SPEXT: A Decoupled Multi-Spectral Foundation Model for Earth Observation and Vision-Language Grounding

Sannara Ek
University of Cambridge
Cambridge, United Kingdom
sannara.ek@cl.cam.ac.uk

Chong Tang*
University of Bristol
University of Southampton
Bristol, United Kingdom
chong.tang@bristol.ac.uk

Dirk Koch
University of Manchester
Manchester, United Kingdom
dirk.koch@manchester.ac.uk

Alex Weddell
University of Southampton
Southampton, United Kingdom
asw@ecs.soton.ac.uk

Jagmohan Chauhan
University College London
UCL AI Centre
London, United Kingdom
jagmohan.chauhan@ucl.ac.uk

Robert Mullins
University of Cambridge
Cambridge, United Kingdom
robert.mullins@cl.cam.ac.uk

Abstract

Vision-Language Models for Earth Observation are bottlenecked at the visual encoder. Existing pipelines either compress multi-spectral signals into RGB pixels or collapse them into a single CLIP-style global vector, losing spectral information or spatial structure respectively. **SPEXT** resolves this through architectural decoupling. A wavelength-parameterized encoder produces multi-spectral patch tokens for dense tasks and VLM grounding, while a dedicated alignment token receives an image-text alignment loss for retrieval. SPEXT simultaneously leads the PANGAEA segmentation benchmark ($+2.51$ mIoU), supports zero-shot retrieval from the same forward pass (48.4 mAP@100 averaged over three benchmarks), and enables a frozen VLM, paired with a small projection adapter, to answer Earth-Observation questions directly from its tokens, outperforming RGB-pixel and multi-spectral CLIP baselines on the RSVQAxBEN benchmark.

1 Introduction

Earth Observation (EO) is increasingly accessed through Vision-Language Models (VLMs) and Retrieval-Augmented Generation (RAG) pipelines Wen et al. [2025], Sudmanns et al. [2020], Mai et al. [2024], where a user issues a free-form natural-language query, a retriever surfaces relevant satellite imagery, and a frozen or lightly-tuned VLM produces a grounded answer. The value of such pipelines is limited by a single component, the visual encoder that feeds the VLM. Existing encoders take one of two forms, each compromising the system in a different way. RGB-rendering encoders pass percentile-stretched pixels with at most three of the ten or more available bands, discarding the Red-Edge and Short-Wave Infrared signatures that separate crop varietals, mineral classes, and other spectrally similar surfaces Clevers and Gitelson [2013], Lillesand et al. [2015]. CLIP-style multi-spectral encoders Radford et al. [2021], Mall et al. [2024], Marimo et al. [2025], Liu et al. [2024] collapse the spectral signal into a single global vector that is text-searchable but loses the spatial structure needed for dense downstream reasoning. Either way, the fine-grained physical content that makes multi-spectral imagery valuable is lost before it reaches the language model.

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

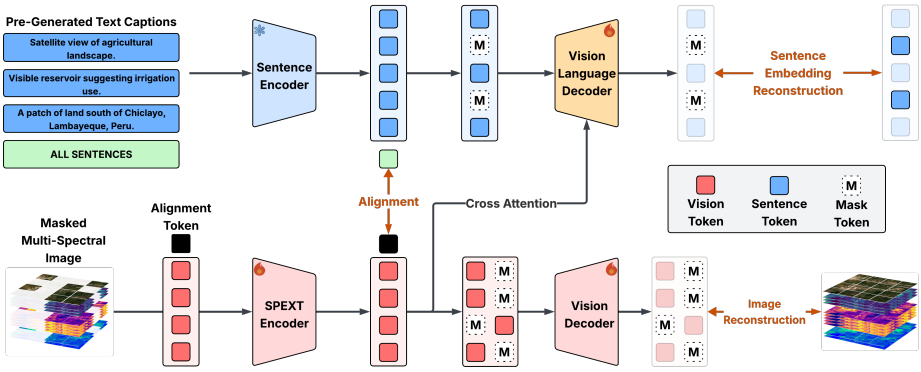


Figure 1: **SPEXT pre-training architecture. (Bottom)** The visual branch encodes masked multi-spectral imagery and reconstructs pixels via an MAE Vision Decoder, with a dedicated *Alignment Token* carrying the image-text alignment signal. **(Top)** The linguistic branch encodes captions with a frozen Sentence Encoder; a Vision-Language Decoder reconstructs masked sentence embeddings by cross-attending to the visual patches (Latent MTP). The two text-supervision signals are routed to disjoint feature streams.

Existing remedies treat the two failures separately. Physics-aware encoders such as DOFA Xiong et al. [2024] and PhySwin Tang et al. [2025] preserve spectral content for dense tasks, while CLIP-style multi-spectral models enable retrieval, and the prevailing assumption is that the trade-off is fundamental, that text supervision must compress visual features to be useful for retrieval. We show this is not the case. Standard multi-modal training applies contrastive supervision directly to the dense feature stream, requiring every patch token to align with a global ranking signal that reshapes the feature space toward textual discriminability Yu et al. [2022]. Empirically, an image-text alignment loss on a dedicated token yields essentially no improvement over a pure MAE baseline on downstream segmentation, whereas a generative latent reconstruction objective on the same data yields a substantial improvement.

We introduce SPEXT (Fig. 1), a multi-spectral foundation model designed around two principles, physics-informed spectral encoding and decoupled semantic supervision. SPEXT decouples the two roles: a wavelength-parameterized encoder produces multi-spectral patch tokens that preserve spectral continuity for dense downstream tasks and direct VLM grounding, while a dedicated alignment token receives the image-text alignment gradient for retrieval. Text supervision is applied to the dense stream only through a generative latent reconstruction objective Yuan et al. [2024], which regularizes the visual features without compressing them toward a textual feature space. The result is a single encoder that simultaneously leads the PANGAEA segmentation benchmark, supports zero-shot retrieval, and, paired with a small projection adapter, enables a frozen VLM to answer land-feature questions directly from SPEXT tokens, outperforming RGB-pixel and multi-spectral CLIP baselines on the RSVQAxBEN benchmark. Architecturally, this also collapses the standard two-encoder EO-RAG pipeline into a single forward pass, eliminating the need for per-query re-encoding at generation time.

Our contributions are as follows:

1. **Decoupled architecture.** SPEXT routes generative supervision to the patch-level stream and alignment supervision to a dedicated alignment token, so dense prediction and retrieval do not interfere.
2. **Two technical components enabling the design.** *(i) Latent Masked Text Prediction (MTP)*: a pre-training objective that reconstructs masked sentence-level embeddings via cross-attention over visual patches, avoiding the discrete-vocabulary and autoregressive overhead of CoCa/BLIP-style alternatives. *(ii) Wavelength-parameterized spectral embedding*: continuous RBF mixing and Fourier-feature modulation that preserve spectral continuity and transfer across optical sensors without retraining.
3. **Empirical results across representation, retrieval, and VLM benchmarks.** Using only an 87 M-parameter backbone and $\sim$1.7 M Sentinel-2 pre-training images (compute and

parameter context vs. baselines in App. A.17), SPEXT achieves the best average score on PANGAEA, supports zero-shot retrieval, and enables a frozen VLM, paired with a small projection adapter, to answer EO questions directly from multi-spectral tokens, outperforming RGB-pixel and multi-spectral CLIP baselines on the RSVQAxBEN benchmark.

2 Related Work

Multi-spectral encoders. Two predominant families of encoders are found in the EO literature. The first emphasizes spectral and spatial fidelity. SatMAE Cong et al. [2022], SpectralGPT Hong et al. [2024], Prithvi Jakubik et al. [2023], and Scale-MAE Reed et al. [2023] adopt masked image modeling on multi-spectral inputs, as do SMARTIES Sumbul et al. [2025] and AnySat Astruc et al. [2025] (the latter JEPA-style); DOFA Xiong et al. [2024], FlexiMo Li et al. [2025], and PhySwin Tang et al. [2025] add wavelength-aware mechanisms such as hypernetworks, soft smoothness losses, or grouped-band convolutions Noman et al. [2024] to respect the continuous nature of multi-spectral signals Goetz et al. [1985], Dehghan-Shoar et al. [2024]. TESSERA Feng et al. [2026] encodes annual time series, and RingMo Sun et al. [2023], Cross-Scale MAE Tang et al. [2023] and SkySense V2 Zhang et al. [2025b] use different input bands. These transfer well to dense prediction but produce no text-aligned representation. SPEXT mixes bands with a fixed RBF kernel computed from wavelengths, rather than a hypernetwork (DOFA) or discrete per-band projections (SMARTIES, App. A.6.1). The second family adapts CLIP-style contrastive learning Radford et al. [2021] to EO captions, e.g. RemoteCLIP Liu et al. [2024], MS-CLIP and its variants Marimo et al. [2025], and related models Huang et al. [2023], Mall et al. [2023], enabling retrieval but compressing the spectral signal into a single global vector without directly supervising the patch-level structure needed for dense prediction or VLM grounding. GAIR Liu et al. [2026] contrastively aligns Sentinel-2 imagery with street views and locations. SPEXT unites the two families in a multi-spectral encoder to serve both dense prediction and retrieval from disjoint feature streams produced in a single forward pass.

Contrastive vs. generative multi-modal pre-training. CLIP-style contrastive supervision Radford et al. [2021] reshapes the encoder feature space toward textual discriminability, serving global retrieval but risking a semantic bottleneck Huang et al. [2023], Yu et al. [2022] that compresses high-dimensional visual signals to align with low-dimensional captions. CoOp Zhou et al. [2022], CLIP-Adapter Gao et al. [2024] and VaMP Cheng and Han [2025] adapt a frozen RGB CLIP post hoc. SPEXT acts earlier, in multi-spectral pre-training, so these methods are complementary rather than direct baselines. Generative alternatives such as CoCa Yu et al. [2022] and Flamingo Alayrac et al. [2022] retain language supervision via autoregressive captioning, but operate over discrete word tokens and tightly couple the visual encoder to a linguistic decoder; in EO this is particularly problematic because captions are typically auto-generated and brief Yuan et al. [2024], Marimo et al. [2025]. Latent-target frameworks like I-JEPA Assran et al. [2023] and MAE He et al. [2022] predict in a latent space but stay within a single modality. SPEXT's Latent Masked Text Prediction (MTP) ports this latent-target idea cross-modally: it reconstructs masked sentence-level embeddings produced by a frozen sentence encoder through cross-attention over visual patches, providing dense semantic regularization without discrete vocabulary, autoregressive decoding, or surface-form leakage into the encoder. GeoLink Bai et al. [2025a], MTP's closest relative, regresses BERT embeddings of masked, geolocated OpenStreetMap objects, whereas MTP targets scene-level sentences.

3 Methodology

SPEXT is built around two principles, physics-informed spectral encoding and decoupled semantic supervision. Under the second, *image-text alignment supervision and dense spectral representation must be routed to disjoint feature streams of the encoder*. Standard CLIP-style training applies a contrastive image-text loss directly to the encoder's pooled output, requiring every patch token to align with a global ranking signal that ultimately reshapes the dense feature space toward textual discriminability. We find empirically that an image-text alignment (ITC) loss on a dedicated token provides essentially no benefit to downstream segmentation, while a generative latent text-prediction objective on the same backbone and data yields a substantial improvement (§4.4). The remainder of this section describes the three components that operationalize the two principles: a wavelength-parameterized spectral embedding (§3.1), the Latent Masked Text Prediction (MTP) objective routed

through the dense patch stream (§3.2), and a dedicated alignment token that receives the ITC gradient and serves as the retrieval representation (§3.3). The two-stage training curriculum is in §3.4.

3.1 Wavelength-Parameterized Spectral Embedding

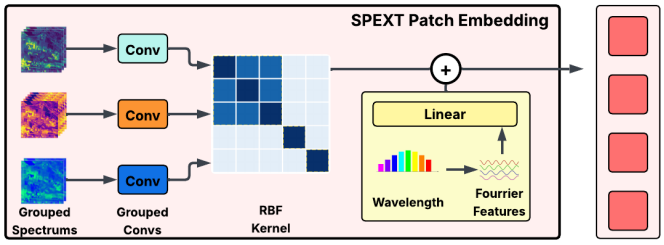


Figure 2: **SPEXT patch embedding.** Grouped band convolutions, smoothed by a wavelength-RBF kernel and modulated by Fourier features of the central wavelengths.

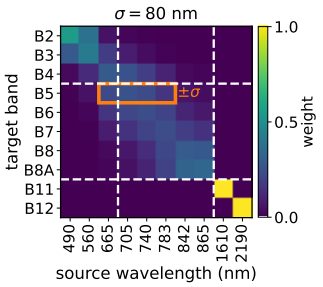


Figure 3: (Right) Normalized RBF kernel across Sentinel-2 bands with σ=80 nm. Dashed lines mark the Visible, RedEdge/NIR, and SWIR groupings. Example (orange box): B5 (705 nm) draws mostly on itself and on B4, B6 and B7, the bands within $\pm\sigma$, and nothing from SWIR.

Multi-spectral observations sample an underlying reflectance function $r(\lambda)$ at discrete band centers $\Lambda = \{\lambda_1, \ldots, \lambda_C\}$. Standard patch embedding treats the C channels as index-ordered and independent, discarding this continuous structure. SPEXT's embedding is parameterized by absolute wavelength so that information sharing is determined by proximity on the wavelength axis rather than channel index (Fig. 2). Following prior multi-spectral architectures Cong et al. [2022], Hong et al. [2024], Tang et al. [2025], we use the 10 primary surface-reflectance bands of Sentinel-2 (excluding B01, B09, B10). The embedding has three stages.

Grouped spectral projection. Bands are partitioned into three physically distinct families, Visible, Red-Edge/NIR, and SWIR, each processed by an independent convolutional projection. This allocates separate capacity to spectral regions with different sensor characteristics while retaining an explicit per-band axis for subsequent interactions.

Continuous spectral message passing. For projected patches $X \in \mathbb{R}^{H\times W\times C}$, we mix information across bands at every spatial location using a row-stochastic RBF kernel $K \in \mathbb{R}^{C\times C}$:

$$K_{ij} = \frac{\exp\!\big(-(\lambda_i - \lambda_j)^2/2\sigma^2\big)}{\sum_{k=1}^{C} \exp(-(\lambda_i - \lambda_k)^2/2\sigma^2)}, \qquad X'_{h,w,i} = \sum_{j=1}^{C} K_{ij}\, X_{h,w,j}. \tag{1}$$

This is a Nadaraya–Watson kernel smoother Nadaraya [1964], Watson [1964] on the sampled spectrum: each band feature becomes a locally-weighted average of its spectral neighbors, suppressing discretization artifacts while keeping mixing local in λ. The resulting wavelength-dependent mixing weights are visualized in Fig. 3; downstream performance is stable for σ between 40 and 160 nm (App. A.8).

Wavelength-conditioned modulation. To preserve the unique physical identity of each band after smoothing, we map central wavelengths to Fourier features $\phi(\lambda) = [\sin(2\pi\lambda f_k), \cos(2\pi\lambda f_k)]_{k=0}^{F-1}$, project them to per-band scale and bias parameters $\psi(\lambda), \zeta(\lambda)$, and apply

$$X_{\text{final}} = X' \odot \big(1 + \tanh(\psi(\lambda))\big) + \zeta(\lambda). \tag{2}$$

The layer is parameterized by absolute wavelengths rather than channel indices, so it transfers across optical sensors without architectural changes (App. A.11: 98.8% of native Sentinel-2 performance on Landsat HLS using only the new sensor's wavelengths, with the NIR group convolution re-initialized). SPEXT also preserves spectral structure end-to-end; we report the full reconstruction-fidelity numbers in §4.1 (Tab. 2).

3.2 Latent Masked Text Prediction (MTP)

To supply dense semantic supervision *without* the discriminative bottleneck of CLIP-style training, we introduce **Latent Masked Text Prediction (MTP)**. Unlike masked language modeling, MTP operates entirely at the sentence level in a continuous embedding space. **There is no discrete vocabulary, BPE tokenizer, or word embedding anywhere in MTP.**

Avoiding the contrastive bottleneck. CLIP-style training requires every patch token to align with a global ranking signal, reshaping the entire dense feature space toward textual discriminability. MTP routes the semantic gradient through a cross-attention decoder instead. The gradient flows through the decoder back into the patches, but only requires those patches to be *sufficiently informative* for the decoder to recover the sentence embedding, rather than aligned with the text feature space themselves. Consistent with this, in §4.4 MAE+MTP improves over MAE alone by $+2.85$ F1 on MADOS.

Setup. For each image, we split its caption into sentences and encode each sentence s_i once with the frozen sentence encoder $\mathcal{E}_{\text{txt}}$ (Qwen3-Embedding-0.6B Zhang et al. [2025a]) into a single 1024-dimensional vector $\mathbf{z}_{t,i} = \mathcal{E}_{\text{txt}}(s_i) \in \mathbb{R}^{1024}$. The MTP input is the short sequence of these per-sentence vectors, $[\mathbf{z}_{t,1}, \mathbf{z}_{t,2}, \ldots, \mathbf{z}_{t,N}, \mathbf{z}_{t,\text{addr}}]$, where $\mathbf{z}_{t,\text{addr}}$ is the embedding of a metadata-derived geographic anchor sentence and the total length is typically 2–5 (Fig. 5). We emphasize that each position in this sequence is one whole sentence already collapsed into a dense vector. This sentence-level unit suits EO captions, which vary more in wording than in scene content (Tab. 9). At training time, 50% of the positions are replaced by a single learnable mask vector $\mathbf{m} \in \mathbb{R}^{1024}$. A small transformer decoder, with self-attention over the sentence-vector positions and cross-attention into the visual patches, predicts the masked positions; we supervise with a cosine-plus-MSE objective, $(1 - \cos) + 0.25\,\text{MSE}$, between the prediction $\hat{\mathbf{z}}_{t,i}$ and the original frozen vector $\mathbf{z}_{t,i}$, computed only over the masked positions. Because $\mathcal{E}_{\text{txt}}$ is frozen, paraphrasing and word-tokenization noise are absorbed before any signal reaches the visual encoder; we additionally probe robustness to caption corruption directly in App. A.9 and find that spectrally-driven PANGAEA tasks (Sen1Floods11, HLS, SpaceNet7) remain within 1.16 mIoU from 25% to 75% caption swap or dropout. The geographic anchor sentence is added because SSL4EO-S12 captions are short and would otherwise often yield a length-1 sequence with no maskable position; its construction does not introduce downstream-evaluation leakage. Full caption statistics and the geocoding pipeline are in App. A.4; the MTP loss formula is in App. A.2.

3.3 Decoupled Alignment Token and Dual-Output Design

The ITC supervision is applied not to the patch tokens but to a dedicated **Alignment Token** $\mathbf{v}_{\text{cls}}$ prepended to the visual sequence. Architecturally, its sole role is to receive the ITC loss. We align $\mathbf{v}_{\text{cls}}$ to the holistic sentence embedding produced by the same frozen $\mathcal{E}_{\text{txt}}$ with a cosine objective on matched pairs, $\mathcal{L}_{\text{ITC}} = 1 - \cos(\text{proj}(\mathbf{v}_{\text{cls}}), \mathbf{z}_t)$, without in-batch negatives or a learnable temperature. Keeping $\mathcal{E}_{\text{txt}}$ frozen enables pre-computation of the text side offline. SPEXT uses no CLIP model or CLIP weights. The exact loss is in App. A.2.

Dual-output design. A single SPEXT forward pass produces two output streams used disjointly downstream. The **patch-level features** feed all dense prediction tasks and are also fed to the downstream VLM (§4.3). The alignment token is discarded for these tasks, in the same way a CLS token is dropped before passing patch features to a UperNet. The **alignment token** is the retrieval representation, matched against query sentence embeddings produced by the same frozen $\mathcal{E}_{\text{txt}}$ for zero-shot text-to-image retrieval. Because both outputs come from the same forward pass, a SPEXT-based retrieval-augmented deployment requires neither a separate retrieval encoder nor re-encoding of retrieved imagery at generation time (App. A.16).

3.4 Two-Stage Pre-training

We adopt a two-stage pre-training curriculum following Cong et al. [2022], Hong et al. [2024].

Stage I, spectral warm-up. The encoder is pre-trained from random initialization, without external vision weights, on fMoW-S2 Cong et al. [2022] (around 712 k Sentinel-2 tiles) with a pure MAE pixel-reconstruction objective and spectral-group band dropout for 100 epochs. This optimizes the wavelength-parameterized embedding without any text-supervision interference, establishing a spectrally-faithful feature space that the joint stage builds on.

Stage II, multi-modal semantic grounding. The encoder is then trained jointly on SSL4EO-S12 Wang et al. [2023], around 1 M Sentinel-2 image-caption pairs across 250 k global locations and four seasons Marimo et al. [2025]. Stage II optimizes a joint loss

$$\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{\text{MAE}} + \beta\mathcal{L}_{\text{MTP}} + \gamma\mathcal{L}_{\text{ITC}} \tag{3}$$

with α=0.4, β=0.3, γ=0.3 balancing spectral reconstruction, latent text prediction, and alignment. We train for 50 epochs with a 4-epoch decoder warm-up before unfreezing the encoder; the rationale and full schedule are in App. A.1, training-loss curves in App. A.3, and sensitivities to the masking ratio (Tab. 7) and loss-coefficient triplet (α, β, γ) in App. A.7. Across both stages, SPEXT sees roughly 712k+1M $\approx$ 1.7M Sentinel-2 tiles in total.

4 Experiments

We evaluate SPEXT along the three claims stated in the abstract: **(i)** as a multi-spectral encoder for dense prediction (§4.1), **(ii)** as a zero-shot text-to-image retriever (§4.2), and **(iii)** as the visual encoder that grounds a frozen VLM in multi-spectral imagery, evaluated on the third-party RSVQAxBEN benchmark (§4.3). Each evaluation uses PANGAEA RSFMs or CLIP-style baselines. Component-level ablations isolating the contribution of the decoupling design are presented in §4.4.

Setup. SPEXT is a ViT-Base backbone with 128×128 inputs at patch size 8 and asymmetric masking, 90% on the visual stream and 50% on the text stream in Stage II. Downstream evaluation uses the standard PANGAEA fine-tuning protocol Marsocci et al. [2024] with UperNet decoders for segmentation. Full hyperparameters, optimizer, and dataset details are in App. A.1. Source code and pre-trained Stage II weights are released at `https://github.com/Sannaraek/SPEXT`.

4.1 Multi-Spectral Representation Learning

PANGAEA segmentation. We benchmark fine-tuning on the **PANGAEA benchmark** Marsocci et al. [2024], a unified evaluation framework for RSFMs that controls for fine-tuning recipes across eight datasets covering crops, floods, and buildings, and reports mean mIoU, Average Rank, and Average Distance to the Best (DtB) following Adorni et al. [2025].

Table 1: Comparison of SPEXT against pre-trained RSFMs on the full training data (100%) PANGAEA benchmark. Results in **bold** are best, while underlined are second best for each task. **Avg Rank** indicates mean ranking across tasks, and **Avg DtB** (Average Distance to the Best) represents the mean percentage-point difference between model performance and the best result achieved for each task.

Model (Parameters)	HLS	MADOS	PASTIS	Sen1Floods11	XView2	CropMap	SN7	AI4Farms	Average ↑	Avg Rank ↓	Avg DtB ↓
S12-MAE (22M)	81.91	49.90	32.03	87.79	50.44	45.80	57.13	24.69	53.71	10.38	8.86
S12-Data2Vec (22M)	81.91	44.36	34.32	88.15	51.36	54.03	58.23	24.23	54.57	8.50	7.99
S12-MoCo (22M)	81.58	51.76	34.49	89.26	51.59	48.58	57.64	25.38	55.04	8.13	7.53
S12-DINO (22M)	81.72	49.37	36.18	88.61	50.56	48.66	56.47	25.62	54.65	8.75	7.92
DOFA-CLIP (457M)	75.52	53.95	24.34	86.84	57.02	40.94	61.19	27.25	53.38	9.38	9.19
MS-CLIP (87M)	79.42	67.94	26.18	84.15	60.16	21.98	62.46	27.50	53.72	7.00	8.84
RemoteCLIP (87M)	76.59	60.00	18.23	74.26	57.41	52.05	57.76	25.12	52.68	9.63	9.89
GFM-Swin (87M)	76.90	64.71	21.24	72.60	59.15	46.98	60.89	27.19	53.71	8.25	8.86
SatlasNet (87M)	79.96	55.86	17.51	90.30	52.23	46.97	61.88	25.13	53.73	8.50	8.84
Prithvi (87M)	**83.62**	49.98	33.93	90.37	49.35	43.07	56.54	26.86	54.22	8.25	8.35
Scale-MAE (303M)	76.68	57.32	24.55	74.13	**60.72**	25.42	**62.96**	21.47	50.41	9.63	12.16
SpectralGPT (105M)	80.47	57.99	35.44	89.07	48.40	46.95	58.86	26.75	55.49	8.25	7.08
DOFA (112M)	80.63	59.58	30.02	89.37	59.64	51.33	61.84	27.07	57.44	5.50	5.13
CROMA (303M)	82.42	67.55	32.32	**90.89**	53.27	49.38	59.28	25.65	57.60	5.00	4.97
SPEXT (87M)	81.57	**69.35**	**41.28**	90.04	47.88	**57.53**	59.05	**34.18**	**60.11**	**4.75**	**2.46**

As shown in Tab. 1, SPEXT achieves the highest average mIoU (60.11), the best Average Rank (4.75), and the lowest Avg DtB (2.46), outperforming the strongest baseline CROMA by $+2.51$ mIoU using $3.5\times$ fewer parameters. SPEXT ranks first on four of eight individual tasks (MADOS, PASTIS, CropMap, AI4Farms). SPEXT is evaluated on all eight tasks, whereas recent multispectral encoders without text alignment (SMARTIES Sumbul et al. [2025], AnySat Astruc et al. [2025], TESSERA Feng et al. [2026]) report at most two of them under PANGAEA's protocol (App. A.21). To probe sample-efficiency in the low-data setting that EO practitioners typically face, we additionally evaluate on the standard PANGAEA 10% data-split protocol Marsocci et al. [2024], where SPEXT

also achieves the best average mIoU (49.92) and Avg DtB (5.99); the full per-task table and analysis are in App. A.10.

Spectral fidelity. The dense-prediction gains above are downstream evidence that SPEXT's wavelength-parameterized embedding preserves spectral content. We probe this directly with a reconstruction-fidelity test. We report the Spectral Angle Mapper (SAM, conventionally faithful below 5°) and the pixel-wise R^2 of three common spectral indices (NDVI, NDWI, NDBI; negative R^2 indicates the reconstruction is less informative than the per-band mean). Tab. 2 shows that SPEXT recovers all three indices with $R^2 > 0.89$ and reaches a SAM of 2.69°. At matched training length, the decoupled encoder slightly exceeds the coupled variant of §4.4 on every fidelity metric (App. A.18).

Table 2: Spectral fidelity on the SSL4EO validation set. SPEXT preserves spectral structure with SAM $<5^\circ$ and recovers all three indices with $R^2 > 0.89$.

Metric	SPEXT	20-ep. SPEXT	20-ep. coupled
SAM (deg) ↓	**2.69**	3.08	3.13
NDVI R^2 ↑	**0.8925**	0.870	0.851
NDWI R^2 ↑	**0.8998**	0.874	0.871
NDBI R^2 ↑	**0.8910**	0.872	0.866

4.2 Cross-Modal Retrieval

Text retrieval over EO archives is the principal interface for analysts who cannot fine-tune per task. We evaluate SPEXT's alignment token as a zero-shot text-to-image retriever on three specialized Remote Sensing datasets: EuroSAT Helber et al. [2019], BigEarthNet-S2 Sumbul et al. [2019], and METER-ML Zhu et al. [2022], where each text query ranks the whole corpus. Following MS-CLIP Marimo et al. [2025], we expand discrete class labels into natural-language narrative templates that match SPEXT's pre-training caption distribution (full templates in App. A.14).

SPEXT retrieves directly from its alignment token, so the encoder that serves dense prediction and VLM grounding also answers free-text queries, without a separate retrieval model. From this single forward pass it reaches 68.74 mAP@100 on EuroSAT, 46.44 on BigEarthNet and 29.95 on METER-ML (Tab. 3). We evaluate EuroSAT with the corrected band order (App. A.13), which matters for an embedding that relies on band identity.

Table 3: Zero-shot text-to-image retrieval, mAP@100. Baselines: published values from MS-CLIP Marimo et al. [2025] (their Tab. 1 and protocol, BigEarthNet v1). SPEXT: corrected EuroSAT bands, BigEarthNet reBEN v2 test split Clasen et al. [2025], multi-label METER-ML. Avg.: mean over the three datasets. Best per column in bold.

Model	EuroSAT	BigEarthNet	METER-ML	Avg.
RemoteCLIP	34.34	26.62	20.08	27.01
Llama3-RGB-CLIP	52.72	28.84	25.60	35.72
OpenCLIP L/14	56.92	28.57	28.99	38.16
Llama3-MS-CLIP	63.03	35.62	29.72	42.79
SPEXT (ours)	**68.74**	**46.44**	**29.95**	**48.38**

4.3 VLM Grounding on RSVQAxBEN

In an EO-RAG pipeline, the VLM answers from the visual tokens of the retrieved scenes. We evaluate this answering stage on RSVQAxBEN Lobry et al. [2021], a third-party benchmark of presence (yes/no) and land-cover (LC) questions about BigEarthNet patches, with CORINE-derived answers and a geographic train/test split. In a pre-registered comparison (App. A.15), a frozen Qwen3-VL-8B-Instruct Bai et al. [2025b] answers each question from its own image, receiving the patch tokens of SPEXT, Llama3-MS-CLIP or, for the RGB baseline, Qwen3-VL's own vision tower through the same small MLP adapter.

Table 4: RSVQAxBEN test results (2.72 M questions). Qwen merger: Qwen3-VL's pretrained merger instead of the adapter. OA: overall accuracy. AA: mean of Pres. and LC exact. Pres. bal.: balanced presence accuracy. LC F1: class-set F1. Presence majority baseline: 63.32.

Visual tokens	LLM	OA	AA	Pres.	Pres. bal.	LC exact	LC F1
SPEXT	frozen	**81.09**	**62.46**	**91.47**	**90.97**	**33.45**	**68.25**
Llama3-MS-CLIP	frozen	77.03	55.98	88.76	88.52	23.19	59.50
RGB (adapter)	frozen	75.81	54.54	87.67	87.39	21.41	54.59
RGB (adapter)	LoRA	75.35	53.98	87.26	86.65	20.70	54.09
RGB (Qwen merger)	LoRA	74.61	53.15	86.57	85.65	19.74	52.42

SPEXT leads on every metric (Tab. 4), by $+5.3$ points of overall accuracy (OA) over RGB and $+4.1$ over Llama3-MS-CLIP, with the largest margins on spectrally defined classes ($+9.1$ presence accuracy over RGB on agriculture and $+3.8$ on forest and semi-natural areas, against $+0.2$ on artificial surfaces, Tab. 18). With a frozen language model, SPEXT also exceeds the RGB variants with LoRA Hu et al. [2022]. Because retrieval and answering use tokens from the same forward pass, a SPEXT-based EO-RAG pipeline needs a single visual encoder and never re-encodes retrieved scenes (Tab. 5).

4.4 Component Ablation

We isolate the contribution of each design element with a controlled ablation: each candidate is continually pre-trained from the same fMoW-S2 weights for 20 epochs on SSL4EO captioned data, then fine-tuned on the 10% MADOS split for 20 epochs across three random seeds (Tab. 7).

Decoupling. With ITC routed to the alignment token, adding it on top of MAE leaves dense prediction essentially unchanged ($+0.0$ F1, -0.22 mIoU on MADOS): the objective neither helps nor harms the dense stream. The full decoupled MAE+MTP+ITC reaches the top of the table, and routing its ITC to the alignment token matches routing it to pooled patch features on both PANGAEA splits (Route, and per task in Tab. 6), so the decoupled design adds retrieval at no dense-task cost. The complementary signal that does help dense prediction is the generative MTP objective ($+2.85$ F1 over MAE alone), which routes text supervision through a cross-attention decoder rather than into the patches themselves.

Architectural priors. The wavelength-parameterized embedding contributes incrementally on top of the index-agnostic baseline ($+1.95$ mIoU, $+2.87$ F1 with both RBF mixing and wavelength modulation), with the modulation step accounting for most of the gain. The same embedding transfers across sensors to Landsat with no architectural changes, recovering 98.8% of native Sentinel-2 mIoU on HLS Burn Scars after swapping in the new sensor's wavelengths and inheriting the visible and SWIR group convolutions (Tab. 14, App. A.11).

Bands and Stage I. SPEXT's multi-spectral bands carry much of its accuracy: with the same weights, RGB-only input lowers mIoU by up to 27.6 (Sen1Floods11) and 22.4 (MADOS), and pre-training on all bands rather than RGB alone adds up to 22.9 (Sen1Floods11). Stage I further improves most tasks (Tab. 8, per-seed detail in App. A.19).

Table 5: Per-query visual-encoder cost of the standard EO-RAG pipeline (CLIP-style retriever + VLM's native ViT) versus SPEXT-RAG (single encoder, cached tokens). Numbers are encoder forward passes and instantiated visual-encoder parameters per user query at top-$K{=}3$.

Stage	Standard EO-RAG	SPEXT-RAG
Index-time image encoding (amortized)	$1\times$ CLIP encoder	$1\times$ SPEXT
Query-time text encoding	$1\times$ sentence encoder	$1\times$ sentence encoder
Re-encoding of each top-K image	$1\times$ VLM-ViT	0 (cached)
Visual forward passes per query	$\sim$3 (one per top-K)	**0**
Visual-encoder parameters	$\sim$87 M + 300–500 M	$\sim$**87 M**

Table 6: Matched coupled/decoupled pair on PANGAEA (mIoU, mean ± std over seeds 42/234/1337, 20-epoch Stage II encoders from the same Stage I weights). Δ = decoupled − coupled, from unrounded means. *DICE loss in both arms. The 10% cells use cross-entropy.

	Full split			10% split		
Task	Decoupled	Coupled	Δ	Decoupled	Coupled	Δ
HLS Burn Scars	81.36 ± 0.16	81.61 ± 0.13	-0.25	77.69 ± 1.62	78.01 ± 1.16	-0.32
MADOS	67.08 ± 2.42	63.04 ± 1.68	$+4.03$	47.48 ± 2.39	48.93 ± 1.27	-1.45
PASTIS	38.56 ± 0.62	39.37 ± 0.84	-0.81	37.81 ± 0.68	38.83 ± 0.15	-1.02
Sen1Floods11	89.84 ± 0.56	90.13 ± 0.42	-0.29	85.42 ± 2.26	84.15 ± 2.51	$+1.27$
xView2	47.85 ± 0.32	47.67 ± 0.32	$+0.18$	36.84 ± 2.22	36.86 ± 2.31	-0.02
CropMap	55.57 ± 2.69	54.68 ± 4.39	$+0.89$	26.93 ± 8.89	28.14 ± 11.52	-1.21
SN7	58.59 ± 0.23	58.77 ± 0.11	-0.18	55.94 ± 0.24	55.79 ± 0.24	$+0.15$
AI4Farms	$35.34 \pm 4.49^*$	$32.78 \pm 9.70^*$	$+2.56$	34.89 ± 4.52	32.87 ± 9.23	$+2.03$
Average	59.27	58.51	$+0.77$	50.37	50.45	-0.07

5 Discussion

The results in §4.1 through §4.3 are unified by a single observation: the value of language supervision for multi-spectral encoders depends on *how* that supervision is applied. SPEXT's patch tokens, which receive it through latent text prediction, give a VLM higher RSVQAxBEN accuracy than Llama3-MS-CLIP or RGB tokens, with the largest margins on spectrally defined classes (§4.3). When the same supervision is routed to a dedicated alignment token, the dense features remain spectrally faithful enough to lead PANGAEA on average, while the alignment token simultaneously supports zero-shot retrieval.

6 Conclusion & Limitations

We argued that the visual encoder, not the language model, is the principal bottleneck for VLMs in Earth Observation. SPEXT operationalizes this view through a decoupled architecture that routes generative latent text prediction to the dense patch stream and image-text alignment to a dedicated alignment token. Across segmentation, retrieval, and downstream VLM grounding, the resulting encoder serves all of these roles from a single forward pass. For EO retrieval-augmented generation, the application that motivated SPEXT, a single encoder therefore serves as both the retriever and the VLM's visual front end (App. A.16). The implication is broader than EO. Any multi-modal pre-training setting where dense physical or structural input must coexist with text-driven downstream interaction is a candidate for the same supervision-decoupling principle. Notably, SPEXT achieves this with $3.5\times$ fewer parameters than CROMA and $4.7\times$ less pre-training data than DOFA.

Limitations. SPEXT inherits three constraints common to current EO foundation models. (i) Caption supervision is automatically generated. MTP targets are frozen Qwen3-Embedding outputs over

Table 7: Component ablation on MADOS, mean ± std over 3 seeds. **Top:** architectural components (RBF mixing, wavelength modulation). **Middle:** training objectives. ITC is routed to the alignment token in all MADOS rows, so the MAE+ITC row measures its effect on dense prediction once already decoupled. **Bottom:** † full SPEXT = wavelength priors + MAE+MTP+ITC. **Route:** PANGAEA protocol, 8-task average mIoU on the full and 10% splits.

Type	Configuration	mIoU ↑	F1 ↑
Arch.	Baseline	32.85 ± 2.34	42.94 ± 2.10
	+ RBF	33.00 ± 1.76	43.11 ± 1.21
	+ Wavelength Modulation	34.80 ± 4.83	45.81 ± 5.10
Obj.	MAE	30.70 ± 4.55	41.01 ± 4.63
	MAE + ITC (align. token)	30.48 ± 3.87	41.02 ± 3.86
	MAE + MTP	32.80 ± 1.85	43.86 ± 1.66
Full	SPEXT†	**35.23** ± 4.36	**45.94** ± 4.36

Masking Ratio	mIoU ↑	F1 ↑
90% (SPEXT)	**35.23** ± 4.36	**45.94** ± 4.36
75%	32.80 ± 1.85	43.13 ± 1.58
60%	33.77 ± 2.30	44.18 ± 2.14
Route: ITC on	Full	10%
align. token	59.27	50.37
pooled patches	58.51	50.45

Table 8: Spectral bands and Stage I (PANGAEA full split, mIoU, 20-epoch encoders, seed 42).

	HLS	MADOS	Sen1Fl	CropMap	SN7	AI4Farms
(a) Stage I init., all bands	81.47	66.11	90.38	54.63	58.33	37.89
(b) as (a), RGB input	71.55	43.71	62.77	41.91	58.08	36.60
(c) no Stage I, all bands	79.79	57.93	90.42	37.99	57.11	30.79
(d) no Stage I, RGB only	69.46	49.88	67.57	45.71	57.90	33.41
Bands at task time (a−b)	+9.92	+22.40	+27.61	+12.72	+0.25	+1.29
Stage I (a−c)	+1.68	+8.18	−0.04	+16.64	+1.22	+7.10
Bands at pre-training (c−d)	+10.33	+8.05	+22.85	−7.72	−0.79	−2.62

machine-derived SSL4EO-S12 captions, augmented with a metadata-derived geographic anchor sentence (§3.2); none of the supervision is human-curated. (ii) The spectral prior is parameterized for optical sensors and transfers to Landsat without architectural changes (App. A.11). The wavelength-parameterized embedding is built around continuous reflectance bands and would require adaptation for non-optical modalities such as SAR. (iii) SPEXT does not lead on every individual task in every setting. On the low-data PANGAEA split it regresses on AI4Farms (App. A.10). Natural extensions include region-level text grounding (App. A.20) and multitemporal imagery, which needs a captioned multitemporal corpus rather than a new architecture (App. A.22).

Acknowledgments and Disclosure of Funding

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC), under the project Perfect Recollection for Clearer Insight (grant number EP/Y036077/1). The authors acknowledge the use of resources provided by the Isambard-AI National AI Research Resource (AIRR) McIntosh-Smith et al. [2024]. Isambard-AI is operated by the University of Bristol and is funded by the UK Government's Department for Business, Innovation, Science and Trade (DBIST) via UK Research and Innovation; and the Science and Technology Facilities Council [ST/AIRR/I-A-I/1023]. The authors declare no competing interests.

References

Pierre Adorni, Minh-Tan Pham, Stéphane May, and Sébastien Lefèvre. Towards efficient benchmarking of foundation models in remote sensing: A capabilities encoding approach. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3096–3106, 2025.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023.

Guillaume Astruc, Nicolas Gonthier, Clément Mallet, and Loic Landrieu. AnySat: One Earth observation model for many resolutions, scales, and modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19530–19540, 2025. doi: 10.1109/CVPR52734.2025.01819.

Lubin Bai, Xiuyuan Zhang, Siqi Zhang, Zepeng Zhang, Haoyu Wang, Wei Qin, and Shihong Du. GeoLink: Empowering remote sensing foundation model with OpenStreetMap data. In *Advances in Neural Information Processing Systems*, volume 38, 2025a. doi: 10.52202/085713-5617.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. arXiv preprint arXiv:2511.21631, 2025b.

Silin Cheng and Kai Han. VaMP: Variational multi-modal prompt learning for vision-language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2994.

Kai Norman Clasen, Leonard Hackel, Tom Burgert, Gencer Sumbul, Begüm Demir, and Volker Markl. reBEN: Refined BigEarthNet dataset for remote sensing image analysis. In *IGARSS 2025 - 2025 IEEE International Geoscience and Remote Sensing Symposium*, pages 1264–1268. IEEE, 2025. doi: 10.1109/IGARSS55030.2025.11242834.

Jan GPW Clevers and Anatoly A Gitelson. Remote estimation of crop and grass chlorophyll and nitrogen content using red-edge bands on sentinel-2 and-3. *International Journal of Applied Earth Observation and Geoinformation*, 23:344–351, 2013.

Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 197–211. Curran Associates, Inc., 2022.

Mohammad Hossain Dehghan-Shoar, Gabor Kereszturi, Reddy R. Pullanagari, Alvaro A. Orsi, Ian J. Yule, and James Hanly. A physically informed multi-scale deep neural network for estimating foliar nitrogen concentration in vegetation. *International Journal of Applied Earth Observation and Geoinformation*, 130:103917, 2024. ISSN 1569-8432.

Aayush Dhakal, Srikumar Sastry, Subash Khanal, Adeel Ahmad, Eric Xing, and Nathan Jacobs. Range: Retrieval augmented neural fields for multi-resolution geo-embeddings. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24680–24689, 2025.

Zhengpeng Feng, Clement Atzberger, Sadiq Jaffer, Jovana Knezevic, Silja Sormunen, Robin Young, Madeline C. Lisaius, Markus Immitzer, Toby Jackson, James Ball, David A. Coomes, Anil Madhavapeddy, Andrew Blake, and Srinivasan Keshav. TESSERA: Temporal embeddings of surface spectra for Earth representation and analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 34818–34831, 2026.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. CLIP-Adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024. doi: 10.1007/s11263-023-01891-x.

Alexander FH Goetz, Gregg Vane, Jerry E Solomon, and Barrett N Rock. Imaging spectrometry for earth remote sensing. *science*, 228(4704):1147–1153, 1985.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, Antonio Plaza, Paolo Gamba, Jon Atli Benediktsson, and Jocelyn Chanussot. Spectralgpt: Spectral remote sensing foundation model. *IEEE Trans. Pattern Anal.*

Mach. Intell., 46(8):5227–5244, August 2024. ISSN 0162-8828. doi: 10.1109/TPAMI.2024.3362475.

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.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36: 72096–72109, 2023.

Johannes Jakubik, Sujit Roy, CE Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, Carlos Gomes, Gabby Nyirjesy, Blair Edwards, et al. Foundation models for generalist geospatial artificial intelligence. *arXiv preprint arXiv:2310.18660*, 2023.

Xuyang Li, Chenyu Li, Pedram Ghamisi, and Danfeng Hong. Fleximo: A flexible remote sensing foundation model. *arXiv preprint arXiv:2503.23844*, 2025.

Thomas Lillesand, Ralph W Kiefer, and Jonathan Chipman. *Remote sensing and image interpretation*. John Wiley & Sons, 2015.

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

Zeping Liu, Ni Lao, Zhangyu Wang, Junfeng Jiao, and Gengchen Mai. GAIR: Location-aware self-supervised contrastive pre-training with geo-aligned implicit representations. *ISPRS Journal of Photogrammetry and Remote Sensing*, 237:166–182, 2026. doi: 10.1016/j.isprsjprs.2026.04.035.

Sylvain Lobry, Begüm Demir, and Devis Tuia. RSVQA meets BigEarthNet: A new, large-scale, visual question answering dataset for remote sensing. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 1218–1221. IEEE, 2021. doi: 10.1109/IGARSS47720.2021.9553307.

Gengchen Mai, Weiming Huang, Jin Sun, Suhang Song, Deepak Mishra, Ninghao Liu, Song Gao, Tianming Liu, Gao Cong, Yingjie Hu, Chris Cundy, Ziyuan Li, Rui Zhu, and Ni Lao. On the opportunities and challenges of foundation models for geoai (vision paper). *ACM Trans. Spatial Algorithms Syst.*, 10(2), July 2024. ISSN 2374-0353. doi: 10.1145/3653070.

Utkarsh Mall, Bharath Hariharan, and Kavita Bala. Change-aware sampling and contrastive learning for satellite images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5261–5270, 2023.

Utkarsh Mall, Cheng Perng Phoo, Meilin Kelsey Liu, Carl Vondrick, Bharath Hariharan, and Kavita Bala. Remote sensing vision-language foundation models without annotations via ground remote alignment. In *The Twelfth International Conference on Learning Representations*, 2024.

Clive Tinashe Marimo, Benedikt Blumenstiel, Maximilian Nitsche, Johannes Jakubik, and Thomas Brunschwiler. Beyond the visible: Multispectral vision-language learning for earth observation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 359–375. Springer, 2025.

Valerio Marsocci, Yuru Jia, Georges Le Bellier, David Kerekes, Liang Zeng, Sebastian Hafner, Sebastian Gerard, Eric Brune, Ritu Yadav, Ali Shibli, et al. Pangaea: A global and inclusive benchmark for geospatial foundation models. *arXiv preprint arXiv:2412.04204*, 2024.

Simon McIntosh-Smith, Sadaf Alam, and Christopher Woods. Isambard-AI: a leadership-class supercomputer optimised specifically for artificial intelligence. In *Proceedings of the Cray User Group*, pages 44–54. 2024.

Elizbar A Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1): 141–142, 1964.

Mubashir Noman, Muzammal Naseer, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Fahad Shahbaz Khan. Rethinking transformers pre-training for multi-spectral satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27811–27819, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4088–4099, 2023.

Martin Sudmanns, Dirk Tiede, Stefan Lang, Helena Bergstedt, Georg Trost, Hannah Augustin, Andrea Baraldi, and Thomas Blaschke. Big earth data: disruptive changes in earth observation data management and analysis? *International Journal of Digital Earth*, 13(7):832–850, 2020.

Gencer Sumbul, Marcela Charfuelan, Begüm Demir, and Volker Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019-2019 IEEE international geoscience and remote sensing symposium*, pages 5901–5904. IEEE, 2019.

Gencer Sumbul, Chang Xu, Emanuele Dalsasso, and Devis Tuia. SMARTIES: Spectrum-aware multi-sensor auto-encoder for remote sensing images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5569–5578, 2025. doi: 10.1109/ICCV51701.2025.00528.

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

Chong Tang, Joseph Powell, Dirk Koch, Robert D Mullins, Alex S Weddell, and Jagmohan Chauhan. Physwin: An efficient and physically-informed foundation model for multispectral earth observation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Maofeng Tang, Andrei Cozma, Konstantinos Georgiou, and Hairong Qi. Cross-Scale MAE: A tale of multiscale exploitation in remote sensing. In *Advances in Neural Information Processing Systems*, volume 36, pages 20054–20066, 2023. doi: 10.52202/075280-0879.

Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *Advances in Neural Information Processing Systems*, 36:8690–8701, 2023.

Yi Wang, Nassim Ait Ali Braham, Zhitong Xiong, Chenying Liu, Conrad M Albrecht, and Xiao Xiang Zhu. Ssl4eo-s12: A large-scale multimodal, multitemporal dataset for self-supervised learning in earth observation [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 11(3): 98–106, 2023.

Geoffrey S Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372, 1964.

Congcong Wen, Yiting Lin, Xiaokang Qu, Nan Li, Yong Liao, Hui Lin, and Xiang Li. Rs-rag: Bridging remote sensing imagery and comprehensive knowledge with a multi-modal dataset and retrieval-augmented generation model. *arXiv preprint arXiv:2504.04988*, 2025.

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*, 2024.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856.

Zhenghang Yuan, Zhitong Xiong, Lichao Mou, and Xiao Xiang Zhu. Chatearthnet: A global-scale image-text dataset empowering vision-language geo-foundation models. *Earth System Science Data Discussions*, 2024:1–24, 2024.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025a.

Yingying Zhang, Lixiang Ru, Kang Wu, Lei Yu, Lei Liang, Yansheng Li, and Jingdong Chen. SkySense V2: A unified foundation model for multi-modal remote sensing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9136–9146, 2025b. doi: 10.1109/ICCV51701.2025.00854.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. doi: 10.1007/s11263-022-01653-1.

Bryan Zhu, Nicholas Lui, Jeremy A. Irvin, Jimmy Le, Sahil Tadwalkar, Chenghao Wang, Zutao Ouyang, Frankie Y. Liu, Andrew Y. Ng, and Robert B. Jackson. Meter-ml: A multi-sensor earth observation benchmark for automated methane source mapping. In *CDCEO@IJCAI*, 2022.

A Appendix

A.1 Pre-training Hyperparameters and Schedule

Backbone and inputs. ViT-Base, 128×128 input crops, patch size 8, asymmetric masking (90% on the visual stream, 50% on the text stream) in Stage II, and 96×96 crops with 75% visual masking in Stage I.

Optimization. AdamW with learning rate 1×10^{-4} and cosine decay; global batch size 2048; channel-wise z-standardization per band with statistics of the L2A values of the SSL4EO-S12 v1.1 Zarr release, which include the $+1000$ offset (offset-free inputs need $+1000$ added first).

Hardware. 4 NVIDIA GH200 GPUs.

Stage I. 100 epochs of pure MAE on fMoW-S2 with spectral-group band dropout: per patch, the Red-Edge/NIR and SWIR groups are dropped with probability 0.3 and the Visible group with probability 0.03, and the full image is reconstructed. Its pixel decoder receives the visible tokens in permuted order, a setting the released code reproduces as an option.

Stage II warm-up. The first 4 epochs of Stage II are a decoder warm-up: the SPEXT encoder is held frozen so the vision and vision-language decoders adapt to the pre-trained spectral features without disturbing them. After epoch 4 the encoder is unfrozen and trained jointly with the decoders for the remaining 46 epochs. The transient at epoch 5 visible in Fig. 4 is this scheduled unfreezing.

Frozen sentence encoder. Qwen3-Embedding-0.6B Zhang et al. [2025a]; sentence embeddings are pre-computed once and cached.

A.2 Loss Formulations

The Stage I MAE pixel-reconstruction loss is the standard mean-squared error between the decoded pixels and the masked target,

$$\mathcal{L}_{\text{MAE}} = \|\text{VisionDecoder}(X_{\text{masked}}) - X_{\text{target}}\|_2^2. \tag{4}$$

The Latent Masked Text Prediction (MTP) loss is a cosine distance plus a weighted mean-squared error in the frozen sentence-encoder embedding space, computed only over masked positions $\mathcal{M}$:

$$\mathcal{L}_{MTP} = \frac{1}{|\mathcal{M}|} \sum_{i\in\mathcal{M}} \Big[\big(1 - \cos(\hat{\mathbf{z}}_{t,i}, \mathbf{z}_{t,i})\big) + \tfrac{\mu}{d} \|\hat{\mathbf{z}}_{t,i} - \mathbf{z}_{t,i}\|_2^2\Big]. \tag{5}$$

Here $\mu = 0.25$ weights the squared-error term and $d = 1024$ is the dimension of the sentence embedding.

The ITC loss is a cosine loss on matched pairs between the projected alignment token $\mathbf{v}_{\text{cls}}$ and the holistic sentence embedding $\mathbf{z}_t$ produced by the same frozen $\mathcal{E}_{\text{txt}}$,

$$\mathcal{L}_{\text{ITC}} = \frac{1}{N} \sum_{i=1}^{N} \big(1 - \cos(\text{proj}(\mathbf{v}_{\text{cls},i}), \mathbf{z}_{t,i})\big), \tag{6}$$

with a learned projection head proj.

A.3 Training Dynamics

We analyze the training dynamics over the first 50 epochs to demonstrate the efficiency of our multi-objective strategy. Fig. 4 illustrates the rapid convergence of the Total Loss and its components: Masked Autoencoding (MAE), Masked Text Prediction (MTP), and Image-Text Alignment (ITC) loss.

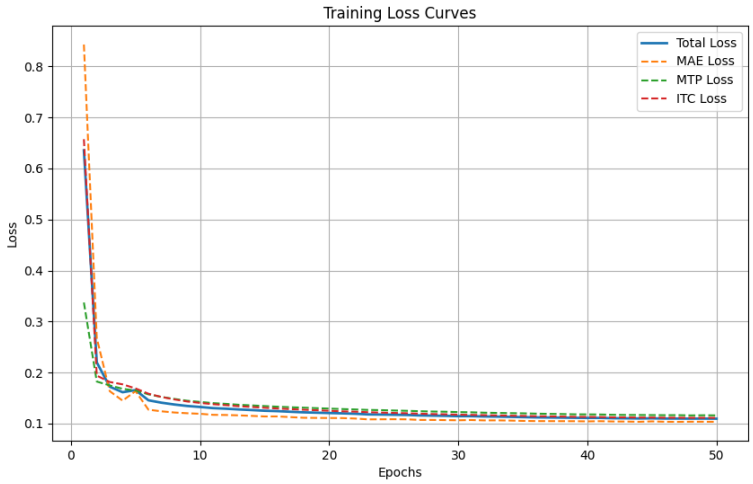


Figure 4: **Pre-training Loss Decomposition.** The total loss (solid blue) and constituent losses (dashed) converge rapidly within 50 epochs. The parallel reduction in generative (MAE/MTP) and alignment (ITC) losses indicates stable optimization without task conflict.

As shown in Fig. 4, the model exhibits a sharp decrease in loss during the initial warmup phase. The transient oscillation visible at epoch 5 corresponds to the **scheduled unfreezing of the encoder**, after which the loss immediately restabilizes and continues to converge.

A.4 SSL4EO Caption-Length Distribution

Fig. 5 shows the cumulative distribution of sentence counts per caption in the SSL4EO-S12 corpus. Over 60% of raw captions contain at most two sentences, and the long tail extending past five sentences is dominated by complex urban or coastal scenes. Without enrichment, MTP would frequently see input sequences of length 1, leaving no positions to mask. The masking objective then degenerates into either a no-op when the single position is unmasked, or a memorization task when it is masked, since the decoder has no other linguistic context to condition on. Adding a metadata-derived geographic anchor sentence (§3.2) guarantees a minimum of two positions per sample, restoring a non-trivial reconstruction problem.

Geographic anchor construction. For each sample we map the centroid coordinates to administrative boundaries, e.g. country and province, and physical landmarks, e.g. nearest populated place or water body, using open-source Natural Earth layers, similar to coordinate-based grounding Vivanco Cepeda et al. [2023], Dhakal et al. [2025]. We infer a season label from the acquisition timestamp and a

latitude-dependent climate band. The attributes are synthesized into the template *"Patch of [Type] during [Season] near [Location]"*. The construction is purely structural and does not introduce downstream-evaluation leakage, since the constructed sentence carries only what is already encoded in the image's geographic metadata. The release includes the resulting anchor sentences.

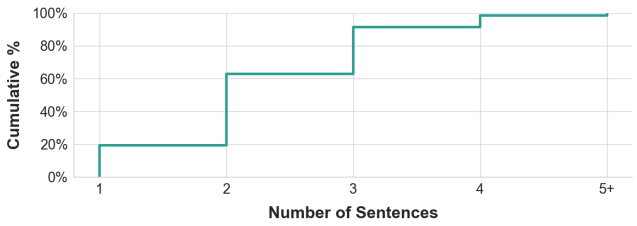


Figure 5: Cumulative distribution of sentence counts per caption in SSL4EO-S12. Over 60% of raw captions contain ≤ 2 sentences, motivating the metadata enrichment described in §3.2.

A.5 SSL4EO-S2 Samples with Generated Metadata

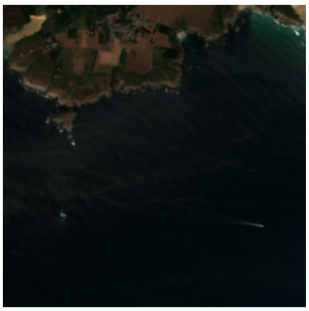
In this section, we present qualitative examples from the SSL4EO-S2 dataset to demonstrate the alignment between visual data and textual descriptions. Tab. 9 displays representative samples, detailing the satellite imagery alongside their generated captions and geolocation-based context.

To explicitly ground the images in their spatio-temporal reality, we constructed natural language descriptions using a custom reverse geocoding pipeline. For each sample, we mapped the centroid coordinates to administrative boundaries (e.g., country, province) and physical landmarks (e.g., nearest populated place or water body) using open-source geospatial data (Natural Earth). We then inferred the season based on the acquisition timestamp and latitude-dependent climate models (tropical vs. temperate). These attributes were synthesized into a structured natural language template:

*"Patch of **[Type]** during **[Season]** near **[Location]**."*

This process ensures consistent, physically grounded metadata for every image in the dataset.

Table 9: Samples from the SSL4EO-S2 dataset including generated captions, reverse-geocoded addresses, and associated latitude/longitude position.

Satellite Image	Generated Caption & Metadata
	**Caption:** The image shows a satellite view of a landscape with farmland areas, a waterway, and a ditch. The farmland is divided into sections with different crops being grown, and the waterway runs through some of the fields. The ditch serves as a drainage system for the farmland. **Address:** Patch of land during dry season near Chiclayo, Lambayeque, Peru. **Coordinates:** -6.701195124457226, -79.83503845676178
	**Caption:** The image shows a large bay surrounded by land on three sides, with residential, agricultural, and natural areas near the water's edge. The bay is connected to the ocean and is located near the coast. **Address:** Patch of water during summer in Bay of Biscay near France. **Coordinates:** 47.28992533366016, -3.185417033419533
	**Caption:** The image is a satellite view of a rural area with various land uses, including agriculture, residential areas, military bases, and training areas. The largest agricultural field is rectangular and covers 1.5 square kilometers, while the three residential areas have an average size of 0.5 square kilometers each. The military bases and training areas cover a total of 0.2 square kilometers. **Address:** Patch of land during autumn near Manisa, Manisa, Turkey. **Coordinates:** 38.62747919969373, 27.461719558481143

A.6 Physics-Informed Spectral Embedding

We provide the pseudo-code to illustrate the forward pass of our Physics-Informed Spectral Embedding layer of SPEXT. Unlike standard patch embedding, which treats channels as arbitrary dimensions, Algorithm 1 explicitly grounds features in physical reality. It proceeds in four stages: (1) grouping bands by spectral region for efficient spatial encoding, (2) applying a wavelength-dependent RBF kernel to enforce spectral continuity, (3) modulating features via Fourier-encoded wavelength metadata, and (4) aggregating the result into a unified token representation.

Algorithm 1 Physics-Informed Spectral Embedding Layer

Input: Image Batch $X \in \mathbb{R}^{B\times 10\times H\times W}$
Constants: Wavelengths $\Lambda \in \mathbb{R}^{10}$, RBF $\sigma = 80.0$, Fourier Dim $F = 128$
Learnable Params: Group Convs $\mathcal{C}_{vis}, \mathcal{C}_{red}, \mathcal{C}_{swir}$, Modulation Heads $\psi_{bias}, \psi_{scale}$
{*1. Grouped Convolutions (Embed Spatial Features)*}
$Y_{vis} \leftarrow \mathcal{C}_{vis}(X_{[:3]}) \in \mathbb{R}^{3D\times H'\times W'}$
$Y_{red} \leftarrow \mathcal{C}_{red}(X_{[3:8]}) \in \mathbb{R}^{5D\times H'\times W'}$
$Y_{swir} \leftarrow \mathcal{C}_{swir}(X_{[8:]}) \in \mathbb{R}^{2D\times H'\times W'}$
$Z \leftarrow \text{Concat}(Y_{vis}, Y_{red}, Y_{swir})$
$Z \leftarrow \text{Reshape}(Z, [B, N_{patches}, 10, D])$ ▷ Align per band
{*2. Physics-Based Spectral Mixing*}
Compute Kernel $K_{ij} = \exp(-||\lambda_i - \lambda_j||^2/2\sigma^2)$
Normalize K: $K_{ij} \leftarrow K_{ij}/(\sum_c K_{ic} + \epsilon)$
$Z_{mixed} \leftarrow \text{Einsum}(\text{"}bnid, ci \to bncd\text{"}, Z, K)$ ▷ Mix bands by proximity
{*3. Wavelength-Conditioned Modulation*}
Fourier Feats: $\phi(\Lambda) = [\sin(2\pi\bar{\Lambda}F), \cos(2\pi\bar{\Lambda}F)]$
$\beta \leftarrow \psi_{bias}(\phi(\Lambda)), \quad \alpha \leftarrow 1 + \tanh(\psi_{scale}(\phi(\Lambda)))$
$Z_{final} \leftarrow Z_{mixed} \odot \alpha + \beta$
{*4. Spectral Aggregation*}
$Z_{out} \leftarrow \sum_{c=1}^{10} Z_{final}[:, :, c, :]$ ▷ Sum across spectral dim
return Z_{out}

A.6.1 Comparison with Other Spectral Stems and Design Insights

DOFA Xiong et al. [2024] *predicts* its patch-projection weights from band wavelengths with a hypernetwork. SMARTIES Sumbul et al. [2025] *enumerates* learned per-band projections tied to fixed wavelength intervals, a lookup table rather than a continuous function of wavelength, so an unseen sensor needs new projections or interpolation. PhySwin Tang et al. [2025] *soft-constrains* spectral structure through a smoothness loss. SPEXT *computes* inter-band mixing from wavelength distances with a fixed RBF kernel and learns the wavelength-conditioned modulation applied after its grouped convolutions. Counted analytically for both patch embeddings in one shared configuration (224×224 input, 16×16 patches, width 1024, 10 bands), SPEXT's spectral mechanism (RBF mixing and modulation) has 526K parameters and needs 0.057 GFLOPs per sample, against about 34M parameters and 0.731 GFLOPs for DOFA's hypernetwork. The whole patch embeddings have 10.5M vs. 34.2M parameters and need 3.96 vs. 1.76 GFLOPs, most of SPEXT's in its grouped convolutions. Three design insights follow.

- **Compute, rather than predict, physics-determined structure.** Within the optical range, inter-band correlation is governed by wavelength proximity, a smooth and known prior, and learning it with a hypernetwork spends capacity on flexibility the physics does not require.
- **Put learning where physics ends.** SPEXT's learnable spectral capacity sits in the modulation, which governs how features respond given a spectral position, not how positions relate to one another. A sensor is described to the stem by its band wavelengths.
- **Structural priors remain available at inference, unlike loss-based priors.** A smoothness penalty shapes the weights only during training, whereas an analytic kernel is recomputed for any list of band wavelengths (App. A.11).

A.7 Sensitivity Analysis

The masking-ratio sweep is reported with the component ablation in Tab. 7. The runs of Tabs. 10, 11 and 12 continue Stage II pre-training from the Stage I weights for 20 epochs, rather than the 50 epochs of the released model.

Tab. 10 investigates the sensitivity of the model to the weighting of our training objectives (α, β, γ). We compare three configurations: (1) an **MAE-dominated** training ($0.8/0.1/0.1$) that prioritizes pixel reconstruction; (2) an **ITC/MTP-dominated** training ($0.2/0.4/0.4$) focusing on high-level semantic alignment; and (3) a **Balanced** approach ($0.4/0.3/0.3$). The results indicate that the balanced distribution maximizes performance, outperforming the MAE-dominated configuration by over 1.3 mIoU points. Pixel reconstruction alone, or alignment alone, both underperform the balanced setting on MADOS.

Table 10: Ablation of loss coefficients ($\alpha/\beta/\gamma$) on the MADOS dataset. We compare reconstruction-heavy (MAE), alignment-heavy (ITC/MTP), and balanced configurations. The balanced setting yields the highest mIoU on MADOS.

		MADOS	
Strategy	Config ($\alpha/\beta/\gamma$)	mIoU ↑	F1 ↑
MAE-Dominated	0.8 / 0.1 / 0.1	33.86 ± 5.03	44.13 ± 5.80
ITC/MTP-Dominated	0.2 / 0.4 / 0.4	35.16 ± 5.51	44.39 ± 5.23
Balanced (Ours)	0.4 / 0.3 / 0.3	**35.23** ± 4.36	**45.94** ± 4.36

A.8 RBF Kernel Bandwidth Sensitivity

The RBF kernel is parameterized by a single hyperparameter σ that controls the wavelength interaction range. Tab. 11 sweeps $\sigma \in \{40, 80, 120\}$ nm across the PANGAEA tasks. Performance is stable across the tested range. An additional run extends the sweep to $\sigma = 160$ nm: averaged over the six tasks of Tab. 11 other than xView2, $\sigma = 40/80/120/160$ nm give 64.58/65.40/64.48/66.24 mIoU, and over seven tasks of the 10% split 53.80/51.20/51.24/51.86, so performance is stable across this range. The kernel does not require per-sensor re-tuning: it only needs the central wavelengths of the new sensor's bands as input, from which the kernel matrix is recomputed analytically (see App. A.11).

Table 11: Sensitivity to the RBF bandwidth σ (in nm) on PANGAEA (mIoU). Performance is stable across the tested range.

σ (nm)	HLS	MADOS	PASTIS	Sen1Fl	xView2	CropMap	SN7	Avg
40	81.07	66.33	37.73	89.86	47.45	53.89	58.63	62.14
80 (Ours)	**81.41**	**70.62**	37.81	89.86	**47.57**	54.05	58.64	**62.85**
120	81.17	64.22	**37.92**	**90.08**	47.25	**54.60**	**58.87**	62.02

A.9 Robustness to Caption Noise

Automated captions are inherently noisy. To verify that the MTP objective does not propagate this noise into the spectral representations, we re-train SPEXT under two corruption settings at three severity levels: **swap** (replacing the caption with one drawn from a different scene) and **dropout** (removing the caption entirely). We then fine-tune on every PANGAEA task and report mIoU (Tab. 12).

MTP gradients reach the spectral patch embeddings only through the key and value projections of a separate cross-attention decoder; the encoder only needs to remain *sufficiently informative*, not aligned with the text feature space itself.

Table 12: PANGAEA mIoU under caption corruption. Spectrally-driven tasks (Sen1Floods11, HLS, SpaceNet7) remain stable (≤ 1.16 mIoU drift) from 25% to 75% corruption. Semantic tasks (CropMap, PASTIS) degrade more under swap noise but the overall average drops only ~ 2 mIoU from 25% to 75%.

Config	Sen1Fl	HLS	SN7	MADOS	CropMap	PASTIS	xView2	AI4Farms	Avg
Swap 25%	83.86	76.23	54.77	45.35	41.06	38.16	38.68	30.62	51.09
Swap 50%	83.33	76.08	55.13	52.81	28.32	36.80	39.45	27.81	49.97
Swap 75%	82.82	77.01	55.59	51.88	23.79	35.96	40.05	25.29	49.05
Dropout 25%	83.28	76.80	54.72	50.93	31.48	39.24	39.05	31.68	50.90
Dropout 50%	83.61	76.61	54.75	47.76	27.89	39.17	38.91	29.62	49.79
Dropout 75%	82.96	75.64	54.99	50.34	29.93	37.66	39.68	28.22	49.93

Table 13: 10% PANGAEA split. SPEXT achieves the best average mIoU and Avg DtB despite using $3.5\times$ fewer parameters than the strongest baseline. CROMA edges SPEXT on Average Rank in this setting, and SPEXT regresses on AI4Farms.

Model (Parameters)	HLS	MADOS	PASTIS	Sen1Floods11	XView2	CropMap	SN7	AI4Farms	Average ↑	Avg Rank ↓	Avg DtB ↓
S12-MAE (22M)	76.60	18.44	31.06	84.81	39.84	35.29	40.51	23.60	43.77	8.88	11.40
S12-Data2Vec (22M)	74.38	17.86	33.09	81.91	41.60	34.11	40.66	22.85	43.31	9.63	11.86
S12-MoCo (22M)	73.11	19.47	32.51	79.58	41.15	36.54	49.46	37.97	46.22	7.63	8.94
S12-DINO (22M)	75.93	23.47	36.62	84.95	41.02	38.44	41.15	37.91	47.44	5.63	7.73
DOFA-CLIP (457M)	65.63	38.07	21.67	75.63	**56.85**	13.54	49.55	25.63	43.32	8.88	11.84
MS-CLIP (87M)	71.39	41.99	24.60	80.77	54.82	20.16	49.60	21.00	45.54	8.38	9.62
RemoteCLIP (87M)	69.40	20.57	17.19	62.22	53.75	19.86	43.11	23.85	38.74	11.25	16.42
GFM-Swin (87M)	67.23	28.19	21.47	62.57	53.45	27.21	39.48	32.88	41.56	10.38	13.61
SatlasNet (87M)	74.79	29.87	16.76	83.92	44.07	29.08	49.78	13.91	42.77	8.38	12.39
Prithvi (87M)	77.73	21.24	33.56	86.28	35.08	27.71	36.78	35.04	44.18	7.88	10.99
Scale-MAE (303M)	75.47	21.47	22.86	64.74	56.06	13.44	49.68	26.66	41.30	8.38	13.87
SpectralGPT (105M)	**83.35**	20.29	34.53	83.12	35.81	31.06	36.31	37.35	45.23	7.88	9.94
DOFA (112M)	71.98	23.77	27.68	82.84	55.60	29.91	46.10	27.74	45.70	7.50	9.46
CROMA (303M)	76.44	32.44	32.80	**87.22**	46.54	36.77	42.15	**38.48**	49.11	**4.38**	6.06
SPEXT (87M)	78.96	**44.23**	**40.05**	84.51	37.62	**40.98**	**56.14**	16.87	**49.92**	4.75	**5.99**

A.10 PANGAEA Per-Task Results: 10% Split

The full per-dataset PANGAEA results under the 10% data split, summarized in the main text (§4.1), are reported in Tab. 13.

A.11 Cross-Sensor Transfer to Landsat

Because the physics-informed layers are parameterized by wavelength rather than band index, transferring SPEXT to a different optical sensor requires no architectural changes. We demonstrate this by evaluating on the **HLS Burn Scars** task using Landsat inputs (6 bands, central wavelengths $\{490, 560, 665, 865, 1610, 2190\}$ nm). The adaptation procedure is:

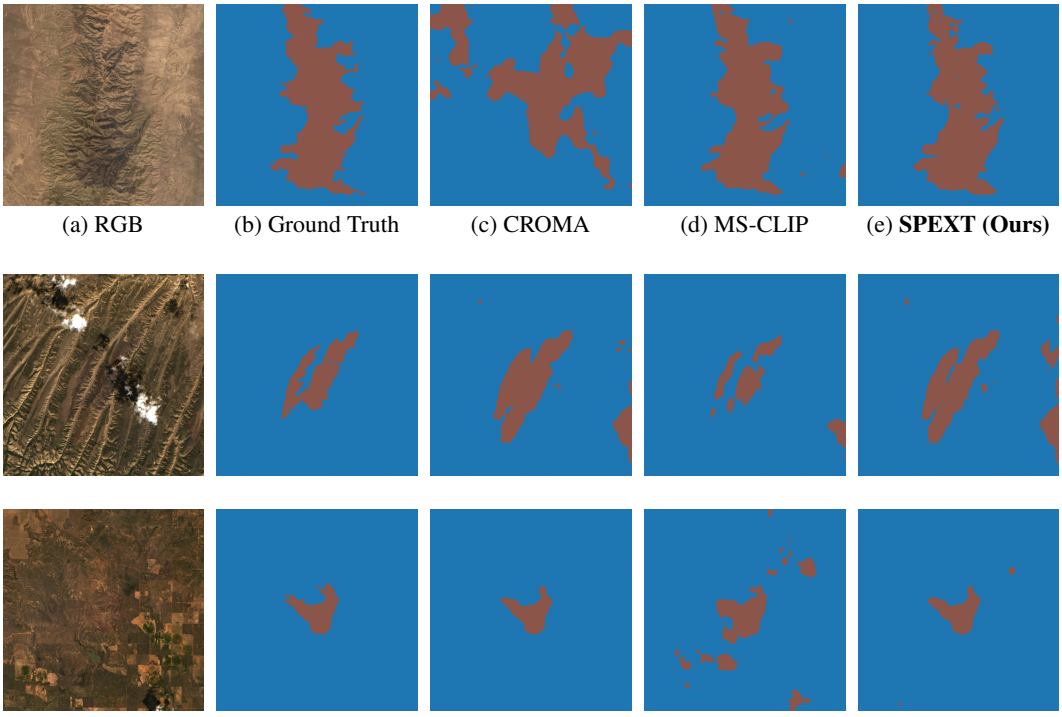
- The **RBF kernel** is recomputed analytically from the 6 Landsat central wavelengths.
- The **Fourier features** feeding the wavelength modulation are recomputed from the same wavelengths.
- The **wavelength modulation weights** transfer directly: they learned "how to modulate given a spectral position" during S2 pre-training and now receive Landsat positions.
- The **Visible** (3 bands) and **SWIR** (2 bands) group convolutions transfer directly (identical bands); only the NIR group (1 band) is randomly initialized.

Table 14: Cross-sensor transfer on HLS Burn Scars: SPEXT pre-trained on Sentinel-2 then fine-tuned on Landsat reaches 98.8% of its native Sentinel-2 performance (mIoU), with only the NIR group convolution randomly initialized.

Metric	Landsat	Sentinel-2
Mean IoU	80.67	81.65
Mean F1	88.51	89.19
Mean Precision	88.03	88.42
Mean Recall	89.01	90.00

A.12 Qualitative Figures on Segmentation Tasks

We compare against CROMA (the strongest masked-image RSFM in our PANGAEA suite) and MS-CLIP (the strongest spectral-CLIP encoder). As illustrated in Fig. 6, SPEXT produces sharper segmentation masks across varying conditions. In the top row, SPEXT recovers fine-grained boundaries that CROMA over-smooths. The middle row shows the same pattern in low-contrast settings, where SPEXT retrieves secondary artifacts missed by both baselines. The bottom row contrasts the two designs in a complex dark scene with high spectral contrast: MS-CLIP produces extensive over-segmentation across the background, while SPEXT suppresses the same false positives and isolates the burn scar with geometric fidelity close to the ground truth.



(a) RGB (b) Ground Truth (c) CROMA (d) MS-CLIP (e) **SPEXT (Ours)**

Figure 6: **Qualitative Segmentation Results on the HLS Burn Scar Scenes Dataset.** Visual comparison of segmentation predictions against baselines (CROMA, MS-CLIP). **Top Row:** A complex burn scar with jagged boundaries; SPEXT captures disconnected artifacts that baselines miss. **Middle Row:** A lower-contrast scene where CROMA misses the secondary scar and MS-CLIP fragments the central detection; SPEXT retrieves both accurately. **Bottom Row:** A challenging dark scene where MS-CLIP exhibits severe over-segmentation (hallucinating false positives across the background), whereas SPEXT suppresses these false positives and isolates the target scar.

A.13 Retrieval Protocol and EuroSAT Band Order

All SPEXT retrieval results (Tab. 3, Fig. 7) use the released weights, with queries and scores computed as in the released retrieval code.

Queries. Each class query inserts the class descriptor (App. A.14) at {c} into two dataset-specific templates, each preceded by the instruction prefix "Instruct: Given a professional remote sensing query, retrieve the most visually similar satellite imagery." and, on a new line, "Query:". The EuroSAT templates are "The image is a satellite view of a developed or natural area featuring {c}." and "A satellite image showing a mix of land uses including {c}." The METER-ML and BigEarthNet templates are "What is the dominant land use? The dominant land use is {c}." and "What are the most prominent features in this satellite view? The most prominent features are {c}." The two Qwen3-Embedding Zhang et al. [2025a] embeddings are ℓ_2-normalized, averaged and re-normalized.

Images and scoring. Images are resized bicubically to 128×128 and normalized per band with the pre-training statistics (App. A.1), after adding 1000 to EuroSAT and BigEarthNet inputs. METER-ML is normalized with per-band statistics of its test images, computed without labels. With cosine similarity s_{ic} between image i and class query c and $t = 5$, the score is $\log[\mathrm{softmax}_c(t\, s_{ic})\, \mathrm{softmax}_i(t\, s_{ic})]$ (dual softmax). For a fixed class this ranks images by $s_{ic} - \frac{1}{2t} \log \sum_{c'} \exp(t\, s_{ic'})$, the cosine similarity minus a per-image hub penalty that uses no statistics of the other images. Compared with plain cosine ranking, dual softmax changes mAP@100 from 63.39 to 68.74 on EuroSAT, from 27.70 to 29.95 on METER-ML and from 48.69 to 46.44 on BigEarthNet (full-ranking AP from 34.79 to 36.28). We use it for all datasets.

Metrics and datasets. mAP@100 averages over classes the average precision within the top 100 retrieved images, normalized by the relevant images among them. Full-ranking AP is computed over the whole ranking. EuroSAT Helber et al. [2019]: all 27,000 images, 10 classes. METER-ML Zhu et al. [2022]: 1,018 test images, whose 14 annotated types are merged into the 7 classes of Tab. 15, and the 27 images with two facility types count as relevant to both. BigEarthNet: the test split of reBEN (BigEarthNet v2) Clasen et al. [2025], 119,825 patches, 19-class multi-label nomenclature.

EuroSAT band order. The EuroSAT GeoTIFFs store B8A after B12. Our earlier EuroSAT evaluations dropped the stored bands at positions 0, 9 and 10, the selection for the fMoW-Sentinel band order, which MS-CLIP's public benchmark code also uses. On EuroSAT it feeds B09, B12 and B8A into the B8A, B11 and B12 slots. The released code selects the ten bands by explicit index. Correcting the band order raises SPEXT by $+15.49$ mAP@100 (53.25 to 68.74) but Llama3-MS-CLIP, with its own protocol, by only $+2.24$ (75.26 to 77.50), showing that SPEXT's wavelength-parameterized embedding exploits band identity. The METER-ML, BigEarthNet, PANGAEA and SSL4EO-S12 loaders select the correct bands, so no other result is affected.

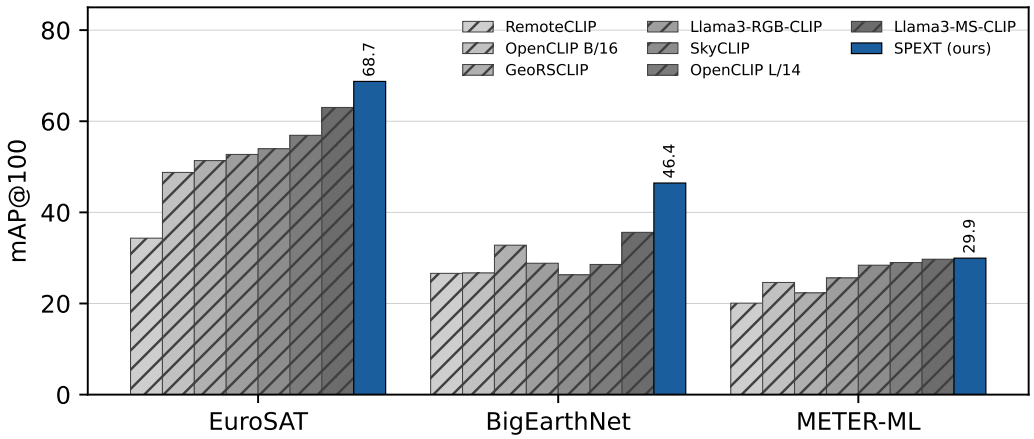


Figure 7: Zero-shot text-to-image retrieval, mAP@100. Hatched: published values from MS-CLIP's Table 1 Marimo et al. [2025] (their protocol, BigEarthNet v1). Solid: SPEXT, with the protocol above (corrected EuroSAT bands, BigEarthNet reBEN v2).

A.14 Descriptive Templates for Retrieval Tasks

To bridge the gap between sparse class labels and our model's pre-training on dense captions, we transform discrete category names into rich physical descriptions. Tables 15, 16, and 17 detail the exact mapping from official class labels to the class descriptors that fill the query templates of App. A.13.

Table 15: Narrative Prompts for the METER-ML Dataset.

Class Label	**Narrative Description**
CAFOs	An industrial agricultural area featuring multiple long, parallel rectangular barns and dark green manure lagoons.
Landfills	A waste management site with visible terraced earth mounds and piles of brownish excavation material.
Mines	A rugged mining site characterized by massive stepped excavations, jagged rocky surfaces, and a series of deep terraced layers.
Negative	A satellite photograph of a natural landscape with a mix of brown earth, green forest cover, and irregular grassy patches.
Processing Plants	An aerial view of a large factory with numerous rooftop ventilation units, external industrial pipes, and clustered rectangular grey buildings.
Refineries	A dense cluster of many white circular tank structures surrounded by asphalt roads and rectangular industrial buildings.
WW Treatment	A municipal utility site with round settlement tanks and large open-air rectangular water basins.

Table 17: Narrative Prompts for the BigEarthNet-S2 Dataset.

Class Label	**Narrative Description**
	Artificial Surfaces
Urban fabric	A built-up area showing a fine-grained urban texture of closely spaced homes and local street networks, creating a heterogeneous grey-toned surface that differs from the large blocks of industrial zones.
Industrial units	A developed area featuring large-scale industrial buildings, warehouses with expansive flat roofs, and wide impervious surfaces for commercial activities and transportation.

(continued on next page)

(Tab. 17 continued)

Class Label	Narrative Description
	Agricultural Areas
Arable land	A landscape of organized agricultural fields in a grid-like pattern, showing a mix of green and brown areas that indicate different stages of crop growth, tilled soil, or harvest.
Permanent crops	An aerial photograph of specialized permanent agriculture, such as a vineyard, fruit orchard, or nut grove, defined by its rigid linear architecture and the absence of seasonal plowing patterns.
Pastures	A landscape of enclosed green meadows divided by irregular hedgerows or wooden fences, appearing as soft-textured grassy carpets that lack the sharp geometric precision of cultivated crop fields.
Complex cultivation	A diverse, patchwork landscape of small, fragmented agricultural plots and garden areas with varying colors and textures, creating a busy and heterogeneous land use pattern.
Land w/ vegetation	A mixed rural environment characterized by fuzzy transitions between crop fields and wild vegetation, appearing as a patchwork of high-biomass forest regrowth and lower-biomass agricultural land.
Agro-forestry	A complex land-use pattern showing high-reflectance crop parcels integrated with the darker, lumpy texture of tree canopies, creating a park-like or orchard-adjacent appearance.
	Forests & Semi-Natural Areas
Broad-leaved forest	A lush forested landscape with high spectral vibrancy, where the dense arrangement of broad leaves creates a bright and highly reflective signature in the infrared bands.
Coniferous forest	A dense and continuous canopy of evergreen needle-leaf trees, appearing as a deep, dark-green expanse with significantly lower near-infrared reflectance than deciduous woodlands.
Mixed forest	A satellite view of a forest with high spectral diversity, where patches of vibrant-green broad-leaved trees are interspersed with the deep-green and darker tones of coniferous species.
Natural grassland	An open, natural landscape with native grasses, scattered rocky outcrops, and minimal tree cover, appearing predominantly in earth tones with thin green vegetation.
Moors & Heathland	A landscape of Mediterranean-style maquis or northern moorland, featuring a dense cover of short, hardy shrubs that create a consistent but dull green or brown carpet over poor, rocky soils.
Transitional woodland	A landscape dominated by dense brushwood and immature forest regrowth, showing high spatial heterogeneity and a lack of organized human-made patterns or tall, consistent tree canopies.
	Wetlands & Water Bodies
Beaches & Dunes	A highly reflective, bright pale land feature composed of fine sand or silt, appearing as a luminous white or light-beige strip along coastal edges or as vast, barren inland expanses.
Inland wetlands	A semi-aquatic land cover that is neither fully terrestrial nor open water, appearing as a dark, textured expanse of saturated organic matter and dense, swampy vegetation.
Coastal wetlands	A low-lying coastal zone characterized by a mix of grey-brown mudflats and salt-resistant halophytic vegetation, appearing as a mottled, high-moisture expanse where the land meets the sea.
Inland waters	A stationary or flowing water feature entirely embedded within the land, appearing as a sharp-edged, dark-toned reservoir, pond, or lake that contrasts strongly with the surrounding vegetation and soil.
Marine waters	An expansive, deep-blue satellite view of open ocean water, appearing as a smooth and uniform dark surface with near-zero reflectance in the infrared bands and minimal surface features.

Table 16: Narrative Prompts for the EuroSAT Dataset.

Class Label	Narrative Description
Annual Crop	Agricultural land with crop cultivation and visible farming patterns.
Forest	Natural landscape with dense tree cover.
Herbaceous Veg.	Natural land cover featuring wild grasses and sparse shrubs.
Highway	Transport infrastructure featuring long linear roads and highways.
Industrial	Industrial area with large buildings, warehouses, and developed facilities.
Pasture	Open green grassland with livestock grazing.
Permanent Crop	Perennial orchards, vineyards and plantation fields with regular tree rows.
Residential	Developed suburban area with residential houses and neighborhood buildings.
River	Natural waterway featuring a winding river or stream.
Sea/Lake	Large expanse of blue ocean water or deep lake surfaces.

A.15 RSVQAxBEN Evaluation Details

This section details the pre-registered RSVQAxBEN study of §4.3: its arms, recipe, metrics, class groups and predictions were fixed before any arm was trained.

Data. RSVQAxBEN Lobry et al. [2021] (Zenodo record 5084904, CC-BY-4.0) pairs each BigEarthNet patch Sumbul et al. [2019] with 25 questions generated from its CORINE labels: presence (yes/no) questions and land-cover (LC) questions, whose answers list the classes present. Patches are split geographically (training in the west, test in the east). We read each patch's 10 Sentinel-2 bands (B02–B08, B8A, B11, B12) from reBEN Clasen et al. [2025], matched through its `s2v1_name` column, and use the full official splits. After mapping, training has 8,120,275 questions on 324,811 patches (94.3% of 8,609,625) and test has 2,720,525 questions on 108,821 patches (92.1% of 2,953,125), with no patch in both. The test split holds 2,233,705 presence questions (36.68% "yes", majority baseline 63.32%) and 486,820 LC questions with 1,834 distinct answers. No hyper-parameter is tuned, so the validation split is not used.

Arms and recipe. Visual tokens come from SPEXT and Llama3-MS-CLIP Marimo et al. [2025] (patch tokens, 256×768 and 196×768 per patch) or, in the RGB arms, from Qwen3-VL's own vision tower on an RGB rendering, taken before its merger (1152-d tokens). Each encoder receives its native input: SPEXT the $+1000$ L2A offset and SSL4EO-S12 statistics at 128 px, Llama3-MS-CLIP its own offset-free statistics at 224 px, and the RGB arms B04/B03/B02, percentile-stretched, at 128 px through Qwen3-VL's processor. A freshly initialized 2-layer MLP adapter maps the tokens into the language model in every arm except RGB (Qwen merger) + LoRA, which uses Qwen3-VL's pretrained merger (4096-d) instead and is therefore not capacity-matched. The tokens are injected as a prefix into Qwen3-VL-8B-Instruct Bai et al. [2025b] (visual tokens, question, answer, with the loss on the answer only), whose language model is frozen or, in the "+ LoRA" arms, adapted with LoRA Hu et al. [2022] (rank 32, $\alpha = 64$, dropout 0.05, on the q, v, o and gate projections). Every arm uses AdamW (learning rate 10^{-3} for the adapter and 10^{-4} for LoRA, weight decay 0.01), a 1,000-step linear warm-up followed by cosine decay to 10%, a global batch of 64 and one epoch over all training questions, with seed 0.

Scoring and class groups. A presence answer is whichever of "yes" and "no" has the higher next-token logit. LC answers are generated greedily (at most 96 tokens) and scored by exact match of the normalized string (primary, RSVQAxBEN's standard) and by CORINE class-set F1 (secondary). OA is the accuracy over all test questions and AA the mean of presence and LC exact-match accuracy. Treating the 2.72 M test questions as independent, the sampling error of OA is about 0.03 points. Even if the 25 questions of a patch were fully correlated, it would stay below 0.15 points, well below SPEXT's margin over every other arm. Presence questions are grouped by the CORINE classes they mention, using pre-registered keywords matched in the lower-cased question: *artificial surfaces* (artificial, urban, industrial, commercial, transport, road, rail, port area, airport, mineral extraction, dump, construction, sport, leisure), *agriculture* (agricultur, arable, irrigated, rice, permanent crop, vineyard, fruit tree, olive, pasture, annual crop, cultivation, agro-forestry), *forest and semi-natural areas* (forest, grassland, heath, moor, sclerophyllous, woodland, shrub, scrub, herbaceous, beach,

dune, sand, bare rock, sparsely vegetated, burnt, glacier, snow, open spaces, semi natural, semi-natural) and *water and wetland* (water, wetland, marsh, peat, saline, intertidal, lagoon, estuar, sea and ocean, ocean). "Non-agricultural" does not count as agriculture, and "agro-forestry" counts as agriculture only. Questions that match several groups are *mixed*. The last three groups are the *spectral* groups, for which red-edge and SWIR bands were expected to carry more of the signal than for artificial surfaces.

Table 18: RSVQAxBEN presence accuracy per pre-registered class group (overall results in Tab. 4): artificial surfaces (Art.), agriculture (Agr.), forest and semi-natural areas (For.), water and wetland (Wat.), the three spectral groups together (Spec.) and questions spanning several groups (Mix.).

Arm	Art.	Agr.	For.	Wat.	Spec.	Mix.
SPEXT	**97.86**	**86.70**	**95.59**	**91.08**	**91.20**	**90.36**
Llama3-MS-CLIP	97.68	79.48	95.02	90.14	88.55	86.99
RGB (adapter)	97.70	77.60	91.78	90.37	87.08	86.12
RGB (adapter) + LoRA	97.15	77.99	90.61	89.65	86.54	85.92
RGB (Qwen merger) + LoRA	96.75	77.47	88.77	89.55	85.77	85.24

Pre-registered predictions. **P1** (frozen language model: SPEXT $>$ RGB (adapter) $>$ Llama3-MS-CLIP on OA) holds for SPEXT's first place, and Llama3-MS-CLIP places second, ahead of RGB (adapter). **P2** (SPEXT's presence margin over RGB (adapter) is larger on the spectral groups than on artificial surfaces) holds: $+4.1$ points against $+0.2$ (Tab. 18). **P3** (RGB (Qwen merger) + LoRA is the strongest arm, and SPEXT + LoRA $>$ RGB (adapter) + LoRA) does not hold, because its first part fails: RGB (Qwen merger) + LoRA has the lowest OA of the reported arms. LoRA also does not improve the RGB adapter arm.

A.16 Single-Encoder Pipeline Efficiency

Analytically, a SPEXT-based RAG pipeline is structurally cheaper than the standard EO-RAG stack because it removes two compute steps that competing pipelines must perform. Tab. 5 contrasts the two designs at the level of encoder forward passes and instantiated visual-encoder parameters per query, assuming a top-$K{=}3$ retrieval budget.

The standard pipeline relies on three distinct visual stages: a CLIP-style encoder (e.g., MS-CLIP, $\sim$87 M parameters) computes the index-time image embedding, the same encoder must remain available at query time to score the corpus, and the VLM's native ViT (the CLIP-ViT shipped with most open VLMs, $\sim$300–500 M parameters depending on the model) re-encodes each retrieved RGB image at generation time. SPEXT collapses these into a single encoder pass: the alignment token serves the retrieval role, and the same patch-level features cached at indexing time are forwarded through a small projection MLP and consumed directly by the VLM's language stack at generation time, with no separate vision encoder.

A SPEXT-RAG deployment therefore operates with a single visual encoder rather than two, and avoids invoking the VLM's native ViT on every retrieved image at query time. The latter is the dominant per-query visual cost in standard EO-RAG stacks, since the VLM's ViT is typically larger than the CLIP-style retrieval encoder and runs once per retrieved image. The gap widens with the retrieval depth K.

A.17 Efficiency Context

Tab. 19 situates SPEXT against recent leading EO foundation models in terms of parameter count, pre-training data scale, and downstream PANGAEA performance. With 87 M parameters and $\sim$1.7 M Sentinel-2 images for pre-training, SPEXT attains the highest PANGAEA average (60.11) while using $3.5\times$ fewer parameters than CROMA and roughly $4.7\times$ less pre-training data than DOFA. Among 87 M-parameter baselines, SPEXT exceeds the next best by $+5.9$ mIoU.

The data column above tallies only EO-specific pre-training and is therefore not directly comparable across families. The CLIP-style baselines (RemoteCLIP, MS-CLIP, Llama3-MS-CLIP, DOFA-CLIP) initialize from a web-scale pre-trained CLIP backbone before EO fine-tuning, so their effective representational data exposure is orders of magnitude larger than the listed corpus. SPEXT, in

Table 19: Parameter, pre-training, and downstream-performance comparison of SPEXT against open multi-spectral and CLIP-style EO foundation models. PANGAEA Avg is the 8-task mIoU from Tab. 1. Pre-training data figures are approximate and follow each model's source paper. CLIP-style baselines such as RemoteCLIP, MS-CLIP, and DOFA-CLIP additionally initialize from a web-scale pre-trained CLIP backbone, with the data column reflecting only their EO-specific fine-tuning corpus.

Family	Model	Params	Pre-training data	PANGAEA Avg ↑
Multi-spec. MIM	Prithvi	87M	∼250k HLS (S2+L30)	54.22
	SpectralGPT	105M	∼1M S2	55.49
	DOFA	112M	∼8M images (5 modalities)	57.44
	CROMA	303M	2M S1+S2 pairs (SSL4EO-S12)	57.60
CLIP-style contrastive	RemoteCLIP	87M	∼165k RGB image-text pairs	52.68
	MS-CLIP	87M	∼1M SSL4EO-S12 S2 image-text pairs	53.72
	DOFA-CLIP	457M	∼8M (5 modalities) + captions	53.38
	SatlasNet	87M	∼10M tiles (multi-task)	53.73
S2-MAE baselines	S12-MAE	22M	∼1M SSL4EO-S12 S2	53.71
	Scale-MAE	303M	∼370k FMoW-RGB	50.41
	SPEXT (ours)	**87M**	**∼1.7M S2 (fMoW-S2 + SSL4EO-S12)**	**60.11**

contrast, is trained from random initialization on Sentinel-2 imagery only, with text targets from a frozen, off-the-shelf sentence encoder.

A.18 Matched Coupled versus Decoupled Pair

The routing comparison of §4.4 uses two encoders that share the architecture, spectral embedding, Stage I initialization, data, objectives (MAE, MTP and ITC) and 20-epoch Stage II schedule. They differ only in where the ITC loss is applied: to the dedicated alignment token (*decoupled*, as in SPEXT) or to the mean of the visible patch embeddings (*coupled*). Tab. 6 reports every PANGAEA task in both label regimes. Downstream training follows the PANGAEA protocol rather than the 20-epoch MADOS schedule of Tab. 7, and absolute values are not comparable with Tab. 1, which uses the released 50-epoch model. Tab. 2 compares the spectral fidelity of the two encoders. All its columns are measured on 8,704 SSL4EO validation tiles with 75% of patches masked, with SAM on masked pixels and the index R^2 on all pixels, using offset-included values.

Most per-task differences are small relative to the spread across seeds. The largest, MADOS with full labels, favors the decoupled encoder and reverses sign with 10% labels. Routing the alignment loss to a dedicated token therefore has no dense-task cost outside seed noise.

A.19 Spectral Bands and Stage I

All arms of Tab. 8 use 20-epoch Stage II pre-training with the SPEXT recipe and are evaluated under the PANGAEA protocol at seed 42. Arm (a) is the decoupled encoder of App. A.18, whose three-seed means are in Tab. 6. Arm (b) evaluates the weights of (a) on RGB input, with every non-RGB band held at its dataset mean (zero after per-band standardization), so the architecture is unchanged. Arms (c) and (d) start from random initialization, and (d) uses the same RGB input in pre-training and evaluation. CropMap is seed-sensitive: at seed 234, arm (a) reaches 53.47 and arm (c) 50.74. On MADOS, three seed-42 runs of arm (b) gave 43.46, 44.83 and 43.71, and the table lists the last. Stage I's contribution is measured by arm (c), and that of the Stage II objectives by Tab. 7, whose rows share one Stage I initialization.

A.20 MTP Cross-Attention, Pooled Conditioning and Feature Geometry

Cross-attention. Using the full-schedule SPEXT model with all patches visible, we mask one caption sentence at a time in 24 multi-sentence SSL4EO-S12 validation scenes (79 masked sentences) and record the MTP decoder's cross-attention from the masked slot over the 16×16 patches, averaged over heads and the four decoder blocks. The 10% most-attended patches receive on average 0.145 of the attention mass (range 0.125–0.174), against 0.10 under uniform attention, and the mean attention entropy is 5.52 nats, against 5.55 for uniform attention. Sentences that name an object span the same range as scene-level ones (Fig. 8). The decoder thus aggregates over the whole scene rather than localizing individual sentences, consistent with MTP's sentence-to-scene supervision (§3.2): MTP acts as dense semantic regularization, keeping the patches informative about the scene, rather than as spatial grounding.

Pooled conditioning. To test whether MTP needs patch-level access, we pre-train a variant whose MTP decoder cross-attends to one globally pooled visual token instead of the patch tokens, against its patch-conditioned counterpart, arm (c) of Tab. 8. Both are 20-epoch pre-trainings from random initialization. Pooled conditioning is ahead on HLS and MADOS, slightly behind on Sen1Floods11, level on SN7 and behind on AI4SmallFarms, and on CropMap the sign of the difference flips with the seed (Tab. 20). The two conditionings thus perform comparably, consistent with MTP acting as scene-level semantic regularization.

Table 20: MTP conditioning control (PANGAEA full split, mIoU, seed 42 unless stated, both arms 20-epoch pre-trainings from random initialization). Δ = pooled − patch. *DICE loss.

MTP decoder attends to	HLS	MADOS	Sen1Floods11	CropMap (42 / 234)	SN7	AI4Farms*
All patch tokens	79.79	57.93	90.42	37.99 / 50.74	57.11	30.79
One pooled token	81.31	60.09	89.60	57.00 / 39.81	57.08	26.46
Δ	$+1.52$	$+2.16$	-0.82	$+19.01$ / -10.93	-0.03	-4.33

Feature geometry. Tab. 21 probes whether the text objectives pull the patch space toward the text embedding or compress it, for five 20-epoch variants on 1,024 SSL4EO-S12 validation scenes.

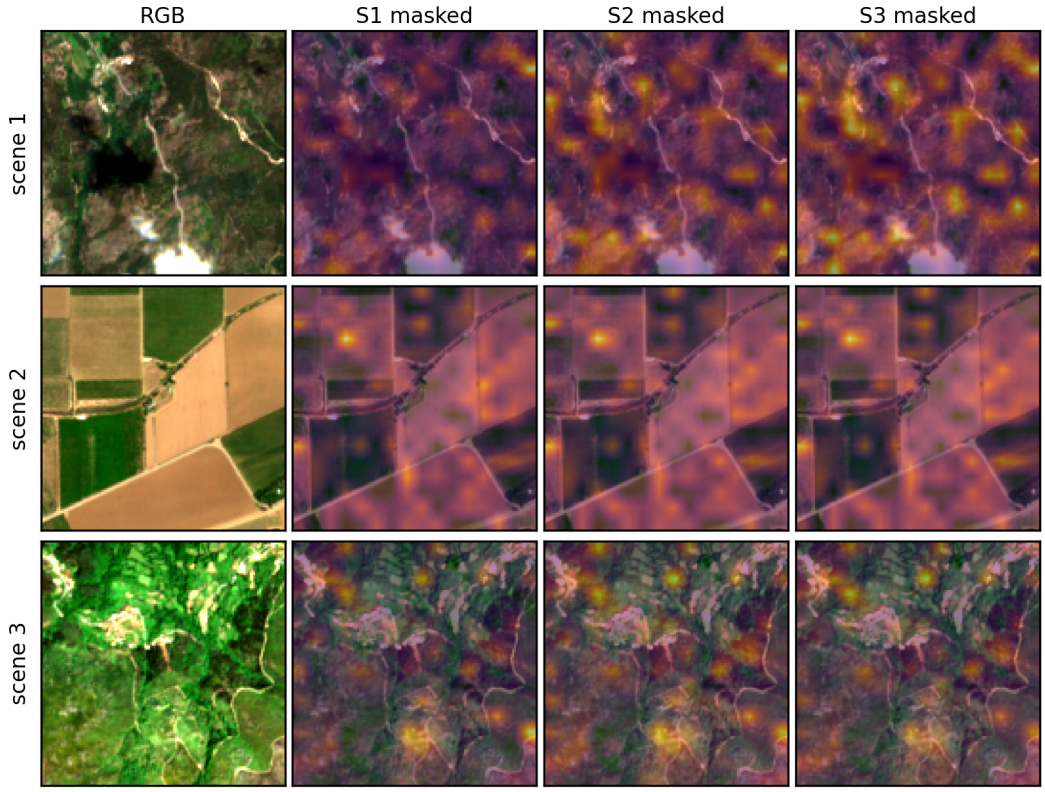


Figure 8: MTP decoder cross-attention over patches for the first three of the 24 analyzed scenes in file-name order (not selected). Columns: the RGB composite, then the attention when the scene's first, second or third caption sentence (S1–S3) is masked. For example, S2 of scene 1 is "The landscape is dominated by a large forested area and a small stream." The attention spreads over the whole scene, consistent with scene-level supervision.

Relative to MAE alone, every text-supervised variant has a higher effective rank and a lower top-10 eigenvalue share, so no text objective compresses the patch space. CKA rises with each text objective and is highest for the coupled variant, whose patch features receive the ITC loss directly.

Table 21: Patch-space geometry of five 20-epoch variants (1,024 SSL4EO-S12 validation scenes, all patches visible). CKA: linear CKA between mean-pooled patch features and the frozen text embedding that serves as the ITC target. Effective rank (participation ratio) and top-10 share: of the eigenvalue spectrum of the patch-feature covariance. Each ITC row names the representation that receives the ITC loss.

Variant	CKA	Effective rank	Top-10 eigenvalue share
MAE only	0.203	16.2	0.696
MAE + MTP	0.252	16.7	0.652
MAE + ITC (alignment token)	0.282	17.4	0.653
MAE + MTP + ITC (alignment token, decoupled)	0.323	17.0	0.646
MAE + MTP + ITC (pooled patches, coupled)	0.394	21.4	0.584

A.21 Recent Multispectral and RGB Encoders

Tab. 22 collects published results of recent encoders on the PANGAEA tasks we report, each with the protocol under which it was obtained.

Multispectral encoders. SMARTIES Sumbul et al. [2025] reports three PANGAEA tasks under the benchmark's default protocol, two of them in our suite. AnySat Astruc et al. [2025] uses its

Table 22: Recent encoders on the PANGAEA tasks (mIoU), published numbers only. MADOS and CropMap are omitted because none of these models reports them. [a]Frozen encoder with UPerNet under PANGAEA's default protocol, from the SMARTIES paper. [b]AnySat's own probe with 3M free parameters (full fine-tuning: HLS 90.6, SN7 58.8). [c]Probe with 33K free parameters on PASTIS-HD fold 1 with S1, S2 and SPOT-6 input (fine-tuning: 65.9). All of PASTIS-HD is in AnySat's pre-training corpus. [d]Linear probe with 6K free parameters on S1+S2 input. [f]Run by the GeoLink authors with PANGAEA-bench's UPerNet decoder and training configuration.

Model	Params	Input	HLS	PASTIS	Sen1Fl.	xView2	SN7	AI4Farms
Multispectral encoders								
SMARTIES	ViT-B	S2, S1, RGB	82.8[a]	–	–	–	62.2[a]	–
AnySat	125M	11 sensors	87.7[b]	42.7[c]	91.1[d]	–	58.1[b]	–
SPEXT (ours)	87M	S2	81.57	41.28	90.04	47.88	59.05	34.18
Reported in the GeoLink paper (frozen encoders)[f], not comparable with the rows above								
GeoLink			–	–	–	59.94	63.22	43.26
Cross-Scale MAE			–	–	–	58.44	60.23	40.33
CROMA			–	–	–	57.34	58.97	41.16

own probing and fine-tuning protocols, with Sen1Floods11 on S1+S2 and PASTIS-HD inside its pre-training data, and TESSERA Feng et al. [2026] encodes annual pixel time series without a single-date mode. None of the three has a text encoder or retrieval capability. A shared-protocol comparison over our eight tasks is therefore not possible from published numbers. For SMARTIES, its announced integration into PANGAEA-bench will enable the complete comparison.

RGB and other-band encoders. Cross-Scale MAE Tang et al. [2023], RingMo Sun et al. [2023], SkySense V2 Zhang et al. [2025b] and GeoLink Bai et al. [2025a] are not evaluated in the PANGAEA benchmark itself, and RingMo and SkySense V2 report no PANGAEA-protocol results (SkySense V2's PASTIS-MM is not PANGAEA's PASTIS-HD). Their band access also differs from SPEXT's, a confound that the RGB-only arms of Tab. 8 quantify for SPEXT. The GeoLink paper reports frozen-encoder results for several encoders, among them GeoLink, Cross-Scale MAE and CROMA, on four PANGAEA datasets with PANGAEA-bench's UPerNet decoder, and the bottom block of Tab. 22 reproduces the three datasets in our suite. These runs are not comparable with official PANGAEA results: CROMA scores 41.16 on AI4SmallFarms there and 25.65 in the official results (Tab. 1).

A.22 Extension to Multitemporal Imagery

SPEXT is pre-trained on single-date imagery. The temporal PANGAEA tasks (PASTIS, CropMap) aggregate per-date SPEXT features with the benchmark's L-TAE decoder. A multitemporal version would extend the existing design rather than replace it. On the vision side, patch tokens would receive a temporal encoding alongside the spectral one, such as a per-date Fourier feature analogous to the wavelength Fourier features of the wavelength-conditioned modulation, and MAE masking would extend to spatio-temporal tubes. On the text side, MTP's supervision unit is already the sentence, so a multitemporal caption becomes a sequence of dated sentences ("in March, bare fields" and "by July, dense canopy"), each embedded by the frozen sentence encoder with its date folded into the sentence, as location and season are folded into the geographic anchor sentence (§3.2). Recovering a masked dated sentence from the features of the whole temporal stack would then supervise change directly. The main requirement is a captioned multitemporal corpus, which we leave to future work.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and the contribution list in §1 (physics-informed spectral encoding, decoupled semantic supervision; empirical results) match the per-task results reported in §4.1–§4.4 and the appendix; we explicitly bound each claim (e.g. "best on PANGAEA average", "supports zero-shot retrieval").

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: §6 ("Conclusion & Limitations") states three explicit constraints: dependence on automatically generated captions, the optical-only spectral prior (would require adaptation for SAR), and that SPEXT does not lead on every individual PANGAEA task in every setting.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not present formal theorems or proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: §3 fully specifies the architecture (wavelength-parameterized embedding, Latent MTP, alignment token, two-stage curriculum); App. A.1 lists all pre-training hyperparameters; App. A.2 gives the loss formulations; downstream evaluation uses the published PANGAEA fine-tuning protocol Marsocci et al. [2024].

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All datasets used (SSL4EO-S12, fMoW-S2, PANGAEA, EuroSAT, METER-ML, BigEarthNet v2 (reBEN), RSVQAxBEN) and all baseline checkpoints (S12-MAE, S12-Data2Vec, S12-MoCo, S12-DINO, GFM-Swin, SatlasNet, Prithvi, Scale-MAE, SpectralGPT, DOFA, CROMA, DOFA-CLIP, MS-CLIP, Llama3-MS-CLIP, RemoteCLIP) are publicly available; source code and pre-trained SPEXT weights are released at `https://github.com/Sannaraek/SPEXT`.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: App. A.1 reports architecture (ViT-Base, 128×128 inputs, patch size 8), asymmetric masking (90% visual, 50% text), AdamW with learning rate 1×10^{-4}, global batch 2048, hardware, and per-stage epoch counts, as well as Stage I's 96×96 inputs and 75% visual masking; downstream protocol matches the standard PANGAEA fine-tuning recipe Marsocci et al. [2024].

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: The component ablation and the masking-ratio sweep (Tab. 7), the matched decoupling pair (App. A.18) and the loss-weight sensitivity analyses (App. A.7) report mean $\pm$ std over 3 random seeds. The headline PANGAEA and retrieval tables follow the single-seed protocol used by their respective benchmark suites.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: App. A.1 reports 4 NVIDIA GH200 GPUs for pre-training, with Stage I 100 epochs on fMoW-S2 ($\sim$712k tiles) and Stage II 50 epochs on SSL4EO-S12 ($\sim$1M tiles); downstream PANGAEA fine-tuning uses 4 GPUs per task. Each RSVQAxBEN arm trains on one node with 4 NVIDIA GH200 GPUs, for about 23 h with the language model frozen and 34–36 h with LoRA.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses only publicly released remote-sensing imagery and benchmark datasets, follows the NeurIPS Code of Ethics, and involves no human subjects or scraped personal data.

   Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Multi-spectral foundation models support agriculture, environmental monitoring, and disaster response. Potential misuse (e.g. surveillance) is a property of remote-sensing imagery in general rather than something introduced by our architecture; we discuss scope and constraints in §6.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The released model is a multi-spectral image encoder pre-trained on public Sentinel-2 imagery; it does not generate text or images, contains no scraped personal data, and presents no high-risk dual-use concerns beyond those already inherent to EO imagery.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Each dataset (SSL4EO-S12, fMoW-S2, PANGAEA tasks, EuroSAT, METER-ML, BigEarthNet v2 (reBEN), RSVQAxBEN (Zenodo 5084904, CC-BY-4.0)) and each pre-trained model (Qwen3-Embedding-0.6B, Qwen3-VL-8B-Instruct; both Apache-2.0, and Llama3-MS-CLIP under its own license) is cited at first use, and each is used within its license terms (the EO datasets are predominantly CC-BY-4.0 or research-permissive).

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The SPEXT weights (final Stage-II checkpoint), training and evaluation code, retrieval templates and geographic-anchor metadata are released at `https://github.com/Sannaraek/SPEXT` and documented in the release README. Architecture and training config are documented in §3 and App. A.1.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

Justification: All RSVQAxBEN questions are generated automatically by the benchmark's authors from public CORINE land-cover labels; no crowdsourcing or human subjects are involved.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: No human-subjects research.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs are core components of the methodology: Qwen3-Embedding-0.6B (frozen sentence encoder used for the MTP target and for retrieval queries, §3.2/§3.3) and Qwen3-VL-8B-Instruct (answering VLM in the RSVQAxBEN study, §4.3). Each is identified at first use with citation. We additionally used LLMs for writing assistance (copy-editing and prose tightening) on the manuscript itself. This writing-only use does not affect the methodology, scientific rigor, or originality of the work.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.