mathrm{SV}}_{\mathbf{\Delta}_{\mathbf{u},\mathbf{v}}} = \mathrm{blockdiag}\left(\left\{\mathbf{r}(\phi^{\mathrm{SV},x}_{\mathbf{u},\mathbf{v},m}), \mathbf{r}(\phi^{\mathrm{SV},y}_{\mathbf{u},\mathbf{v},m})\right\}_{m=0}^{M-1}\right). \tag{9}$$

Thus, SV-RoPE turns spatial variation into an explicit conditioning signal for positional adaptation. Collecting these induced pairwise rotations over all valid token pairs, we denote $\mathcal{R}^{\mathrm{SV}} = \left\{\mathbf{R}^{\mathrm{SV}}_{\mathbf{u},\mathbf{v}}\right\}_{\mathbf{u},\mathbf{v}}$. When $\rho_\ell = 1$ and $\alpha_{\mathbf{u},\mathbf{v},m} = 1$, $\mathbf{R}^{\mathrm{SV}}_{\mathbf{u},\mathbf{v}}$ reduces to the standard RoPE rotation $\mathbf{R}_{\mathbf{\Delta}_{\mathbf{u},\mathbf{v}}}$.

3.3.2 Relation Coordinate Calibration

In RCC, the factor ρ_ℓ in Eq. (7) provides a ground-scale reference for the pretrained rotary frequency. Let ℓ_{ref} denote a reference token-step ground distance, set by default to the finest such distance among the RSFM pretraining inputs. For the current token-step ground distance ℓ, RCC defines the physical-scale factor:

$$\rho_\ell = \left(\frac{\ell}{\ell_{\mathrm{ref}}}\right)^{1-\gamma}, \qquad \gamma \in [0, 1], \tag{10}$$

where ℓ_{ref} denotes the reference ground distance. The pretrained frequency bank is correspondingly scaled by γ to preserve its frequency structure under coordinate calibration.

Equivalently, the physical-scale factor can be interpreted as rescaling the indexed displacement into a ground-referenced displacement:

$$\widetilde{\mathbf{\Delta}}_{\mathbf{u},\mathbf{v}} = \rho_\ell \mathbf{\Delta}_{\mathbf{u},\mathbf{v}}. \tag{11}$$

The exponent γ governs calibration strength: $\gamma = 1$ preserves raw grid offsets ($\rho_\ell = 1$), $\gamma = 0$ expresses displacements relative to ℓ_{ref}, and intermediate values smoothly interpolate between them.

3.3.3 Relation Frequency Calibration

RFC derives the variation-dependent factor $\alpha_{\mathbf{u},m}$ in Eq. (7) from token-wise SV-Profiles.

Spatial-variation profile. Given the spatial token map $\mathbf{Z} \in \mathbb{R}^{N_x \times N_y \times D}$ from the frozen encoder, we characterize the local representation variation around token $\mathbf{u}$ by measuring how its feature affinity to surrounding tokens changes over multiple spatial extents $\mathcal{W} = \{w_1, \ldots, w_K\}$. For extent w_k, we define its neighborhood affinity as:

$$s_{\mathbf{u}}^{(k)} = \frac{1}{|\mathcal{N}_{w_k}(\mathbf{u})|} \sum_{\mathbf{v} \in \mathcal{N}_{w_k}(\mathbf{u})} \frac{\mathbf{Z}_{\mathbf{u}}^{\top} \mathbf{Z}_{\mathbf{v}}}{\|\mathbf{Z}_{\mathbf{u}}\|_2 \|\mathbf{Z}_{\mathbf{v}}\|_2}, \tag{12}$$

where $\mathcal{N}_{w_k}(\mathbf{u})$ denotes the neighborhood of $\mathbf{u}$ at spatial extent w_k. Repeating this computation from local to progressively larger neighborhoods produces a multi-scale affinity profile $\mathbf{s}_{\mathbf{u}} = \left[s_{\mathbf{u}}^{(1)}, \ldots, s_{\mathbf{u}}^{(K)}\right]$. where $\mathbf{s}_{\mathbf{u}}$ summarizes how the feature affinity surrounding $\mathbf{u}$ changes as the spatial extent increases.

Profile-conditioned frequency calibration. RFC maps each SV-Profile to frequency-wise modulation factors:

$$\alpha_{\mathbf{u},m} = \exp\left(\tanh\left(\mathcal{F}_m(\mathbf{s}_{\mathbf{u}})\right)\right), \tag{13}$$

where $\mathcal{F}_m$ is a lightweight profile encoder and $\alpha_{\mathbf{u},m} \in (e^{-1}, e)$ denotes the modulation at frequency m. The bounded positive modulation adjusts the rotary sensitivity at each frequency according to the spatial variation associated with $\mathbf{u}$.

For a relation $(\mathbf{u}, \mathbf{v})$, we symmetrically apply token-wise modulations to both endpoints and compute their relative interaction via the phase difference in Eq. (8), where the symmetric formulation ensures that reversing the direction of the relation negates the relative phase:

$$\phi^{\mathrm{SV}}_{\mathbf{v},\mathbf{u},m} = -\phi^{\mathrm{SV}}_{\mathbf{u},\mathbf{v},m}. \tag{14}$$

As each token modulates its own coordinates based on its local context, the interaction is coordinate-aware rather than purely translation-invariant. This enables identical grid offsets to yield diverse rotary sensitivities based on the SV-Profiles of the interacting tokens.

3.4 Parameter-Efficient SV-RoPE Injection

SV-RoPE is integrated into each Transformer block through a lightweight attention-parallel adapter (Zhang et al., 2026a), while the pretrained attention and positional pathway remain frozen. Given input $\mathbf{Z}$, the frozen branch produces $\mathbf{Y}_{\mathrm{frozen}}$, and the trainable SV-RoPE branch computes

$$\Delta\mathbf{Y} = \mathrm{Attn}_{\text{SV-RoPE}}\left(\mathbf{Z}; \mathcal{R}^{\mathrm{SV}}\right). \tag{15}$$

The block output is then calculated as: $\mathbf{Y} = \mathbf{Y}_{\mathrm{frozen}} + \Delta\mathbf{Y}$. During adaptation, only the SV-RoPE and prediction head are optimized, yielding a parameter-efficient update to the pretrained RSFM.

4 Experiment

4.1 Experimental Settings

Datasets and metrics. Our evaluation is conducted on the commonly used GeoBench (Lacoste et al., 2023), including six *classification* datasets: m-bigearthnet (BE), m-brick-kiln (BK), m-eurosat (ES), m-forestnet (FN), m-pv4ger (PV), m-so2sat (SS), and six *semantic segmentation* datasets: m-cashew-plantation (CW), m-NeonTree (NT), m-nz-cattle (CT), m-chesapeake-landcover (CP), m-pv4ger-seg (PS) and m-SA-crop-type (SC), spanning GSDs from 0.1 m (RGB aerial) to 15 m (Landsat-8). For classification, we report the mean Accuracy score. For segmentation, we report Intersection over Union (IoU) for 2-classes datasets and mean IoU (mIoU) otherwise. More details are in Appendix B.1.

Backbones and baselines. We evaluate SV-RoPE along two complementary dimensions. First, to examine its compatibility with different pretrained priors, we instantiate SV-RoPE on representative RSFMs with diverse native positional encodings: *DOFA* (Xiong et al., 2024a) with absolute positional encoding (APE), *Scale-MAE* (Reed et al., 2023) with GSD-aware positional encoding, and *DiNOv3-SAT* (Siméoni et al., 2025) with vanilla RoPE for heterogeneous EO inputs. Across all backbones, SV-RoPE is attached as an attention-parallel rotary adapter and does not require backbone-specific

Table 1: *Comparison of performances on GeoBench.* Avg.C, Avg.S, and Avg denote the macro-average performance over classification, segmentation, and all datasets, respectively. The best and second-best results within each backbone are highlighted in **bold** and underline, respectively.

Method	Classification						Segmentation						Average		
	BE	BK	ES	FN	PV	SS	CW	NT	CT	CP	PS	SC	Avg.C	Avg.S	Avg
DoFA															
LP/UP	87.90	96.80	87.00	41.47	96.99	55.10	61.20	57.40	73.80	56.92	92.60	31.00	77.54	62.15	69.85
BitFit	89.41	98.30	95.82	50.60	97.99	57.20	67.06	55.64	74.99	59.23	94.03	33.48	81.55	64.07	72.81
NormTuning	89.45	98.49	94.04	52.44	98.10	58.62	66.38	55.49	74.78	58.39	93.81	32.88	81.86	63.62	72.74
LoRA	88.23	98.30	98.13	54.79	97.90	60.24	60.67	55.89	75.14	60.86	94.64	35.37	82.93	63.76	73.35
SLR	88.86	98.30	97.20	53.74	98.00	61.56	68.19	56.16	75.29	60.34	94.10	34.10	82.94	64.70	73.82
SV-RoPE	**89.68**	**98.50**	**98.31**	**57.19**	**98.80**	**63.49**	**68.57**	**60.34**	**75.64**	**63.36**	**95.01**	**35.68**	**84.33**	**66.43**	**75.38**
ScaleMAE															
LP/UP	**72.50**	84.78	72.76	33.98	87.60	32.96	58.34	56.96	72.60	53.47	90.63	29.16	64.10	60.19	62.15
BitFit	66.39	96.39	91.75	35.78	91.19	47.16	58.74	55.27	72.60	55.49	90.88	**29.43**	71.44	60.40	65.92
NormTuning	68.23	91.69	79.11	35.03	90.49	40.97	58.48	55.33	72.83	53.08	90.72	28.54	67.59	59.83	63.71
LoRA	68.37	97.49	94.73	46.71	95.29	47.57	59.96	55.09	73.24	61.73	92.55	25.79	75.03	61.39	68.21
SLR	68.12	97.20	96.05	**47.75**	94.90	53.04	59.32	54.59	73.26	65.12	95.20	26.31	76.18	62.30	69.24
SV-RoPE	69.18	**97.80**	**97.80**	**47.75**	**97.70**	**53.96**	**63.34**	**57.21**	**74.64**	**73.41**	**95.29**	26.74	**77.37**	**65.11**	**71.24**
DINOv3-SAT															
LP/UP	**74.98**	96.40	95.33	48.50	95.40	60.55	68.48	60.05	74.92	68.87	94.69	35.07	78.53	67.01	72.77
BitFit	71.22	97.10	98.76	51.05	97.60	60.45	69.78	61.31	74.87	71.91	94.78	36.20	79.36	68.14	73.75
NormTuning	71.99	97.30	98.69	52.54	97.10	60.85	69.72	60.53	74.85	71.57	94.84	36.28	79.75	67.97	73.86
LoRA	74.06	97.90	99.06	53.12	97.90	63.59	70.45	63.55	75.00	74.81	95.00	36.89	80.94	69.28	75.11
SLR	74.88	97.90	**99.17**	54.34	97.90	64.02	70.23	64.01	75.23	75.09	94.95	37.88	81.37	69.57	75.47
SV-RoPE	74.86	**98.20**	99.16	**55.84**	**98.20**	**64.30**	**72.51**	**66.25**	**75.65**	**75.49**	**95.16**	**38.30**	**81.76**	**70.56**	**76.16**

redesign. Second, to assess SV-RoPE as a parameter-efficient adaptation method, we compare it with representative PEFT baselines, including *BitFit* (Zaken et al., 2022), *NormTuning* (Zhao et al., 2023), *LoRA* (Hu et al., 2022), and *Scaled Low-Rank* (*SLR*) (Scheibenreif et al., 2024). For all LoRA-based methods, the rank is set to 8 following the common setting of Hu et al. (2022).

Implementation details. Following the base model settings, for classification, we follow a linear probing protocol, in which a `[CLS]` token is prepended to the average-pooled token embeddings for prediction; for segmentation, we use the UperNet (Xiao et al., 2018) as decoder on top of the frozen backbone. The total number of training epochs is fixed to 100 and SV-RoPE is optimized in the attention layers with $\gamma = 0.5$, following the standard PEFT setup. More details are in Appendix B.2.

4.2 Overall Performance

Table 1 compares SV-RoPE with representative PEFT methods across three RSFMs. SV-RoPE matches or exceeds the strongest baseline on 12/12, 10/12, and 10/12 downstream tasks with DoFA, ScaleMAE and DINOv3-SAT respectively. This consistency highlights the benefit of explicitly modeling spatial variation beyond generic parameter-efficient updates. The improvement on Scale-MAE is particularly informative. Its pretraining already combines GSD-aware positional encoding with multi-frequency reconstruction, yet SV-RoPE improves Avg over SLR by 2.00 points. The additional gains on DOFA and DINOv3-SAT, at 1.56 and 0.69 points, respectively, support the value of using spatial variation as a target-specific adaptation signal for pretrained representations. The gains also exhibit a consistent task-level pattern. Across DOFA, Scale-MAE, and DINOv3-SAT, the increases in Avg.S exceed those in Avg.C, indicating particularly consistent benefits for dense prediction. This pattern is consistent with our motivation to make adaptation responsive to spatial variation throughout the target observation. Moreover, qualitative visualizations are shown in Fig. 3.

4.3 Ablation Study

We analyze the contributions of SV-RoPE components and alternative designs on DoFA as in Table 2.

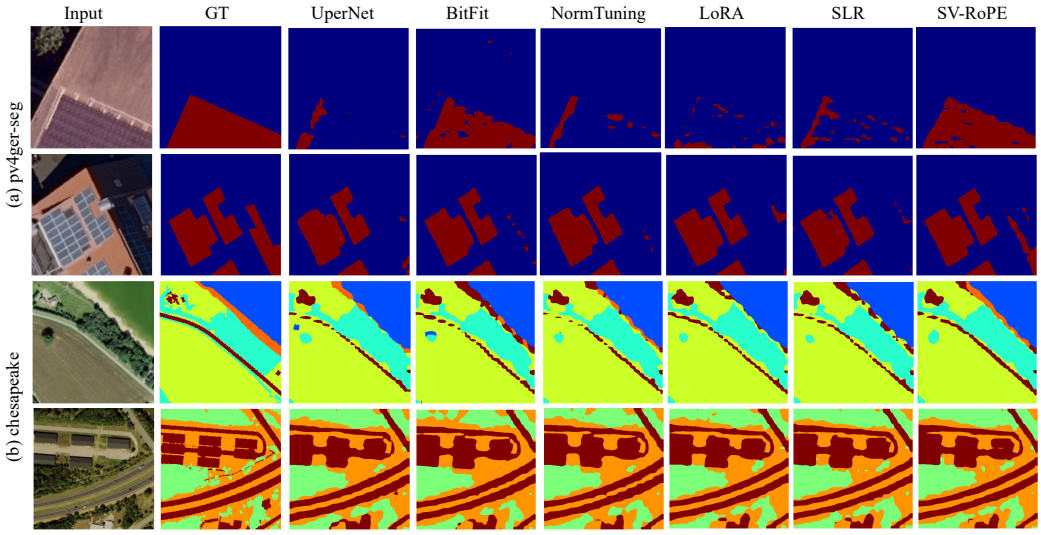


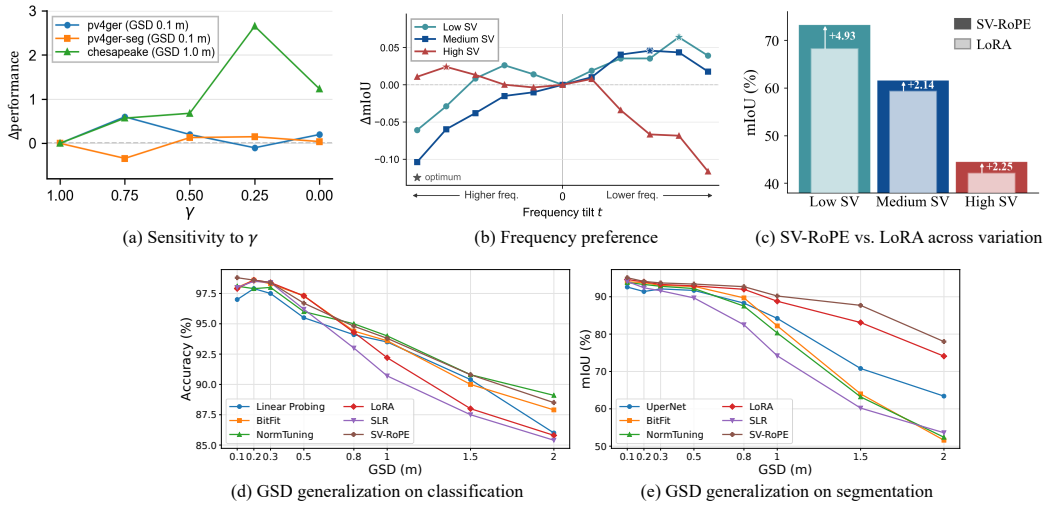
Figure 3: Qualitative comparison of different PEFT methods on representative segmentation datasets.

Table 2: *Ablation study of SV-RoPE.* We evaluate each individual components and alternative designs.

Variant	Classification						Segmentation						Average		
	BE	BK	ES	FN	PV	SS	CW	NT	CT	CP	PS	SC	Avg.C	Avg.S	Avg
RCC Design Ablation															
LoRA	88.23	98.30	98.13	54.79	97.90	60.24	60.67	55.89	75.14	60.86	94.64	35.37	82.93	63.76	73.35
+ RoPE	88.52	98.40	97.78	54.19	98.10	62.40	62.78	59.21	75.60	62.25	94.73	34.72	83.23	64.88	74.06
+ RCC	88.69	98.40	98.19	55.89	98.30	63.59	63.70	59.09	75.48	63.39	94.99	33.59	83.84	65.04	74.44
RFC Design Ablation															
Global scaling	88.60	98.30	97.87	54.94	98.10	**64.71**	64.41	59.36	75.56	**64.39**	95.14	33.60	83.75	65.41	74.58
DAPE	89.21	98.20	97.85	52.25	97.90	62.78	64.36	59.09	75.13	61.44	94.96	**35.75**	83.03	65.12	74.08
SV-RoPE	**89.68**	**98.50**	**98.31**	**57.19**	**98.80**	63.49	**68.57**	**60.34**	**75.64**	63.36	**95.01**	35.68	**84.33**	**66.43**	**75.38**

Why calibrate indexed displacement with RCC? We first evaluate the contribution of coordinate calibration. Adding standard RoPE to LoRA improves the overall average from 73.35 to 74.06, RCC further raises the score to 74.44 by rescaling indexed displacement according to the token-step ground extent. The gains are particularly evident on FN (+1.70) and CP (+1.14), suggesting that accounting for scale-induced variation can improve how indexed relations are interpreted across observations. Fig. 4(a) further reveals two complementary effects regarding γ in RCC. When the target scale is closer to the pretrained reference, RCC provides a more direct performance gain through ground-distance calibration (chesapeake, GSD 1.0m). Under larger scale shifts (e.g., GSD $= 0.1$m), however, stronger physical calibration does not monotonically improve performance, suggesting that scale information alone cannot determine the appropriate positional response. This motivates RFC to incorporate SV-Profiles for variation-aware rotary adaptation.

Why condition frequency on SV-Profile with RFC? RCC calibrates the spatial extent of indexed relations, while the frequency response remains shared across tokens. To examine whether token-specific conditioning is necessary, we first average the SV-Profiles over all tokens and map the aggregated profile to a single frequency scaling factor. This global variant yields only a modest improvement over RCC (74.44→74.58). We further compare with DAPE (Zheng et al., 2024), which dynamically adapts positional interactions according to input-dependent context. Despite this finer-grained content adaptivity, DAPE reaches 74.08, below the RCC baseline and SV-RoPE. The gap indicates that content-dependent adaptation alone does not explicitly capture how spatial structure varies across extents, suggesting the SV-Profile as a dedicated conditioning signal.



(a) Sensitivity to γ (b) Frequency preference (c) SV-RoPE vs. LoRA across variation

(d) GSD generalization on classification (e) GSD generalization on segmentation

Figure 4: **Analysis of SV-RoPE calibration.** (a) Task-dependent sensitivity to the RCC exponent γ. (b) Controlled spectral-tilt probing reveals distinct frequency preferences across low-, medium-, and high-SV tokens. (c) SV-RoPE improves over LoRA across all three SV groups. (d) Performance curves under varying GSDs for classification and segmentation.

4.4 Analysis of Spatial-Variation-Aware Adaptation

Frequency preference across spatial-variation levels. To examine whether tokens with different spatial variation prefer different frequency adjustments, we conduct a diagnostic probe using a controlled RoPE variant in which only the frequency-band scaling is varied. We partition tokens from Chesapeake (CP) into equal-sized low-, medium-, and high-SV groups according to their SV-Profiles, and probe each group with a controlled frequency scaling $\alpha_m(t) = \exp\left(t(m - \bar{m})\right)$, where m indexes the rotary frequency bands from high to low frequency and $\bar{m}$ denotes the center index. When $t = 0$, all frequency bands remain unchanged; negative t emphasizes higher-frequency bands, whereas positive t emphasizes lower-frequency bands. In Fig. 4(b), the three groups exhibit distinct response curves and achieve their best performance at different values of t. This shows that frequency preference varies with spatial variation. Token-wise SV-Profiles thereby provide a flexible signal for adapting frequency scaling to the spatial variation of each observation.

Performance gains across spatial-variation levels. We further examine whether the benefit of SV-RoPE extends across different degrees of spatial variation. Using the same low-, medium-, and high-SV partition on Chesapeake (CP), we compare the prediction performance of LoRA and SV-RoPE within each group. As shown in Fig. 4(c), SV-RoPE improves all three groups, indicating that exploiting this variation-dependent frequency structure translates into better predictions across the SV spectrum. The larger gain for low-SV tokens further suggests that regions with slowly varying representations leave more room for correcting an overly rapid positional response.

Robustness under GSD shifts. We evaluate scale robustness by resampling PV4GER across varying GSDs on DoFA (Fig. 4). While classification degrades gracefully with marginal gaps among methods, segmentation exhibits marked sensitivity to coarser resolutions. Here, SV-RoPE degrades far more gracefully, widening its lead over competing PEFT baselines especially at 1.5–2.0 m. This confirms that SV-RoPE excels at mitigating scale shifts, particularly for dense prediction.

5 Conclusion

We presented SV-RoPE, a parameter-efficient method that explicitly incorporates local spatial variation into the downstream adaptation of RSFMs. By characterizing multi-scale neighborhood affinities via token-wise SV-Profiles, SV-RoPE adaptively calibrates rotary positional frequencies through a lightweight attention-parallel branch Our results highlight spatial variation as an effective signal for adapting pretrained spatial modeling to downstream EO imagery.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing of the manuscript. We have not used generative AI tools for generating research ideas and experimental results. We have reviewed all AI-assisted content, and we take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work adapts pretrained remote sensing foundation models on publicly available benchmark datasets (GEO-Bench) that do not contain personally identifiable information. The proposed method is a parameter-efficient positional adaptation mechanism and does not introduce new data collection or human-subject studies. We do not foresee direct negative societal impacts beyond those common to Earth observation research.

REPRODUCIBILITY STATEMENT

We describe the SV-RoPE formulation in Section 3. Dataset details, training hyper-parameters, and compute resources are provided in Appendix B. Our code will be publicly available upon acceptance.

REFERENCES

Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. Omnisat: Self-supervised modality fusion for earth observation. In *European Conference on Computer Vision*, pp. 409–427. Springer, 2024.

Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. Anysat: One earth observation model for many resolutions, scales, and modalities. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Conference*, pp. 19530–19540, 2025.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. Longrope: Extending llm context window beyond 2 million tokens. *arXiv preprint arXiv:2402.13753*, 2024.

Minh Kha Do, Kang Han, Phu Lai, Khoa T Phan, and Wei Xiang. Robsense: A robust multi-modal foundation model for remote sensing with static, temporal, and incomplete data adaptability. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Conference*, pp. 7427–7436, 2025.

Zhe Dong, Yanfeng Gu, and Tianzhu Liu. Upetu: A unified parameter-efficient fine-tuning framework for remote sensing foundation model. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–13, 2024.

Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, et al. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27672–27683, 2024.

Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *European Conference on Computer Vision*, pp. 289–305. Springer, 2024.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pp. 2790–2799. PMLR, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/pdf/2106.09685.

Jie Huang, Xuejing Liu, Sibo Song, Ruibing Hou, Hong Chang, Junyang Lin, and Shuai Bai. Revisiting multimodal positional encoding in vision–language models. In *International Conference on Learning Representations*, volume 2026, pp. 124522–124539, 2026.

Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan Sherwin, Hannah Kerner, Björn Lütjens, Jeremy Irvin, David Dao, Hamed Alemohammad, Alexandre Drouin, et al. Geo-bench: Toward foundation models for earth monitoring. *Advances in Neural Information Processing Systems*, 36: 51080–51093, 2023.

Haobo Li, Eunseo Jung, Zixin CHEN, Zhaowei Wang, Yueya WANG, Huamin Qu, and Alexis Lau. Pipe: Physics-informed position encoding for alignment of satellite images and time series in typhoon forecasting. *Advances in Neural Information Processing Systems*, 38:52668–52698, 2025a.

Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. Cameras as relative positional encoding. *Advances in Neural Information Processing Systems*, 38:15984–16009, 2025b.

Xuyang Li, Chenyu Li, and Danfeng Hong. Any-optical-model: A universal foundation model for optical remote sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6539–6547, 2026.

Yu Luo, Xiaogang Zhu, Shan Zeng, Wei Xiang, Thomas Francis Bishop, Zhiyong Wang, and Kun Hu. Phenoyieldnet: Learning crop-aware phenological responses for multi-crop yield prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15870–15879, June 2026.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*, 2023.

Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *IEEE/CVF International Conference on Computer Vision*, pp. 4088–4099, 2023.

Shreelekha Revankar, Cheng Perng Phoo, Utkarsh Kumar Mall, Bharath Hariharan, and Kavita Bala. Scale-aware recognition in satellite images under resource constraints. In *International Conference on Learning Representations*, volume 2025, pp. 54725–54737, 2025.

Linus Scheibenreif, Michael Mommert, and Damian Borth. Parameter efficient self-supervised geospatial domain adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27841–27851, 2024.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Romain Thoreau, Valerio Marsocci, and Dawa Derksen. Parameter-efficient adaptation of geospatial foundation models through embedding deflection. *arXiv preprint arXiv:2503.09493*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing systems*, 30, 2017.

Yi Wang, Zhitong Xiong, Chenying Liu, Adam J. Stewart, Thomas Dujardin, Nikolaos Ioannis Bountos, Angelos Zavras, Franziska Gerken, Ioannis Papoutsis, Laura Leal-Taixé, and Xiao Xiang Zhu. Towards a unified copernicus foundation model for earth vision. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9888–9899, October 2025.

Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Jian Tong, Haodong Duan, Qipeng Guo, Jiaqi Wang, et al. Videorope: What makes for good video rotary position embedding? *arXiv preprint arXiv:2502.05173*, 2025.

Kan Wu, Houwen Peng, Minghao Chen, Jianlong Fu, and Hongyang Chao. Rethinking and improving relative position encoding for vision transformer. In *IEEE/CVF International Conference on Computer Vision*, pp. 10033–10041, 2021.

Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *European Conference on Computer Vision (ECCV)*, pp. 418–434, 2018.

Zhitong Xiong, Yi Wang, Fahong Zhang, Adam J Stewart, Joëlle Hanna, Damian Borth, Ioannis Papoutsis, Bertrand Le Saux, Gustau Camps-Valls, and Xiao Xiang Zhu. Neural plasticity-inspired multimodal foundation model for earth observation. *arXiv preprint arXiv:2403.15356*, 2024a.

Zhitong Xiong, Yi Wang, Fahong Zhang, and Xiao Xiang Zhu. One for all: Toward unified foundation models for earth vision. In *IEEE International Geoscience and Remote Sensing Symposium*, pp. 2734–2738. IEEE, 2024b.

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Annual Meeting of the Association for Computational Linguistics*, pp. 1–9, 2022.

Cheng Zhang, Boying Li, Meng Wei, Yan-Pei Cao, Camilo Gambardella, Dinh Phung, and Jianfei Cai. Unified camera positional encoding for controlled video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38027–38037, 2026a.

Yuxiang Zhang, Wei Li, Mengmeng Zhang, Jiawei Han, Ran Tao, and Shunlin Liang. Spectralx: Parameter-efficient domain generalization for spectral remote sensing foundation models. *ISPRS Journal of Photogrammetry and Remote Sensing*, 239:774–792, 2026b.

Bingchen Zhao, Haoqin Tu, Chen Wei, Jieru Mei, and Cihang Xie. Tuning layernorm in attention: Towards efficient multi-modal llm finetuning. *arXiv preprint arXiv:2312.11420*, 2023.

Chuanyang Zheng, Yihang Gao, Han Shi, Minbin Huang, Jingyao Li, Jing Xiong, Xiaozhe Ren, Michael Ng, Xin Jiang, Zhenguo Li, et al. Dape: Data-adaptive positional encoding for length extrapolation. *Advances in Neural Information Processing Systems*, 37:26659–26700, 2024.

Dawei Zhu, Nan Yang, Liang Wang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. Pose: Efficient context window extension of llms via positional skip-wise training. *arXiv preprint arXiv:2309.10400*, 2023.