PFU: A Universal Interface between Earth Observation Data and Frozen Foundation Models

Anonymous authors
Paper under double-blind review

Abstract

RGB foundation models discard measurements that do not fit their three-channel interface. We study a different reuse strategy: a trainable spectral frontend that maps Sentinel-2 bands to pseudo-RGB while leaving the deployed backbone unchanged. Under one matched protocol and three predetermined seeds, PFU-E improves over the RGB baseline by 3.482 micro-mAP points on BEN-MM and 3.644 accuracy points on So2Sat. A 15-parameter spectral projection is already stronger than RGB on BEN-MM, while capacity-matched and test-time B08 controls show that the spatial frontend uses aligned NIR rather than only additional parameters. Beyond classification, pretrained PFU-E reaches 41.935 mIoU on PASTIS-HD, 19.909 points above a bare linear probe and 3.351 above scratch PFU. Twelve-band PFU-D improves Cashew Plant segmentation with both LVD and SAT backbones, and PFU-I dual-path and S1+S2 experiments demonstrate that the interface extends to wider spectra and multimodal inputs.

1 Introduction

EO imagery differs from natural images in both channel semantics and dynamic range. A common reuse strategy drops non-RGB measurements and applies an RGB foundation model, while alternatives modify the input stem, learn spectral tokens, or pretrain a sensor-aware model. Modifying the backbone is effective when one owns a single model and sensor stack, but it creates a new checkpoint for every channel configuration and complicates validation of an already deployed encoder. An external interface instead allows the same signed, quantized, or remotely served RGB backbone to be reused across sensors while isolating sensor-specific updates in a small module.

PFU maps a four-band observation to a three-channel image before the frozen DINOv3-LVD encoder (Siméoni et al., 2025). We separate four factors: (i) RGB versus RGB+NIR access under an identical 4.879M-parameter frontend, (ii) a 15-parameter spectral projection versus a spatial frontend, (iii) aligned NIR versus an equal-capacity RGB-only frontend, and (iv) scratch versus externally pretrained initialization. We then broaden the study to 12 Sentinel-2 bands, S1+S2 fusion, DINOv3-SAT, V-JEPA, AIMv2, visual prompt tuning, and dense segmentation. We use PFU-D for downstream-trained spectral patch adaptation and PFU-I for pretrained frozen 4-band and dual-path variants.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

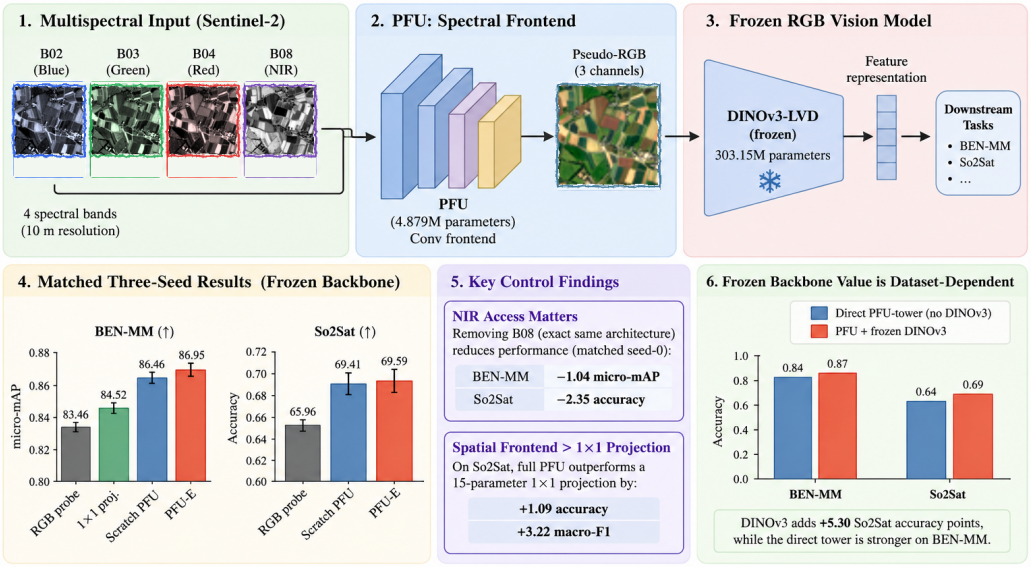


Figure 1: Overview of PFU and the main controlled results. Top: PFU maps four Sentinel-2 bands (B02/B03/B04/B08) through a trainable spectral frontend to a pseudo-RGB image consumed by a frozen DINOv3-LVD backbone. Bottom: matched classification results and completed controls for spectral access, spatial processing, and backbone interaction. Exact protocols and seed structure are reported in Section 5.

The paper makes three contributions:

1. a formal three-seed comparison of RGB probing, a learned 1×1 projection, scratch PFU, and PFU-E, supported by capacity-matched and NIR intervention controls;
2. dense-prediction evidence on PASTIS-HD and 12-band Cashew Plant, where PFU-E, PFU-D, and PFU-I improve over their corresponding stored baselines; and
3. evidence that the interface supports 4- and 12-band optical inputs, S1+S2 multimodal fusion, and both DINOv3-LVD and DINOv3-SAT backbones.

2 Related Work

Multispectral EO representation learning. SatMAE groups spectral and temporal satellite tokens for masked pretraining (Cong et al., 2022); CROMA jointly learns radar and optical representations (Mañas et al., 2023); and DOFA generates wavelength-aware weights to cross sensor modalities (Xiong et al., 2024). OLMoEarth supplies global multimodal temporal data and open EO pretraining resources (Mansoor et al., 2025). Prithvi, Scale-MAE, SSL4EO-S12, and Clay establish the complementary line of general-purpose EO foundation models (Jakubik et al., 2023; Reed et al., 2023; Wang et al., 2023; Clay Foundation, 2024). ChannelViT represents each input channel independently to support varying channel sets (Liu et al., 2023), while SpectralEarth studies hyperspectral pretraining with spectral adapters (Scheibenreif et al., 2025). These approaches internalize EO channel structure in the representation model. PFU instead keeps an existing RGB encoder unchanged and learns a low-level image interface.

Spectral adapters and learned interfaces. SpectraDINO adapts frozen self-supervised vision features to satellite spectra with a spectral adapter and distillation (Nalcakan et al., 2026); HSI-Adapter extends vision foundation models to hyperspectral segmentation with an explicit spectral adapter (Hurtado et al., 2026). These are the closest methodological precedents. Relative to them, our contribution is not the first use of an adapter, but the comparison of an image-space interface against RGB, minimal spectral mixing, and an equal-

capacity RGB-only control. SpectraDINO and HSI-Adapter modify the spectral adaptation mechanism inside the model; PFU targets deployments in which the backbone endpoint must remain unchanged.

Parameter-efficient transfer. Adapters (Houlsby et al., 2019), low-rank updates (Hu et al., 2022), and visual prompt tuning (Jia et al., 2022) reduce the number of trainable parameters inside a pretrained model. PFU is complementary: it leaves all backbone parameters untouched and trains an external frontend plus task head. This reduces the trainable parameter fraction but does not imply lower runtime or memory than RGB probing.

3 Method

Let $x \in \mathbb{R}^{4\times H\times W}$ contain B02, B03, B04, and B08. The frontend G_ϕ emits $\tilde{x} \in [0,1]^{3\times 224\times 224}$, which is ImageNet-normalized and passed through frozen DINOv3-LVD F:

$$\hat{y} = h_\theta(F(\mathrm{norm}(G_\phi(x)))) \,. \tag{1}$$

Only ϕ and the linear task head θ receive gradients.

Exact frontend. The single S2 tower applies four 3×3 Conv–GroupNorm–GELU blocks with channels $4\to 96\to 96\to 192\to 192$; the first and third convolutions have stride 2. A 1×1 projection gives 384 channels. Because the headline setting has one modality and one date, the temporal mean is the identity and the softmax fusion gate is identically one. Fusion post-processing is a 3×3 384-channel block followed by a 1×1 projection. The shared projection is a 3×3 384-channel block and a 1×1 $384\to 512$ layer. The RGB head is a 3×3 $512\to 256$ block and a 1×1 $256\to 3$ convolution, followed by sigmoid and bilinear resize. All 3×3 convolutions use padding 1 and no bias; normalization is GroupNorm with up to 8 groups, except the final 512-channel projection, which uses up to 32 groups. The resulting frontend has 4,879,300 parameters. Including the task head gives 4.899M trainable parameters on BEN-MM and 4.897M on So2Sat.

Resolution path and diagnosed bottleneck. The evaluated implementation first resizes the sensor image to 224^2, then uses two stride-2 tower stages and predicts pseudo-RGB on a 56^2 grid before bilinear upsampling to 224^2. Thus the backbone never receives more than 56^2 learned spatial detail. This is especially restrictive for BEN-MM, whose native S2 crop is 120^2; for So2Sat, the 32^2 crop is enlarged before the same downsampling. A resolution-preserving successor should process native resolution and add a stride-1 skip branch to the upsampled deep feature. We retain the evaluated path for all reported numbers so that every comparison uses the same implementation; resolution-preserving variants remain a targeted follow-up rather than an unreported change to the evaluated model.

PFU-E initialization. PFU-E changes only the frontend initialization. Because the retained pretraining run selected a checkpoint from training loss and used z-scored inputs while the downstream tasks use reflectance, we treat it as an initialization ablation and move its full objective and resource accounting to Appendix C.

PFU-D and PFU-I extensions. PFU-D replaces the RGB patch projection with a trainable 12-band adapter while keeping the transformer frozen; the linear variant uses a pointwise 12-to-3 mapping, whereas the spectral variant injects learned multispectral patch features. PFU-I loads an externally pretrained spectral adapter and compares a frozen four-band path with a dual-path configuration that combines the retained four-band representation with a wider spectral path. These variants share the frozen-backbone principle but use separate training protocols from the formal classification block.

4 Experimental Protocol

Datasets and preprocessing. BEN-MM (Sumbul et al., 2021) is evaluated as 19-label classification with micro-mAP primary. Raw S2 values are divided by 10,000, clipped to $[0,3]$,

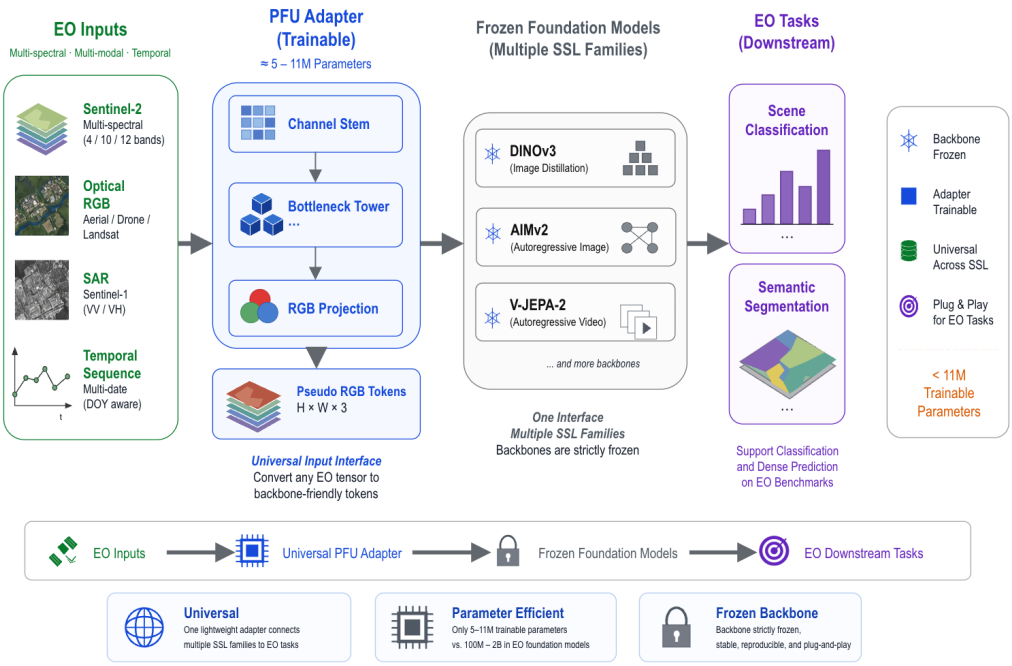


Figure 2: Conceptual overview of the PFU adapter. The diagram summarizes EO inputs, the trainable spectral interface, pseudo-RGB output, frozen vision models, and downstream classification or segmentation. The experiments in this paper use Sentinel-2 optical inputs; the headline setting uses one date and a frozen DINOv3-LVD backbone.

reduced to B02/B03/B04/B08, and bilinearly resized from 120^2 to 224^2. So2Sat LCZ42 (Zhu et al., 2020), using the GFM-Bench export (Si et al., 2025), is evaluated as 17-class classification with accuracy primary. Its reflectance is already scaled; the same bands are clipped to $[0, 3]$ and resized from 32^2 to 224^2. No random downstream image augmentation is applied.

Baselines and reviewer controls. The RGB probe selects $[B04, B03, B02]$, multiplies by 3.5, clips to $[0, 1]$, and trains only the task head. The learned projection is Conv2d(4,3,1) with 15 parameters, initialized to the exact RGB mapping and zero NIR weight. It is part of the formal three-seed block on BEN-MM; the retained So2Sat projection is a separate BF16 seed-0 reviewer control and is not reported as a formal aggregate. The capacity-matched RGB-only control retains the complete four-channel PFU architecture and parameter count but sets B08 to zero in train, validation, and test. Scratch PFU and PFU-E train the same frontend and head; the latter loads the external initialization. No VPT, backbone normalization update, alignment head, or full finetuning is enabled in the headline protocol.

Optimization and statistics. All headline runs use 20 epochs, batch size 32, AdamW, frontend learning rate 3×10^{-4}, head multiplier 3, weight decay 10^{-4}, 1,000 linear warmup batches followed by per-batch cosine decay, and max gradient norm 1. Each checkpoint is chosen using the validation primary metric before one test evaluation. The original headline block reports predetermined seeds $\{0, 1, 2\}$, sample standard deviation, and seed-paired Student-t intervals; with two degrees of freedom, these intervals are descriptive. Reviewer controls use BF16 and the same global batch size and optimizer schedule. Full versus RGB-only and So2Sat full versus 1×1 are matched seed-0 comparisons. We do not pool the BF16 controls with the original aggregates. Seeds and data-loader generators were fixed, but deterministic kernels were not requested.

Extension protocols. The PASTIS-HD comparison uses seed 0, 30 epochs, a frozen DINOv3-LVD backbone, and a U-Net decoder. PFU-D Cashew Plant rows use all 12 optical bands, 60 epochs, batch size 16, and a frozen LVD or SAT backbone. PFU-I classification uses

30 epochs and batch size 64; PFU-I segmentation uses the original stored artifacts, whose LVD pair is decoder-matched but whose SAT pair changes from a linear decoder to U-Net. The BEN input-scaling rows are historical 30-epoch single runs. We report these blocks as extension evidence without error bars and do not pool them with the formal three-seed results.

4.1 Metric Definitions and Reproducibility

BEN-MM micro-mAP is computed from sigmoid scores and is threshold-free. Its reported macro-F1 uses a fixed 0.5 threshold for every one of the 19 classes; no threshold is selected on validation, and classes are equally weighted. So2Sat predictions use argmax over 17 classes, followed by an unweighted mean of per-class F1. The retained headline artifacts contain aggregate metrics but not test logits or per-class vectors, so per-class AP/F1 and the classwise source of the So2Sat macro-F1 change cannot be reconstructed without rerunning inference. Appendix E contains the split/hash audit; it finds no exact cross-split duplicate but cannot test geographic near-duplicates or reconstruct overlap with OLMoEarth pretraining.

5 Results

Table 1: Formal matched three-seed classification results with frozen DINOv3-LVD. Values are percentages and report mean ± sample standard deviation over seeds $\{0, 1, 2\}$. The So2Sat 1×1 projection was not completed under this formal three-seed protocol.

Method	BEN-MM micro-mAP (%)	So2Sat accuracy (%)
RGB baseline	83.464 ± 0.001	65.955 ± 0.158
1×1 mapping	84.519 ± 0.002	–
PFU scratch	86.461 ± 0.125	69.411 ± 0.837
PFU-E	86.946 ± 0.076	69.599 ± 0.245

Formal three-seed classification. The concise main table reports the only protocol-matched three-seed comparison. On BEN-MM, the 1×1 mapping improves over RGB by 1.055 points, scratch PFU by 2.997 points, and PFU-E by 3.482 points. On So2Sat, scratch PFU and PFU-E improve accuracy by 3.456 and 3.644 points, respectively. PFU-E is the best mean result on both datasets under this protocol, while scratch PFU shows that external pretraining is not required for a substantial gain.

5.1 Dense Segmentation Extensions

Table 2: PASTIS-HD segmentation with frozen DINOv3-LVD (seed 0). All rows use 30 epochs and a U-Net decoder.

Method	mIoU (%)
LVD bare LP	22.026
LVD VPT-Deep scratch	38.584
LVD VPT-Deep pretrained	41.935

Table 3: PFU-D spectral adaptation on 12-band Cashew Plant segmentation. Each entry is one run with a frozen backbone and U-Net decoder.

Backbone	Linear	Spectral	Gain
DINOv3-LVD	55.360	68.125	+12.765
DINOv3-SAT	70.585	72.496	+1.911

On PASTIS-HD, PFU-E with VPT-Deep reaches 41.935 mIoU, improving over the bare linear probe by 19.909 points and over scratch PFU with the same VPT-Deep setting by 3.351 points. On 12-band Cashew Plant, the PFU-D spectral adapter improves over the linear adapter for both frozen backbones: +12.765 mIoU with DINOv3-LVD and +1.911 with DINOv3-SAT. The consistent direction across backbones supports spectral adaptation, although these are single runs.

Table 4: PFU-I dual-path results from the original pfu_i_eval and pfu_i_seg artifacts. Values are percentages from single runs. The LVD segmentation pair uses U-Net in both rows. The SAT frozen 4-band artifact uses a linear decoder, whereas its dual-path counterpart uses U-Net; the marked difference is therefore a configuration gap, not an isolated dual-path ablation.

Task	Backbone	Frozen 4-band	Dual-path	Gain
BEN classification mAP	DINOv3-LVD	64.746	68.702	+3.956
BEN classification mAP	DINOv3-SAT	67.964	69.922	+1.958
Cashew segmentation mIoU	DINOv3-LVD	54.046	64.632	+10.586
Cashew segmentation mIoU†	DINOv3-SAT	49.745	73.368	+23.623

5.2 PFU-I Dual-Path Transfer

The dual-path configuration improves BEN classification mAP for both backbones and raises LVD Cashew segmentation mIoU by 10.586 points under a decoder-matched comparison. The larger SAT segmentation gap is visually prominent but not a clean path ablation because the stored frozen baseline uses a linear decoder and the dual-path run uses U-Net.

5.3 Multi-band and Multimodal Inputs

Table 5: BEN-MM input scaling with PFU-E. These are single historical 30-epoch runs and demonstrate support for wider optical inputs and S1+S2 multimodal fusion; they are not pooled with the formal three-seed block.

Backbone / configuration	Input	micro-mAP (%)	Macro-F1 (%)
LVD / M1L	S2 4-band	88.544	72.822
LVD / M3L	S2 12-band	89.297	73.436
LVD / M8L	S1+S2	89.078	73.234
SAT / SPB	S2 12-band	89.516	73.568
SAT / SPBmm	S1+S2	89.607	73.817

Moving from 4 to 12 S2 bands improves the LVD configuration by 0.753 micro-mAP and 0.614 macro-F1 points. Adding S1 to the 12-band optical input changes micro-mAP by -0.219 for LVD and +0.091 for SAT. Thus the main conclusion is input and modality adaptability rather than a universal SAR gain; the strongest stored configuration is SAT/SPBmm at 89.607 micro-mAP and 73.817 macro-F1.

5.4 Capacity and Spectral Controls

Table 6: BF16 capacity and spectral controls. Full PFU, RGB-only PFU, and the So2Sat 1×1 projection are paired seed-0 runs. Cells show primary metric / macro-F1.

Dataset	Control	Evidence	Control P/F1	Full P/F1	Paired difference P/F1
BEN-MM	Capacity-matched RGB-only PFU	seed 0; paired	85.215 / 66.551	86.250 / 67.874	+1.035 / +1.323
So2Sat	Capacity-matched RGB-only PFU	seed 0; paired	67.220 / 55.807	69.572 / 59.584	+2.352 / +3.778
	1x1 Conv 4-to-3	seed 0; paired	68.481 / 56.363	69.572 / 59.584	+1.091 / +3.222

The equal-capacity RGB-only control isolates spectral access from model size. At seed 0, aligned B08 raises BEN-MM micro-mAP by 1.035 points and So2Sat accuracy by 2.352 points with the architecture and optimizer fixed. On So2Sat, full PFU also exceeds the learned 1×1 projection by 1.091 accuracy and 3.222 macro-F1 points, separating spatial processing from simple channel mixing.

Architecture considerations. The evaluated DINOv3 path uses only the class token. BEN-MM assigns up to 19 land-cover labels to a 120^2 scene, so future work should test whether patch aggregation better preserves spatially distributed evidence. An auxiliary 3,000-sample feature study motivates this ablation: replacing DINO class-token pooling with patch-token mean pooling raises average cross-backbone CKA from 0.544 to 0.590. This is a representation diagnostic, not a downstream accuracy result, but it points directly to mean, CLS+mean, or attention pooling as the next controlled ablation. PFU also predicts on a 56^2 grid and then upsamples, so the frozen transformer receives interpolated pseudo-RGB rather than native BEN-MM detail. These observations identify specific pooling and resolution ablations, but are not used as evidence for the main classification gains.

Table 7: Seed-paired primary-metric differences. Intervals are 95% Student-t intervals over three paired seeds and should be interpreted cautiously at $n = 3$.

Dataset	Comparison	Mean difference	95% CI
BEN-MM	1x1 Conv 4-to-3 - RGB probe	1.06	[1.05, 1.06]
BEN-MM	Scratch PFU - RGB probe	3.00	[2.69, 3.30]
BEN-MM	Scratch PFU - 1x1 Conv 4-to-3	1.94	[1.64, 2.25]
BEN-MM	PFU-E init - Scratch PFU	0.49	[0.12, 0.85]
So2Sat	Scratch PFU - RGB probe	3.46	[1.02, 5.89]
So2Sat	PFU-E init - Scratch PFU	0.19	[-2.10, 2.47]

5.5 PFU Uses Aligned NIR Signal

Table 8: Test-time NIR interventions. Values are mean ± sample standard deviation over three seeds. Zeroing is an out-of-distribution mask; sample shuffling preserves the marginal NIR distribution while breaking sample correspondence.

Dataset	Method	Full	NIR shuffled	NIR zeroed
BEN-MM	RGB probe	83.464 ± 0.001	83.464 ± 0.001	83.464 ± 0.001
	1x1 Conv 4-to-3	84.519 ± 0.002	73.181 ± 0.030	77.419 ± 0.010
	Scratch PFU	86.461 ± 0.125	54.650 ± 0.483	22.836 ± 1.651
	PFU-E init	86.946 ± 0.076	54.912 ± 1.164	23.818 ± 0.503
So2Sat	RGB probe	65.955 ± 0.158	65.955 ± 0.158	65.955 ± 0.158
	Scratch PFU	69.411 ± 0.837	57.582 ± 0.719	16.543 ± 2.913
	PFU-E init	69.599 ± 0.245	57.505 ± 0.879	18.980 ± 2.725

The equal-capacity training control shows that the gain is not explained by frontend size: enabling B08 improves both datasets. Test-time interventions then show how strongly the trained representation depends on alignment. B08 shuffling lowers BEN-MM scratch PFU from 86.46 to 54.65 micro-mAP and So2Sat from 69.41 to 57.58 accuracy, while the RGB probe is invariant. Zeroing is even more destructive. PFU is therefore not merely using NIR as a global intensity prior; it learns sample-specific RGB–NIR correspondences. The one-seed capacity-matched training contrast is the appropriate estimate of incremental spectral access, while the perturbations diagnose dependence after training.

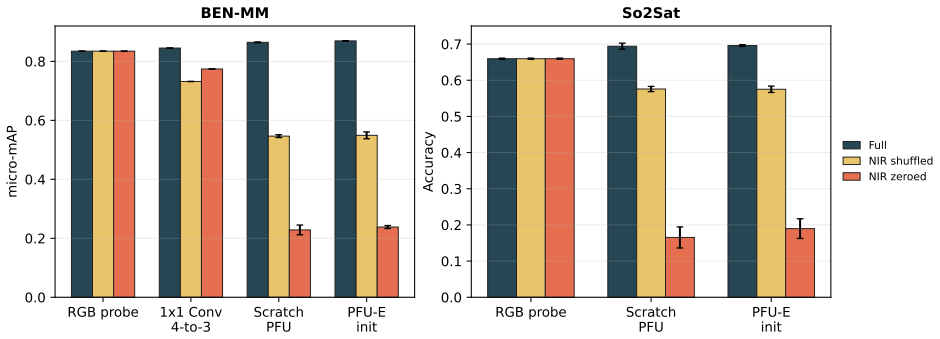


Figure 3: Test-time NIR dependence over three predetermined seeds. Shuffling breaks sample correspondence while approximately preserving the B08 marginal; zeroing measures missing-band robustness.

5.6 Parameter and Compute Accounting

Table 9: Resource accounting for the matched runs. GPU time and peak reserved memory are means over three seeds on one RTX PRO 6000 Blackwell GPU.

Dataset	Method	Trainable	Backbone share	GPU h	Peak GiB
BEN-MM	RGB probe	0.019M	0.006%	7.57	2.64
	1x1 Conv 4-to-3	0.019M	0.006%	12.19	9.70
	Scratch PFU	4.899M	1.616%	14.71	14.52
	PFU-E init	4.899M	1.616%	14.69	14.52
So2Sat	RGB probe	0.017M	0.006%	6.54	2.03
	Scratch PFU	4.897M	1.615%	14.59	14.21
	PFU-E init	4.897M	1.615%	14.59	14.21

The 4.879M frontend is 1.61% of DINOv3-LVD and leaves all 303.2M backbone parameters untouched. This deployment advantage does not make training free: matched PFU runs use 14.6–14.7 GPU-hours and 14.2–14.5 GiB peak reserved memory, compared with 6.5–7.6 GPU-hours and 2.0–2.6 GiB for RGB probes. The trade-off is therefore a stable backbone interface and a small update surface in exchange for an additional image-generation pass. PFU-E adds 34.2 aggregate GPU-hours of pretraining and is not required for the scratch-PFU gains.

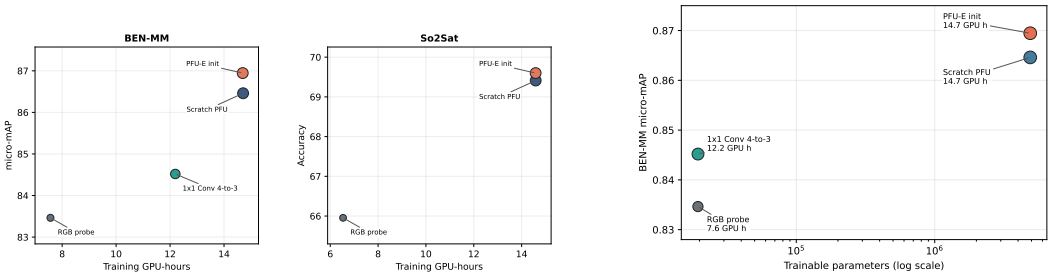


Figure 4: Efficiency views from the matched three-seed runs. Left: performance versus training GPU-hours on both datasets, with marker area encoding mean peak reserved memory. Right: BEN-MM performance versus trainable parameters, with GPU-hours annotated. PFU is parameter-efficient but uses more time and memory than RGB probing; neither panel measures inference latency or FLOPs.

6 Limitations

The formal classification block uses three seeds, but the capacity-matched RGB-only control, the So2Sat 1×1 mapping, PASTIS-HD, PFU-D, PFU-I, and BEN input-scaling experiments are single runs. The extension tables therefore show breadth and observed configuration gaps rather than uncertainty-aware rankings. In particular, the original PFU-I SAT segmentation contrast changes both the input path and decoder, so it cannot isolate the dual-path contribution.

The current architecture still lacks an inflated four-channel DINOv3 patch embedding, a capacity-matched pointwise MLP, and protocol-matched SpectraDINO or HSI-Adapter baselines. PFU-E has a z-score/reflectance pretraining mismatch, no held-out pretraining validation, and no loss-component ablation; its best three-seed downstream mean supports the initialization empirically but does not identify which pretraining component causes the gain. Resolution-preserving processing, alternative token pooling, multi-seed dense experiments, and a fully matched PFU-I decoder study remain the most important next tests.

7 Conclusion

PFU provides a reusable multispectral interface for frozen RGB foundation models. Under the formal three-seed protocol, PFU-E improves over RGB by 3.482 micro-mAP points on BEN-MM and 3.644 accuracy points on So2Sat, while scratch PFU confirms that most of the gain does not require external initialization. PASTIS-HD, 12-band PFU-D, PFU-I dual-path, and BEN S1+S2 results extend the same principle to dense prediction, wider spectra, two backbone families, and multimodal inputs. These extensions are promising single-run evidence rather than a universal ranking; matched multi-seed dense and decoder-controlled experiments are the next requirement.

AI use statement

Generative AI tools (GPT) were used for code generation, experiment orchestration and log analysis, literature search, and drafting and editing the text. All reported numbers were recomputed from local experiment artifacts. The authors remain responsible for the experimental design, interpretation, and final text.

References

Clay Foundation. Clay foundation model, 2024. Technical report.

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 Advances in Neural Information Processing Systems, 2022.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In International Conference on Machine Learning, 2019.

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

Juana Valeria Hurtado, Rohit Mohan, and Abhinav Valada. Hyperspectral adapter for semantic segmentation with vision foundation models. IEEE Robotics and Automation Letters, 11(3):3606–3613, 2026.

Johannes Jakubik, Sujit Roy, C. E. Phillips, et al. Foundation models for generalist geospatial artificial intelligence. arXiv preprint arXiv:2310.18660, 2023.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In European Conference on Computer Vision, 2022.

Dianbo Liu, Alex Lamb, Kenji Kawaguchi, Anirudh Goyal, Chen Sun, Michael Mozer, and Yoshua Bengio. ChannelViT: Multiple input channels for efficient vision transformer models. In International Conference on Learning Representations, 2023.

Oscar Mañas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vázquez, and Pau Rodriguez. CROMA: Remote sensing representations with contrastive radar-optical masked autoencoders. In Advances in Neural Information Processing Systems, 2023.

Peter Mansoor, Sabrina Nan Lin, Nathan Inkawhich, et al. OlmoEarth: A stable latent transformer for global, multi-modal, multi-temporal earth observation. arXiv preprint arXiv:2509.25125, 2025.

Yagiz Nalcakan, Hyeongjin Ju, Incheol Park, Sanghyeop Yeo, Youngwan Jin, and Shiho Kim. SpectraDINO: Modality-conditioned adaptation of rgb vision foundation models across infrared bands. arXiv preprint arXiv:2605.02258, 2026.

Colorado 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, 2023.

Daniel Scheibenreif, Michael Mommert, and Damian Borth. SpectralEarth: Training hyperspectral foundation models at scale. In Proceedings of the AAAI Conference on Artificial Intelligence, 2025.

Haozhe Si, Yuxuan Wan, Minh Do, Deepak Vasisht, Han Zhao, and Hendrik F. Hamann. Towards scalable foundation models for multimodal and hyperspectral geospatial data. arXiv preprint arXiv:2503.12843, 2025.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, et al. DINOv3. arXiv preprint arXiv:2508.10104, 2025.

Gencer Sumbul, Arne de Wall, Tristan Kreuziger, Filipe Marcelino, Hugo Costa, Pedro Benevides, Mario Caetano, Begüm Demir, and Volker Markl. BigEarthNet-MM: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval. IEEE Geoscience and Remote Sensing Magazine, 2021.

Yi Wang, Nils A. A. Braham, Zhitong Xiong, Chenying Liu, and Conrad M. Albrecht. SSL4EO-S12: A large-scale multimodal, multitemporal dataset for self-supervised learning in earth observation. IEEE Geoscience and Remote Sensing Magazine, 2023.

Zhitong Xiong, Yi Wang, Fahong Zhang, Adam J. Stewart, Joelle 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.

Xiao Xiang Zhu, Jingliang Hu, Chunping Qiu, et al. So2Sat LCZ42: A benchmark data set for the classification of global local climate zones. IEEE Geoscience and Remote Sensing Magazine, 2020.

A Architecture and Reproducibility Details

This appendix records the exact evaluated architecture, per-seed results, and extended experiments referenced by the main paper.

Table 10: Exact four-band PFU path used in all headline runs. GN denotes GroupNorm and all activations are GELU unless stated otherwise.

Stage	Operator	Channels	Output grid
Tower	$3 \times 3/2$ Conv–GN	$4 \to 96$	112^2
	$3 \times 3/1$ Conv–GN	$96 \to 96$	112^2
	$3 \times 3/2$ Conv–GN	$96 \to 192$	56^2
	$3 \times 3/1$ Conv–GN	$192 \to 192$	56^2
	1×1 Conv–GN	$192 \to 384$	56^2
Fusion post	3×3 then 1×1	$384 \to 384$	56^2
Projection	3×3 then 1×1	$384 \to 512$	56^2
RGB head	3×3, 1×1, sigmoid	$512 \to 256 \to 3$	56^2
Output	bilinear resize	3	224^2

The backbone uses a 224^2 input, ImageNet mean $(0.485, 0.456, 0.406)$ and standard deviation $(0.229, 0.224, 0.225)$, and the DINOv3 class token. BEN-MM uses binary cross entropy; So2Sat uses cross entropy. Validation is evaluated at the declared cadence (every epoch for the original headline runs and every five epochs for the BF16 controls), and the primary-metric checkpoint is restored for test. All raw result JSONs, histories, configurations, logs, generated CSVs, and source paths are indexed in the supplementary evidence manifest.

Table 11: Primary metric for every predetermined seed. BEN-MM reports micro-mAP and So2Sat reports accuracy. No seed is omitted or substituted.

Dataset	Method	Seed 0	Seed 1	Seed 2
BEN-MM	RGB probe	83.463	83.463	83.465
	1x1 Conv 4-to-3	84.517	84.518	84.520
	Scratch PFU	86.327	86.481	86.574
	PFU-E init	86.914	87.033	86.892
So2Sat	RGB probe	65.982	66.098	65.785
	Scratch PFU	69.746	68.458	70.029
	PFU-E init	69.319	69.704	69.775

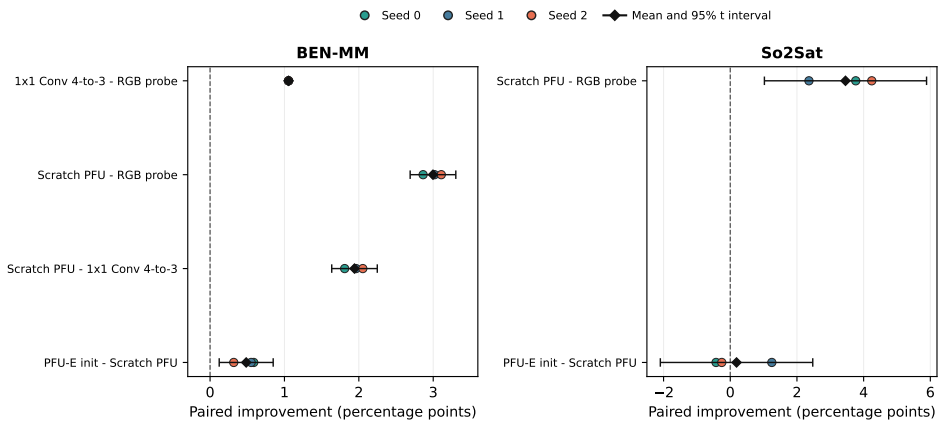


Figure 5: Per-seed paired primary-metric differences and descriptive 95% Student-t intervals. With only three paired seeds, the intervals have two degrees of freedom and are not confirmatory.

B Extended Spectral and Backbone Results

Table 12: Retained single-run scope checks. RGB/PFU classification pairs use 20 epochs within each backbone, but seeds were not recorded. PFU-E plus VPT-Shallow uses 30 epochs and is not protocol-matched to those pairs. PASTIS-HD PFU rows are matched to each other at seed 0 with VPT-Deep, not to the RGB row in trainable surface. All rows are excluded from headline claims.

Task	Dataset	Backbone	Method	Input/PEFT	Epochs	Seed	Metric
Classification	BEN-MM	DINOv3-LVD	RGB probe	RGB	20	–	83.15
Classification	BEN-MM	DINOv3-LVD	Scratch PFU	RGB+NIR	20	–	85.40
Classification	BEN-MM	DINOv3-SAT	RGB probe	RGB	20	–	84.82
Classification	BEN-MM	DINOv3-SAT	Scratch PFU	RGB+NIR	20	–	82.97
Classification	So2Sat	DINOv3-LVD	RGB probe	RGB	20	–	65.52
Classification	So2Sat	DINOv3-LVD	Scratch PFU	RGB+NIR	20	–	68.98
Classification	So2Sat	DINOv3-SAT	RGB probe	RGB	20	–	67.04
Classification	So2Sat	DINOv3-SAT	Scratch PFU	RGB+NIR	20	–	67.59
Classification	BEN-MM	DINOv3-LVD	PFU-E + VPT-Shallow	RGB+NIR; VPT-Shallow	30	–	87.47
Classification	So2Sat	DINOv3-LVD	PFU-E + VPT-Shallow	RGB+NIR; VPT-Shallow	30	–	66.94
Segmentation	PASTIS-HD	DINOv3-LVD	RGB probe	RGB	30	0	22.03
Segmentation	PASTIS-HD	DINOv3-LVD	Scratch PFU	RGB+NIR; VPT-Deep	30	0	38.58
Segmentation	PASTIS-HD	DINOv3-LVD	PFU-E	RGB+NIR; VPT-Deep	30	0	41.94

Table 13: Historical BEN-MM spectral configurations. All rows use 30 epochs and VPT-Deep; seeds were not recorded. Bold marks the best observed value in this exploratory block.

Configuration	Transfer	micro-mAP	Macro-F1	Protocol status
4-band single tower	PFU-E	88.54	72.82	single historical run
12-band single tower	PFU-E	89.30	73.44	single historical run
3+7 two towers	PFU-E	88.78	72.29	single historical run
4+8 two towers, scratch	Scratch	88.63	72.69	single historical run
4+8 two towers, PFU-E	PFU-E	88.82	72.26	single historical run
3+9 two towers, scratch	Scratch	88.67	72.74	single historical run
3+9 two towers, PFU-E	PFU-E	88.93	72.76	single historical run

Table 14: Exploratory PASTIS-HD results. Bold marks the best observed value, not a multi-seed controlled result. Rows with different bands, epochs, or PEFT are not directly comparable.

Backbone	Method	Input	PEFT	Epochs	Seed	mIoU
DINOv3-LVD	RGB probe	RGB	None	30	0	22.03
DINOv3-LVD	RGB probe + LayerNorm	RGB	LayerNorm	30	0	27.27
DINOv3-LVD	Scratch PFU + VPT-Deep	RGB+NIR	VPT-Deep	30	0	38.58
DINOv3-LVD	PFU-E + VPT-Deep	RGB+NIR	VPT-Deep	30	0	41.94
DINOv3-LVD	Scratch PFU	10 S2 bands	None	40	0	41.61
DINOv3-SAT	Scratch PFU	10 S2 bands	None	40	0	12.62
V-JEPA	Scratch PFU	10 S2 bands	None	40	0	40.09
AIMv2	Scratch PFU	10 S2 bands	None	40	0	37.80

The extended experiments expose two consistent patterns. First, broader spectra help when the protocol can exploit them: the 12-band BEN-MM PFU-E/VPT run is the best observed classification result, while the 10-band PASTIS-HD runs transfer to DINOv3-LVD, V-JEPA, and AIMv2. Second, the backbone remains the dominant interaction term: DINOv3-SAT is strong as an RGB probe on BEN-MM but degrades after pseudo-RGB remapping and fails sharply on the retained PASTIS-HD setup. These runs motivate backbone-specific pooling and interface selection rather than one universal pseudo-RGB recipe.

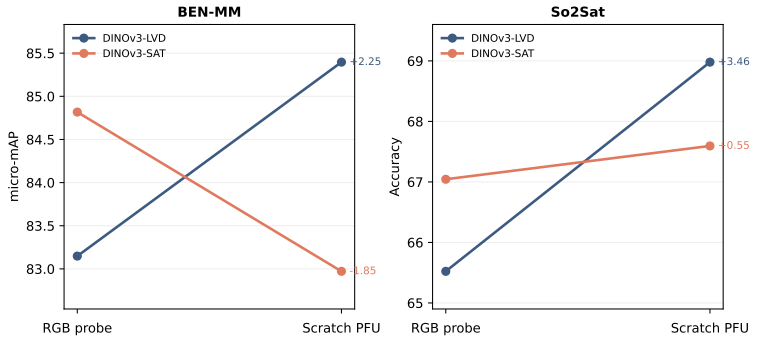


Figure 6: Retained RGB-to-PFU changes for DINOv3-LVD and DINOv3-SAT. PFU helps the natural-image LVD backbone on both classification datasets, whereas the satellite-specialized backbone is mixed.

C PFU-E Pretraining Details

PFU-E uses the same downstream architecture and parameter count as scratch PFU. Its initialization was trained on the OLMoEarth 10_sentinel2_l2a_monthly stream with a frozen DINOv3-LVD teacher:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + 0.25\mathcal{L}_{\text{patch}} + 0.05\mathcal{L}_{\text{RGB}} + \mathcal{L}_{\text{temporal}}. \tag{2}$$

Token terms are cosine distances, the RGB target is $\text{clip}(3.5[B04, B03, B02], 0, 1)$, and the RGB term is ℓ_1. Pretraining uses 50,000 micro-steps on 8 GPUs, per-rank batch size 1, gradient accumulation 8, AdamW with learning rate and weight decay 10^{-4}, and gradient clipping at 1. This is 34.2 aggregate GPU-hours. Step 31,243 was selected by the lowest 50-step smoothed training loss. Student inputs were z-scored during pretraining, whereas downstream inputs are reflectance; correcting this mismatch and selecting checkpoints on held-out validation are prerequisites for a stronger PFU-E claim.

D Language-Backbone Stress Test

The local archive also contains a 10-epoch exploratory interface to frozen Qwen2.5-7B token features on a 43-label BigEarthNet variant. A pretrained language backbone reaches 29.96–34.74 mAP across two retained runs, while a random-weight control reaches 43.49 mAP and the no-LLM frontend reaches 25.01. The random control outperforming the pretrained model is a negative result: a large frozen backbone is not useful merely because an interface can match its input dimension. We therefore use the LLM experiment only as evidence that backbone semantics and pooling must match the downstream visual task.

E Data Audit Artifact

Running audit_data_integrity.py –hash-so2-content regenerates data_integrity_audit.json. The audit records metadata hashes, split counts, path intersections, index ranges, missing files, TIFF georeference probes, content-hash duplicate groups, and the retained OLMoEarth index status. In summary, 400,673/400,673 So2Sat paths are present and unique; 400,592 contents are unique; the 81 extra files occur in 77 same-label groups within test only; and no exact content hash crosses a split. The complete, disjoint train/validation source-index range is 0–352,365. These checks do not resolve nearby-scene overlap because geographic identifiers and OLMoEarth draw IDs were not retained. Running build_revision_assets.py regenerates the quantitative tables and original-result figures from the 21 headline runs, 15 reviewer controls, and retained scope-check JSONs. Running plot_revision_figures.py regenerates the reviewer-control, audit, cross-backbone, and efficiency figures without model inference or training.