AUTOREGRESSIVE EMBEDDING PREDICTION FOR TEMPORAL EARTH OBSERVATION FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-spectral remote sensing time series contain rich temporal and spectral signals. However, existing remote sensing foundation models often struggle to learn robust temporal semantics from these data due to their sensitivity to observation-specific noise and insufficient modeling of fine-grained temporal structure. We propose AREO, a feature-predictive foundation model that addresses these limitations by learning semantic representations while preserving fine-grained temporal structure throughout the prediction process. We introduce Autoregressive Embedding Prediction, in which a causal predictor, combined with temporally consistent high-ratio masking, estimates each masked patch's representation from current and preceding context. This design avoids noisy observation-level reconstruction and instead encourages temporally stable, spatially localized semantic representations. Pretrained on SSL4EO-S12, AREO is evaluated under both frozen-backbone and full fine-tuning protocols across three temporal and two static datasets. It outperforms the baselines on the temporal benchmarks under both evaluation protocols and achieves competitive results on static tasks. These findings suggest that temporally directed embedding prediction can support transferable representations for Earth observation.

1 INTRODUCTION

The growing volume of Earth observation (EO) data creates unprecedented opportunities to monitor the Earth system, with applications spanning agricultural phenology and urban expansion (Bodnar et al., 2025; Mai et al., 2025; Zhang et al., 2024). Remote sensing foundation models increasingly support this progress by pre-training on large-scale EO data and adapting to diverse downstream tasks without extensive labeling. Multi-spectral time series are especially valuable because they capture both temporal dynamics and spectral signatures of the land surface. These signals support core remote-sensing tasks such as land-cover monitoring and anomaly detection (Cong et al., 2022; Guo et al., 2026; Li et al., 2026). In practice, however, such data are routinely corrupted by observation-specific nuisances such as clouds, atmospheric effects, and acquisition geometry. Existing foundation models aim to learn transferable representations from this data, yet a central question remains unresolved: *How can a model capture fine-grained temporal semantics while remaining robust to observation-specific noise?*

Existing pre-training objectives leave this question only partially addressed. Reconstruction-based methods, such as masked autoencoders (MAE), learn fine-grained representations by reconstructing local details. However, this objective also encourages the model to encode nuisance details that do not reflect the underlying land-surface state (Assran et al., 2023; Astruc et al., 2025). This is particularly problematic for Earth observation time series: land-surface states often evolve gradually, whereas observation-specific nuisances are typically transient. Feature-predictive methods provide a more semantic and noise-robust alternative by predicting in representation space rather than reconstructing raw observations. However, existing approaches typically define prediction across views or modalities rather than across time (Astruc et al., 2025; Danish et al., 2026; Tseng et al., 2025). Even when temporal inputs are used (Astruc et al., 2025), they are often compressed before prediction to support multimodal alignment, which weakens fine-grained modeling (Cao et al., 2026). This mo-

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.

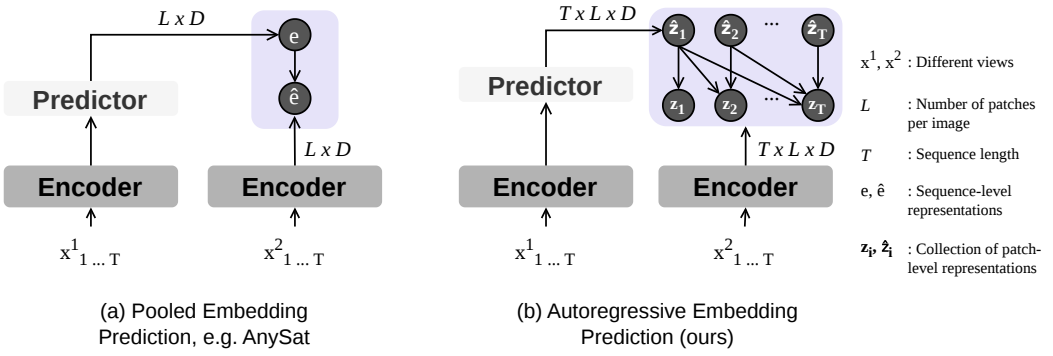
tivates a feature-predictive approach that preserves patch-level temporal structure while leveraging causal latent prediction to reduce sensitivity to observation-specific noise.

We propose AREO, a feature-predictive foundation model based on Autoregressive Embedding Prediction for Earth Observation. The proposed model preserves patch-level structure during prediction, thereby avoiding premature temporal aggregation during representation learning. At its core is Autoregressive Embedding Prediction (AEP), a strategy that trains the model to infer missing patch representations by using information from the current observation and preceding time steps. By conditioning prediction on both current and historical evidence, AREO is encouraged to recover temporally consistent semantics rather than reproduce noisy details from individual observations. Unlike prior feature-predictive approaches that aggregate temporal inputs before prediction, AREO preserves per-observation, patch-level representations throughout the training process. Operating in representation space improves robustness to observation-specific noise, while patch-level prediction retains localized temporal-spectral structure that sequence-level aggregation can obscure. As a result, AREO learns representations that are both semantically grounded and temporally informed.

Our contributions are summarized as follows:

- We propose AREO, a feature-predictive foundation model for temporal Earth observation that learns semantic representations while preserving fine-grained temporal structure during prediction.
- We introduce Autoregressive Embedding Prediction, a predictive objective that estimates the teacher-induced prototype distribution of the current observation from its masked temporal history, yielding representations that are both temporally informed and robust to transient observation noise.
- Experiments on five downstream datasets show that AREO improves over the strongest compared baseline on each temporal benchmark, with average gains of 2.2 and 1.4 mIoU points under frozen-backbone evaluation and full fine-tuning, respectively, while achieving competitive results on static tasks.

2 Related Work



(a) Pooled Embedding Prediction, e.g. AnySat

(b) Autoregressive Embedding Prediction (ours)

Figure 1: Comparison of feature-predictive architectures for temporal remote sensing imagery. (a) Prior models, *e.g.*, AnySat (Astruc et al., 2025), compress the input sequence into a single sequence-level representation, thereby weakening explicit temporal modeling. (b) In contrast, the proposed AEP keeps per-frame, patch-level representations throughout the prediction pipeline and applies autoregressive prediction along the temporal dimension. This enables the model to capture temporal dependencies and dynamic scene evolution more effectively.

Self-supervised Learning in Remote Sensing. Self-supervised learning has become the dominant paradigm for pre-training foundation models, leveraging the abundance of unlabeled data to learn transferable representations (He et al., 2022; Song et al., 2026). A prominent instantiation is masked image modeling, pioneered by masked autoencoder (MAE) (He et al., 2022; Gui et al., 2024), which trains a model to reconstruct randomly masked regions of an input image. Due to the large volume

of available imagery, MAE has been widely adopted in remote sensing (Hong et al., 2026; Jiao et al., 2023), with subsequent work exploring masking strategies across spatial (Cong et al., 2022; Reed et al., 2023; Tang et al., 2023), spectral (Hong et al., 2024; Xiong et al., 2025; Zhao et al., 2026), and temporal (Dumeur et al., 2024; Labatie et al., 2026; Szwarcman et al., 2026) scales for multi-spectral images. Beyond multi-spectral imagery, several models extend MAE to additional modalities such as SAR, DEM, *etc*., often combining it with complementary objectives, including contrastive learning (Fuller et al., 2023; Mañas et al., 2021; Wang et al., 2024) and multimodal fusion (Jakubik et al., 2025; Li et al., 2026) to handle the heterogeneity inherent in multimodal EO data.

Despite their flexibility and robustness across scales and modalities, MAE-based models are trained to reconstruct pixel-level observations in full detail, including sensor noise, cloud contamination, and atmospheric artifacts that are irrelevant to the underlying land-surface state. As a result, their representations tend to encode low-level statistics rather than high-level semantic content (Assran et al., 2023; Do et al., 2026), often requiring heavy task-specific fine-tuning to achieve strong performance on semantic downstream tasks (Szwarcman et al., 2026). This limitation motivates the exploration of pre-training objectives that explicitly target semantic structure rather than raw-signal reconstruction.

Feature-predictive Architectures in Remote Sensing. Joint-Embedding Predictive Architectures (JEPAs) have recently emerged as a promising alternative for self-supervised learning (Assran et al., 2023; 2025). Instead of reconstructing pixels, JEPA-style methods train a student network to predict teacher representations, encouraging the model to capture higher-level semantic structure (Assran et al., 2023; Maes et al., 2026). In remote sensing, SAR-JEPA (Li et al., 2024) is the first JEPA-based model designed for SAR imagery, while DINOv3Sat (Siméoni et al., 2025) and SATtxt (Do et al., 2026) focus on optical imagery. More recent models, such as Galileo (Tseng et al., 2025) and TerraFM (Danish et al., 2026), extend this paradigm to multimodal data.

Although these methods are promising, important limitations remain. Most JEPA-based remote sensing models are designed for single images or carefully curated multimodal inputs and often rely on multiple local-view variants during training (Danish et al., 2026; Siméoni et al., 2025; Zhou et al., 2022). These models (Astruc et al., 2025; Tseng et al., 2025) extend this setting to temporal inputs, but neither explicitly models temporal dynamics. For example, AnySat uses a Lightweight Temporal Attention Encoder within the patch encoder to collapse each temporal sequence into a single fixed-size embedding per patch before the main encoder, removing explicit temporal structure prior to representation learning. As shown in Figure 1, this design choice means that neither method explicitly models temporal dynamics within the JEPA framework, and both depend on curated multimodal pre-training setups, limiting robustness under single-modality inputs or temporal sampling. V-JEPA (Assran et al., 2025) preserves per-frame tokens but is tailored to dense, fast-evolving natural video, pairing a non-causal predictor with an L_1 loss on continuous EMA features. Both choices transfer poorly to EO time series: stable backgrounds across adjacent frames let a bidirectional predictor exploit static content as a shortcut, and transient acquisition noise makes continuous-target regression sensitive to observation-specific detail.

3 METHODOLOGY

3.1 OVERVIEW

Figure 2 illustrates the overview of AREO, illustrating its key components and the flow of information during training. The model consists of a student encoder that processes a masked sequence and, together with a causal predictor, is trained to match the patch-level prototype distributions produced by an EMA teacher encoder operating on the unmasked sequence. Overall, our method differs from existing predictive approaches in two ways. First, it preserves frame-level temporal information throughout the full pipeline, rather than collapsing the sequence into a single representation before prediction. Second, it introduces autoregressive embedding prediction, which trains the model to estimate teacher-induced semantic targets from masked temporal context.

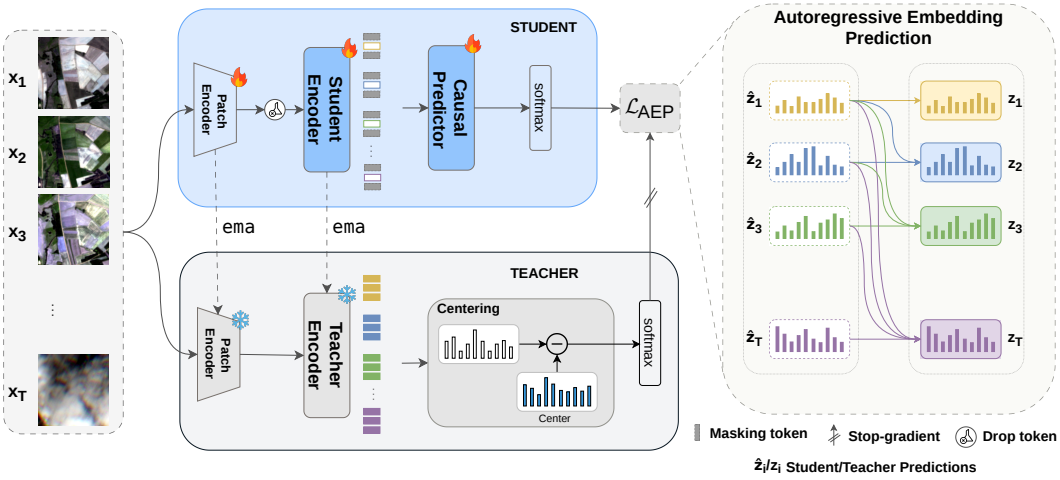


Figure 2: **Overview of AREO**. Our method preserves temporal information throughout the entire pipeline for multi-spectral remote sensing time series. The student encoder processes a consistently masked sequence with a high masking rate, while the teacher encoder operates on the corresponding unmasked sequence to provide target distributions. A causal predictor explicitly models dependencies from preceding observations, and an autoregressive distribution prediction objective aligns student predictions with teacher targets at masked patch locations.

3.2 Architecture

The model contains three key components *i.e.* Student Encoder (E_θ), Teacher Encoder ($E_{\bar{\theta}}$), and Causal Predictor. Given a multi-spectral sequence $\mathbf{x} = \{x_t\}_{t=1}^{T}$, where $x_t \in \mathbb{R}^{H\times W\times C}$, T is the number of temporal observations, $H \times W$ is the spatial resolution, and C is the number of spectral bands, our goal is to learn a representation that captures both spectral content and temporal evolution. We divide each frame into non-overlapping patches of size $P \times P$, producing $N = (\frac{H}{P})(\frac{W}{P})$ patches per frame. After patch embedding, the sequence is represented as a tensor of $T \times N$ spatiotemporal tokens with dimension D. Following Prithvi-2 (Szwarcman et al., 2026), we encode the acquisition date of each observation, enabling the predictor to distinguish irregular revisit times through calendar-aware temporal embeddings.

Student-Teacher Encoders. The student encoder E_θ processes a masked temporal sequence, while the teacher encoder $E_{\bar{\theta}}$ processes the corresponding unmasked sequence and provides stable prediction targets. The teacher parameters are updated by an exponential moving average (EMA) of the student parameters:

$$\bar{\theta} \leftarrow m\bar{\theta} + (1-m)\theta, \tag{1}$$

where $m \in [0, 1)$ is the EMA momentum. This momentum update makes the teacher evolve smoothly over training, stabilizing the target representations and avoiding direct optimization of the teacher branch. Given the unmasked input sequence $\mathbf{x}$, the teacher produces patch-level representations $\mathbf{H}^{\text{tea}} = E_{\bar{\theta}}(\mathbf{x}) \in \mathbb{R}^{T\times N\times D}$, where T is the number of observations, N is the number of patches per observation, and D is the embedding dimension. These representations serve as dense temporal-spatial targets for the student. In contrast, the student branch receives a masked version of the same sequence, forcing it to infer the missing patch representations from the visible context rather than directly copying the input.

Consistent Temporal Masking. We apply spatial masking and share the same masking indices across all temporal frames. Formally, for each sample we draw a visible patch set

$$\mathcal{V} \subset \{1, \ldots, N\}, \qquad |\mathcal{V}| = N_{\text{keep}}, \tag{2}$$

with $N_{\text{keep}} = \lfloor (1-r)N \rfloor$, where r is the masking ratio. The same visible index set $\mathcal{V}$ is used for all T observations. Similar to the MAE approach (Do et al., 2025; Hong et al., 2024), we adopt a high masking ratio (we use r = 0.9 in our experiments). Let $\mathbf{u}_{t,n} \in \mathbb{R}^D$ denote the embedded token for

patch n at time t. The student input is the set of visible tokens $\mathbf{U}^{\text{vis}} = \{\mathbf{u}_{t,n} \mid n \in \mathcal{V},\ t = 1, \ldots, T\}$. The student encoder processes these visible tokens and produces

$$\mathbf{H}^{\text{ctx}} = E_\theta(\mathbf{U}^{\text{vis}}) \in \mathbb{R}^{T \times N_{\text{keep}} \times D}. \tag{3}$$

Sharing the masking pattern across time ensures that a masked spatial location remains unobserved throughout the entire temporal sequence. Together with the 90% masking ratio, this restricts direct access to target trajectories and exposes only a small fraction of the remaining spatiotemporal context to the student. This limits shortcut solutions based on directly observed content and encourages the model to infer masked representations from broader spatial and temporal dependencies.

3.3 Autoregressive Embedding Prediction

Autoregressive Embedding Prediction learns masked patch representations by predicting semantic targets using contextualized tokens associated with the current and preceding temporal positions. This latent prediction objective preserves fine-grained temporal structure while avoiding direct reconstruction of observation-specific input details.

Causal Predictor. The causal predictor aims to recover dense spatiotemporal representations from the masked sequence. To this end, we insert learnable mask tokens at the masked spatial locations and restore the original token order within each frame. This yields a full token tensor $\tilde{\mathbf{H}} \in \mathbb{R}^{T \times N \times D}$. We then apply a causal predictor P_ϕ:

$$\hat{\mathbf{H}} = P_\phi(\tilde{\mathbf{H}}) \in \mathbb{R}^{T \times N \times D}. \tag{4}$$

Unlike prior methods that compress the temporal sequence before prediction, our predictor operates on the full frame-level token sequence. This preserves temporal information throughout the model and enables explicit temporal modeling at the representation level. The predictor uses transformer blocks with a temporal autoregressive attention mask. Within the predictor, token (t, n) may attend only to contextualized token positions associated with the same or earlier observations.

$$\mathcal{A}_{(t,n),(t',n')} = \begin{cases} 0, & t' \le t, \\ -\infty, & t' > t. \end{cases} \tag{5}$$

This attention pattern makes the prediction temporally directional: the prediction at position t may consume only contextualized tokens indexed at $t' \le t$. We adopt a causal rather than bidirectional design because adjacent frames in remote sensing time series often share large regions of stable content (e.g., unchanging land cover). Bidirectional attention would allow the predictor to exploit these overlaps as shortcuts, reducing the training loss without learning meaningful temporal dependencies. We empirically validate this choice in Section 4.4.

Following self-distillation methods, we project both student predictions and teacher targets into a shared prototype space. Let g_s and g_t denote the student and teacher heads. These heads produce

$$\mathbf{Z}^{\text{stu}} = g_s(\hat{\mathbf{H}}) \in \mathbb{R}^{T \times N \times K}, \qquad \mathbf{Z}^{\text{tea}} = g_t(\mathbf{H}^{\text{tea}}) \in \mathbb{R}^{T \times N \times K}, \tag{6}$$

where K is the number of prototypes. These prototype distributions define the training target and avoid reconstructing raw spectral values.

Objective. We train AREO by aligning the student prototype distribution with the teacher prototype distribution over all spatiotemporal tokens. The student logits are sharpened with temperature τ_s, while the teacher logits are centered and sharpened with temperature τ_t.

For each token (t, n), the student and teacher probability distributions are:

$$p^{\text{stu}}_{t,n} = \text{softmax}\left(\frac{\mathbf{Z}^{\text{stu}}_{t,n}}{\tau_s}\right), \qquad q^{\text{tea}}_{t,n} = \text{softmax}\left(\frac{\mathbf{Z}^{\text{tea}}_{t,n} - \mathbf{c}}{\tau_t}\right), \tag{7}$$

where $\mathbf{c}$ is a running center. The training loss is the average cross-entropy between teacher and student distributions, computed only over masked patch locations $\mathcal{M} = \{1, \ldots, N\} \setminus \mathcal{V}$:

$$\mathcal{L}_{\text{AEP}} = -\frac{1}{T|\mathcal{M}|} \sum_{t=1}^{T} \sum_{n \in \mathcal{M}} q^{\text{tea}}_{t,n} \cdot \log(p^{\text{stu}}_{t,n}). \tag{8}$$

To stabilize training and prevent collapse, we maintain a running center over teacher logits:

$$\mathbf{c} \leftarrow \lambda \mathbf{c} + (1-\lambda)\,\frac{1}{BTN}\sum_{b,t,n}\mathbf{Z}^{\text{tea}}_{b,t,n}, \tag{9}$$

where B is the batch size and λ is the center momentum.

4 EXPERIMENTS

4.1 PRETRAINING SETUP

Pretraining dataset. The model is pretrained on SSL4EO-S12 v1.1 (Blumenstiel et al., 2025a), a comprehensive global-scale remote sensing dataset designed for temporal Earth observation. This dataset is widely used for multimodal and temporal pre-training (Blumenstiel et al., 2025b; Jakubik et al., 2025; Li et al., 2026). It consists of approximately 1 million samples, each comprising four multi-seasonal temporal frames. Following the configuration established by Prithvi-2 (Szwarcman et al., 2026), we utilize six Sentinel-2 spectral bands: Blue (B2), Green (B3), Red (B4), Narrow NIR (B8A), SWIR-1 (B11), and SWIR-2 (B12). All input observations are processed at a spatial resolution of 224×224.

Implementation details. We adopt a ViT-L/16 backbone for pre-training and train the model for 100 epochs. The predictor is implemented as a 6-layer causal transformer. For the predictive objective, we use a prototype of size 8192. We study the effect of predictor depth and prototype size in the Appendix. Following DINO (Caron et al., 2021), the student and teacher temperatures are set to 0.1 and 0.04, respectively, while the center momentum used to update the teacher distribution is set to 0.9. Data augmentation includes resizing and normalization. We optimize the model with a peak learning rate of 1×10^{-4}, using a 10-epoch linear warm-up followed by cosine decay. The EMA momentum m follows a cosine schedule from 0.996 to 1. Training is performed in a distributed setting on eight NVIDIA H200 GPUs, with a batch size of 256 per GPU. Consistent with high-capacity MAE frameworks (Hong et al., 2024), we set the masking ratio to 90%. The total pre-training time is about 30 hours.

4.2 EVALUATION SETUP

Downstream datasets. We evaluate models on five land-cover datasets from PANGAEA (Marsocci et al., 2026): three temporal datasets (PASTIS (Fare Garnot & Landrieu, 2021), Crop Type Mapping (M Rustowicz et al., 2019), and DynamicEarthNet (Toker et al., 2022)) and two static datasets (m-BigEarthNet (Lacoste et al., 2023) and Sen1Floods11 (Bonafilia et al., 2020)). These datasets include cloud contamination and atmospheric variability, providing a realistic setting for evaluating the models. We aimed to include all temporal datasets in PANGAEA but excluded BioMassters because we could not reproduce its reported results, an issue also noted by (Jakubik et al., 2025).

Implementation details. We adopt two evaluation protocols: frozen-backbone evaluation and full fine-tuning. Although we report results under both settings, we primarily emphasize frozen-backbone evaluation, as it more directly assesses the transferability and quality of the learned representations while reducing the influence of model-specific factors. Under both protocols, a task-specific prediction head is trained for each downstream task. For temporal datasets, when a model does not support temporal inputs, we follow Galileo (Tseng et al., 2025) and perform multiple forward passes followed by feature aggregation. For all models, the extracted features are processed by a lightweight temporal attention encoder (L-TAE), and the temporally aggregated features are subsequently passed to a UPerNet decoder. We evaluated under the PANGAEA benchmark (Marsocci et al., 2026). The additional evaluation details are provided in the Appendix.

4.3 RESULTS

We compare AREO with recent remote sensing foundation models, including MAE-based models, temporally aware JEPAs (AnySat, Galileo), and the multimodal JEPA TerraFM.

As shown in Table 1, AREO performs favorably across temporal and static benchmarks under both evaluation protocols. The frozen-backbone results indicate strong representation transferability,

Table 1: Comparison of models on temporal and static remote sensing datasets under Frozen-Backbone evaluation (FB) and full Fine-tuning (FT). Metrics are mIoU ↑, except for m-bigearthnet which uses F1 ↑. **Bold**: best; underline: second best.

	Temporal datasets						Static datasets			
	PASTIS		Crop Type Mapping		DynamicEarthNet		m-bigearthnet		sen1floods11	
Model	FB	FT	FB	FT	FB	FT	FB	FT	FB	FT
DOFA (Xiong et al., 2025)	29.8	54.5	40.4	48.8	39.3	44.2	67.5	68.0	82.8	89.4
Satlas (Bastani et al., 2023)	28.0	57.4	40.0	46.2	30.1	34.2	63.5	72.8	83.9	88.2
SatMAE (Cong et al., 2022)	29.6	55.3	37.2	44.4	28.2	32.5	62.1	68.3	74.1	87.4
TerraFM (Danish et al., 2026)	34.5	58.6	41.1	50.7	39.3	46.2	67.8	**73.1**	85.2	89.6
Galileo (Tseng et al., 2025)	39.2	61.2	41.8	49.4	37.2	42.6	65.1	70.7	79.4	87.2
DEO (Wolf et al., 2026)	37.7	57.1	40.7	44.2	36.2	40.4	62.5	68.2	76.2	86.4
Prithvi-V2 (Szwarcman et al., 2026)	39.8	53.4	43.1	52.5	40.0	47.4	65.9	69.0	86.3	90.3
AnySat (Astruc et al., 2025)	39.2	58.2	45.1	50.2	40.4	44.9	63.8	70.3	77.9	88.2
AREO (ours)	**41.8** ± **0.65**	**63.2** ± **0.3**	**47.6** ± **0.4**	**54.1** ± **0.4**	**42.4** ± **0.27**	**48.1** ± **0.32**	**68.5** ± **0.35**	72.1 ± 0.42	**87.1** ± **0.17**	**91.2** ± **0.32**

while the fine-tuning results demonstrate effective downstream adaptation. Gains are generally more pronounced on temporal benchmarks, consistent with our motivation to preserve patch-level temporal structure within the JEPA framework. AREO also consistently outperforms the JEPA-based baselines AnySat (Astruc et al., 2025) and Galileo (Tseng et al., 2025) under both protocols. In addition, the relatively weaker performance of MIM-based models, including SatMAE and DOFA, on temporal benchmarks is consistent with prior findings that reconstruction-based objectives may emphasize low-level statistics rather than higher-level semantic content (Assran et al., 2023).

As illustrated in Figure 3, the temporal inputs often contain substantial cloud contamination, yet AREO maintains coherent parcel boundaries and predictions that are closer to the ground truth than competing baselines. Together with the quantitative results on naturally noisy downstream datasets, this suggests that AEP is less sensitive to transient observation-specific corruption.

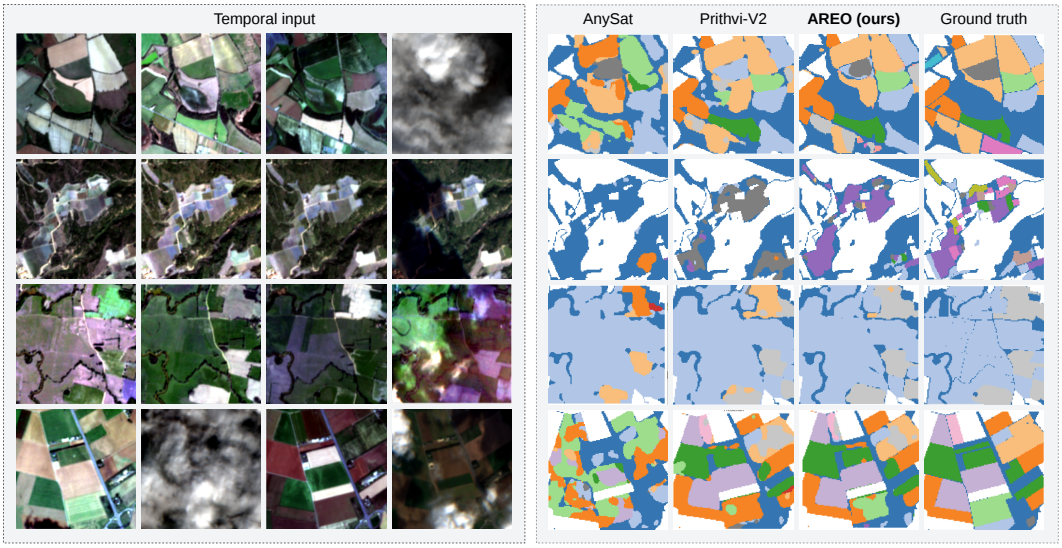


Figure 3: Qualitative segmentation results on PASTIS. Compared with baseline foundation models, AREO produces sharper field boundaries and more temporally coherent predictions, and remains robust under cloud-contaminated frames.

4.4 Ablation Study

We study the impact of two key design choices: the masking strategy and the proposed AEP.

Masking strategy. We analyze two factors in the masking design: the masking ratio and the masking method.

Masking ratio. Figure 4 reports the results for different masking ratios (25%, 50%, 75%, and 90%). We observe a consistent performance improvement as the masking ratio increases, with the best re-

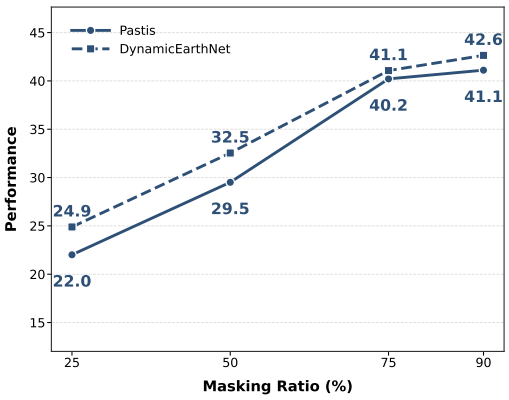


Figure 4: Ablation on different masking ratios under the full backbone frozen evaluation. A 90% masking ratio yields the highest overall mIoU.

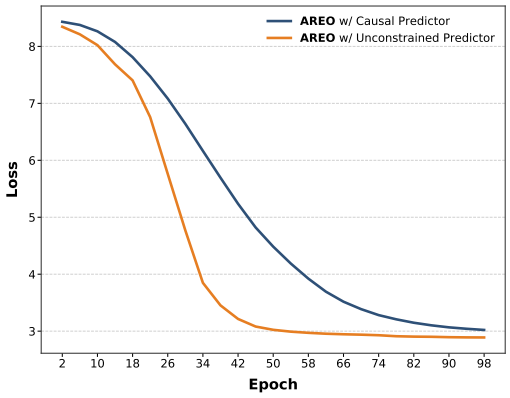


Figure 5: The unconstrained predictor converges faster by attending to encoded tokens from the full sequence. The causal predictor has slightly higher loss but restricts direct predictor attention to tokens indexed at the same or earlier timestamps, imposing directional prediction.

Table 2: Effect of masking ratio and masking method under frozen-backbone (FB) and full fine-tuning (FT) evaluation. Consistent masking outperforms independent masking in all settings.

Masking Ratio	Masking Method	PASTIS		DynamicEarthNet		Crop Type Mapping	
		FB	FT	FB	FT	FB	FT
75%	Independent Masking	35.2	54.2	36.8	41.6	40.0	46.1
	Consistent Masking	**40.2**	**60.8**	**41.1**	**46.4**	**44.5**	**51.3**
90%	Independent Masking	37.3	57.4	39.0	44.3	44.3	49.6
	Consistent Masking	**41.1**	**63.3**	**42.6**	**48.2**	**47.8**	**54.3**

sults achieved at 90%. This suggests that limiting the visible input to the student encoder encourages it to rely more on contextual and temporal information, rather than local cues, to predict the teacher representations. In other words, a higher masking ratio makes the prediction task more challenging, leading to the learning of more informative representations. This observation is consistent with findings in Masked Autoencoder (MAE) (Hong et al., 2024), where higher masking ratios are shown to improve representation quality by increasing task difficulty.

Masking method. We further compare two masking schemes: *independent masking* and *consistent masking*. In independent masking, the masked spatial locations differ across timestamps within a time series. In contrast, consistent masking applies the same masking pattern across all timestamps in the sequence. As shown in Table 2, consistent masking consistently outperforms independent masking across all settings. This improvement can be attributed to better temporal alignment: by masking the same spatial regions across time with a high masking ratio, the model is encouraged to learn coherent spatio-temporal representations and leverage temporal dependencies more effectively. In contrast, independent masking introduces misalignment across timestamps, making temporal reasoning more difficult and leading to inferior performance.

Impact of AEP. We isolate the contribution of AEP by comparing it with two external approaches and two alternative temporal prediction strategies. The external approaches are an MAE-based model, SatMAE (Cong et al., 2022), and V-JEPA 2 (Assran et al., 2025), denoted EO-V-JEPA-2, which retains per-frame tokens but uses a bidirectional predictor. Both are pre-trained on SSL4EO-S12 under the same configuration as AREO. Compressed Embedding Prediction (CEP) pools features across time and predicts a single sequence-level target, discarding frame-level temporal struc-

Table 3: Comparison of temporal prediction strategies under frozen-backbone (FB) and full fine-tuning (FT) evaluation.

	PASTIS		DynamicEarthNet		Crop Type Mapping	
Model	FB	FT	FB	FT	FB	FT
MAE-based model	30.2	56.2	28.1	31.9	35.8	42.1
EO-V-JEPA-2	30.4	58.0	30.3	34.3	38.5	44.6
AREO_{CEP}	21.9	47.8	22.1	34.2	25.3	36.5
AREO_{UEP}	29.5	52.3	34.4	41.2	37.0	50.1
AREO_{AEP}	**41.1**	**63.3**	**42.6**	**48.2**	**47.8**	**54.3**

ture. Unconstrained Embedding Prediction (UEP) replaces the causal predictor with a full-attention transformer, allowing each position to attend to tokens from later observations.

As shown in Table 3, AREO_{AEP} consistently outperforms all alternatives across datasets and evaluation protocols. The weaker performance of EO-V-JEPA-2 indicates that retaining per-frame tokens alone does not explain these gains. On PASTIS, the MAE baseline surpasses both CEP and UEP under both protocols, suggesting that embedding prediction is insufficient when temporal structure is either collapsed or modeled non-causally. Overall, these results attribute AEP's advantage to causal, temporally structured prediction with high masking ratio rather than to embedding prediction alone.

Figure 5 further shows that UEP converges substantially faster during pre-training, as each position can aggregate encoded representations from the entire sequence, including later observations. AEP removes this pathway, combined with the 90% masking ratio, this makes the prediction task more challenging and encourages representations that support directional temporal prediction.

5 Limitations

There are two main limitations of AREO. First, AREO is pretrained exclusively on multi-spectral imagery and does not incorporate complementary modalities such as SAR or DEM. Although this design choice isolates the contribution of temporal–spectral representation learning under a single-modality regime, it may constrain transferability to downstream tasks where multimodal fusion provides decisive complementary signal, such as flood mapping under persistent cloud cover or biomass estimation in densely vegetated areas. Extending the AEP objective to heterogeneous multimodal inputs, where observations across modalities are often asynchronously acquired or partially missing, remains an open direction. Second, although our evaluation follows the well-established PANGAEA benchmark, it is confined to predictive tasks defined over existing time-series observations. A promising extension is to situate AEP within a generative world-model framework capable of supporting dynamic tasks, thereby broadening the applicability of temporally predictive pre-training.

6 Conclusion

We have presented AREO, a feature-predictive foundation model for multi-spectral remote sensing time series that retains per-patch embeddings throughout prediction. At its core is AEP, a predictive strategy in which the student predicts the teacher-induced prototype distribution of each observation from its current and preceding temporal context. This formulation aligns the predictive target with the temporal regularity of the land surface while remaining robust to transient acquisition corruptions. Pretrained on SSL4EO-S12 and evaluated with frozen features, AREO achieves the best performance across three temporal and two static benchmarks, with the largest gains observed on temporal tasks. Future work will extend AREO to multimodal inputs and integrate it into generative world models for prediction, planning, and reasoning about Earth-system dynamics.

AI use statement

In this work, we used generative AI tools for writing polish, improving clarity and readability. We have not used generative AI tools to originate the research idea, develop the core scientific contribu-

tions, or formulate original arguments. Additionally, we used generative AI tools to provide feedback on presentation and aspects of the methodology. We have carefully reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility statement

We describe the pre-training configuration in Section 4.1 and the evaluation configuration in Section 4.2. In addition, we provide further details on the evaluation procedure and downstream datasets in the Appendix. The model and evaluation scripts are available at `https://anonymous.4open.science/r/jepa-pangaea-bench-79CC`.

References

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629, 2023. doi: 10.1109/CVPR52729.2023.01499.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba, Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning, 2025. URL `https://arxiv.org/abs/2506.09985`.

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

Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, and Aniruddha Kembhavi. Satlaspretrain: A large-scale dataset for remote sensing image understanding. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16726–16736, 2023. doi: 10.1109/ICCV51070.2023.01538.

Benedikt Blumenstiel, Nassim Ait Ali Braham, Conrad M Albrecht, Stefano Maurogiovanni, and Paolo Fraccaro. Ssl4eo-s12 v1. 1: A multimodal, multiseasonal dataset for pretraining, updated. *arXiv preprint arXiv:2503.00168*, 2025a.

Benedikt Blumenstiel, Paolo Fraccaro, Valerio Marsocci, Johannes Jakubik, Stefano Maurogiovanni, Mikolaj Czerkawski, Rocco Sedona, Gabriele Cavallaro, Thomas Brunschwiler, Juan Bernabe-Moreno, and Nicolas Longépé. Terramesh: A planetary mosaic of multimodal earth observation data, 2025b. URL `https://arxiv.org/abs/2504.11172`.

Cristian Bodnar, Wessel P Bruinsma, Ana Lucic, Megan Stanley, Anna Allen, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A Weyn, Haiyu Dong, et al. A foundation model for the earth system. *Nature*, 641(8065):1180–1187, 2025.

Derrick Bonafilia, Beth Tellman, Tyler Anderson, and Erica Issenberg. Sen1floods11: A georeferenced dataset to train and test deep learning flood algorithms for sentinel-1. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

Bingyi Cao, Koert Chen, Kevis-Kokitsi Maninis, Kaifeng Chen, Arjun Karpur, Ye Xia, Sahil Dua, Tanmaya Dabral, Guangxing Han, Bohyung Han, et al. Tipsv2: Advancing vision-language pretraining with enhanced patch-text alignment. *arXiv preprint arXiv:2604.12012*, 2026.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David B. Lobell, and Stefano Ermon. Satmae: pre-training transformers for temporal and multi-spectral satellite imagery. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Muhammad Sohail Danish, Muhammad Akhtar Munir, Syed Roshaan Ali Shah, Muhammad Haris Khan, Rao Muhammad Anwer, Jorma Laaksonen, Fahad Shahbaz Khan, and Salman Khan. TerraFM: A scalable foundation model for unified multisensor earth observation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=cBxuzdUDNx.

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

Minh Kha Do, Wei Xiang, Kang Han, Di Wu, Khoa Phan, Yi-Ping Phoebe Chen, Gaowen Liu, and Ramana Rao Kompella. Spectrally distilled representations aligned with instruction-augmented llms for satellite imagery, 2026. URL https://arxiv.org/abs/2602.22613.

Iris Dumeur, Silvia Valero, and Jordi Inglada. Self-supervised spatio-temporal representation learning of satellite image time series. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:4350–4367, 2024. doi: 10.1109/JSTARS.2024.3358066.

Vivien Sainte Fare Garnot and Loic Landrieu. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4852–4861, 2021. doi: 10.1109/ICCV48922.2021.00483.

Anthony Fuller, Koreen Millard, and James R. Green. Croma: remote sensing representations with contrastive radar-optical masked autoencoders. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9052–9071, 2024. doi: 10.1109/TPAMI.2024.3415112.

Zhongbin Guo, Yuhao Wang, Ping Jian, Chengzhi Li, Xinyue Chen, Zhen Yang, and Ertai E. TAMMs: Change understanding and forecasting in satellite image time series with temporal-aware multimodal models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=PXDFTeIqMd.

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 (CVPR)*, pp. 16000–16009, June 2022.

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 Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5227–5244, 2024. doi: 10.1109/TPAMI.2024.3362475.

Danfeng Hong, Chenyu Li, Xuyang Li, Gustau Camps-Valls, and Jocelyn Chanussot. Foundation models in remote sensing: Evolving from unimodality to multimodality. *IEEE Geoscience and Remote Sensing Magazine*, pp. 2–27, 2026. doi: 10.1109/MGRS.2026.3669086.

Johannes Jakubik, Felix Yang, Benedikt Blumenstiel, Erik Scheurer, Rocco Sedona, Stefano Maurogiovanni, Jente Bosmans, Nikolaos Dionelis, Valerio Marsocci, Niklas Kopp, Rahul Ramachandran, Paolo Fraccaro, Thomas Brunschwiler, Gabriele Cavallaro, Juan Bernabe-Moreno, and Nicolas Longépé. Terramind: Large-scale generative multimodality for earth observation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7383–7394, October 2025.

Licheng Jiao, Zhongjian Huang, Xiaoqiang Lu, Xu Liu, Yuting Yang, Jiaxuan Zhao, Jinyue Zhang, Biao Hou, Shuyuan Yang, Fang Liu, Wenping Ma, Lingling Li, Xiangrong Zhang, Puhua Chen, Zhixi Feng, Xu Tang, Yuwei Guo, Dou Quan, Shuang Wang, Weibin Li, Jing Bai, Yangyang Li, Ronghua Shang, and Jie Feng. Brain-inspired remote sensing foundation models and open problems: A comprehensive survey. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:10084–10120, 2023. doi: 10.1109/JSTARS.2023.3316302.

Antoine Labatie, Michael Vaccaro, Nina Lardiere, Anatol Garioud, and Nicolas Gonthier. Maestro: Masked autoencoders for multimodal, multitemporal, and multispectral earth observation data. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 212–224, March 2026.

Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan David Sherwin, Hannah Kerner, Björn Lütjens, Jeremy Irvin, David Dao, Hamed Alemohammad, Alexandre Drouin, Mehmet Gunturkun, Gabriel Huang, David Vazquez, Dava Newman, Yoshua Bengio, Stefano Ermon, and Xiao Xiang Zhu. Geo-bench: toward foundation models for earth monitoring. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Weijie Li, Wei Yang, Tianpeng Liu, Yuenan Hou, Yuxuan Li, Zhen Liu, Yongxiang Liu, and Li Liu. Predicting gradient is better: Exploring self-supervised learning for sar atr with a joint-embedding predictive architecture. *ISPRS Journal of Photogrammetry and Remote Sensing*, 218:326–338, 2024. ISSN 0924-2716. doi: https://doi.org/10.1016/j.isprsjprs.2024.09.013. URL https://www.sciencedirect.com/science/article/pii/S0924271624003514.

Xuyang Li, Chenyu Li, Gemine Vivone, and Danfeng Hong. Seamo: A season-aware multimodal foundation model for remote sensing. *Information Fusion*, 125:103334, 2026. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2025.103334. URL https://www.sciencedirect.com/science/article/pii/S1566253525004075.

Rose M Rustowicz, Robin Cheong, Lijing Wang, Stefano Ermon, Marshall Burke, and David Lobell. Semantic segmentation of crop type in africa: A novel dataset and analysis of deep learning methods. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition workshops*, pp. 75–82, 2019.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. Leworldmodel: Stable end-to-end joint-embedding predictive architecture from pixels, 2026. URL https://arxiv.org/abs/2603.19312.

Gengchen Mai, Yiqun Xie, Xiaowei Jia, Ni Lao, Jinmeng Rao, Qing Zhu, Zeping Liu, Yao-Yi Chiang, and Junfeng Jiao. Towards the next generation of geospatial artificial intelligence. *International Journal of Applied Earth Observation and Geoinformation*, 136:104368, 2025. ISSN 1569-8432. doi: https://doi.org/10.1016/j.jag.2025.104368. URL https://www.sciencedirect.com/science/article/pii/S1569843225000159.

Valerio Marsocci, Yuru Jia, Georges Le Bellier, David Kerekes, Liang Zeng, Sebastian Hafner, Sebastian Gerard, Eric Brune, Ritu Yadav, Ali Shibli, Heng Fang, Yifang Ban, Maarten Vergauwen, Nicolas Audebert, and Andrea Nascetti. Pangaea: Assessing geospatial foundation models capabilities through a global and inclusive benchmark. *IEEE Geoscience and Remote Sensing Magazine*, 14(1):245–285, 2026. doi: 10.1109/MGRS.2025.3628194.

Oscar Mañas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vazquez, and Pau Rodríguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9394–9403, 2021. doi: 10.1109/ICCV48922.2021.00928.

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 *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4065–4076, Los Alamitos, CA, USA, October 2023. IEEE Computer Society. doi: 10.1109/ICCV51070.2023.00378. URL https://doi.ieeecomputersociety.org/10.1109/ICCV51070.2023.00378.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

Qiya Song, Yiqiang Xie, Yuan Sun, Renwei Dian, and Xudong Kang. Robust remote sensing image-text retrieval with noisy correspondence. *arXiv preprint arXiv:2603.28134*, 2026.

Daniela Szwarcman, Sujit Roy, Paolo Fraccaro, orsteinn Elí Gíslason, Benedikt Blumenstiel, Rinki Ghosal, Pedro Henrique de Oliveira, Joao Lucas de Sousa Almeida, Rocco Sedona, Yanghui Kang, Srija Chakraborty, Sizhe Wang, Carlos Gomes, Ankur Kumar, Vishal Gaur, Myscon Truong, Denys Godwin, Sam Khallaghi, Hyunho Lee, Chia-Yu Hsu, Ata Akbari Asanjan, Besart Mujeci, Disha Shidham, Rufai Omowunmi Balogun, Venkatesh Kolluru, Trevor Keenan, Paulo Arevalo, Wenwen Li, Hamed Alemohammad, Pontus Olofsson, Timothy Mayer, Christopher Hain, Robert Kennedy, Bianca Zadrozny, David Bell, Gabriele Cavallaro, Campbell Watson, Manil Maskey, Rahul Ramachandran, and Juan Bernabe Moreno. Prithvi-eo-2.0: A versatile multitemporal foundation model for earth observation applications. *IEEE Transactions on Geoscience and Remote Sensing*, 64:1–20, 2026. doi: 10.1109/TGRS.2025.3642610.

Maofeng Tang, Andrei Cozma, Konstantinos Georgiou, and Hairong Qi. Cross-scale mae: A tale of multiscale exploitation in remote sensing. *Advances in Neural Information Processing Systems*, 36:20054–20066, 2023.

Aysim Toker, Lukas Kondmann, Mark Weber, Marvin Eisenberger, Andrés Camero, Jingliang Hu, Ariadna Pregel Hoderlein, Çağlar Şenaras, Timothy Davis, Daniel Cremers, Giovanni Marchisio, Xiao Xiang Zhu, and Laura Leal-Taixé. Dynamicearthnet: Daily multi-spectral satellite dataset for semantic change segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21126–21135, 2022. doi: 10.1109/CVPR52688.2022.02048.

Gabriel Tseng, Anthony Fuller, Marlena Reil, Henry Herzog, Patrick Beukema, Favyen Bastani, James R Green, Evan Shelhamer, Hannah Kerner, and David Rolnick. Galileo: Learning global & local features of many remote sensing modalities. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 60280–60300. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/tseng25a.html.

Yi Wang, Conrad M. Albrecht, and Xiao Xiang Zhu. Multilabel-guided soft contrastive learning for efficient earth observation pretraining. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024. doi: 10.1109/TGRS.2024.3466896.

Filip Wolf, Blaž Rolih, and Luka Čehovin Zajc. Brewing stronger features: Dual-teacher distillation for multispectral earth observation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27815–27826, June 2026.

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, 2025. URL https://arxiv.org/abs/2403.15356.

Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. Rs5m and georsclip: A large-scale vision- language dataset and a large vision-language model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–23, 2024. doi: 10.1109/TGRS.2024.3449154.

Boyu Zhao, Wei Li, Junjie Wang, Yuxiang Zhang, Hong Yang, Haitao Zhao, Ran Tao, and Qian Du. Alliance: All-in-one spectral-spatial-frequency awareness foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):3586–3603, 2026. doi: 10.1109/TPAMI.2025.3639595.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image BERT pre-training with online tokenizer. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=ydopy-e6Dg.

A Appendix

A.1 Additional Evaluation Details

The evaluated datasets contain different sets of spectral bands, and the supported input bands also vary across models. For each model, we use the bands it supports whenever they are available in the dataset. If a dataset contains fewer bands than a model requires, we pad the missing bands with zeros after applying the corresponding normalization. This ensures that all models receive inputs in their expected format while keeping the evaluation protocol consistent across datasets and baselines. We use the Adam optimizer with a learning rate of 1×10^{-4}, and all models are trained for 50 epochs with a default random seed of 234. For the main results (Table 1), we additionally run the proposed model with two further random seeds (2025 and 2026). The ablation results are reported in a single run (seed 234).

A.2 Additional Ablation

Predictor size and Prototype dimension. Table 4 studies the joint effect of causal predictor depth and prototype number K. Increasing the predictor depth from 4 to 6 layers consistently improves performance, indicating that sufficient predictor capacity is important for modeling temporal dependencies in the AEP objective. However, further increasing the depth to 8 layers does not yield consistent gains, suggesting diminishing returns from an overly expressive predictor. Similarly, increasing K from 4,096 to 8,192 improves the average performance across all predictor depths, while 16,384 prototypes generally leads to slightly lower results, likely due to increased optimization difficulty or overly fragmented target distributions. Due to computational constraints, we do not further scale the predictor depth or prototype dimension. Overall, the default configuration with a 6-layer predictor and $K = 8{,}192$ achieves the best average mIoU of 43.85, providing an effective balance between representation quality, predictor capacity, prototype granularity, and computational cost.

Table 4: Ablation on predictor depth and prototype number K. Avg. denotes the mean over studied datasets. The default AREO setting, marked with †, is the most effective overall.

Predictor Depth	Prototype Number K	PASTIS	DynamicEarthNet	Crop Type Mapping	Avg.
4	4,096	36.45	37.82	42.31	38.86
4	8,192	38.20	39.57	44.12	40.63
4	16,384	37.61	38.90	43.44	39.98
6	4,096	39.88	41.30	45.93	42.37
6	8,192†	41.11	**42.64**	47.79	**43.85**
6	16,384	**41.53**	41.92	46.88	43.44
8	4,096	38.47	39.88	44.53	40.96
8	8,192	40.29	41.73	**48.02**	43.35
8	16,384	39.77	41.05	46.14	42.32

A.3 Downstream Dataset Details

We evaluate on five downstream datasets that span temporal and static settings, providing a controlled basis for assessing generalization across remote sensing tasks. The temporal benchmarks comprise PASTIS, Crop Type Mapping, and DynamicEarthNet, which differ in spatial extent (regional to global), thematic focus (parcel-level agriculture vs. continuous land-cover change), and temporal resolution. The static benchmarks, m-bigearthnet and sen1floods11, target multi-label land-cover classification and binary flood segmentation, respectively. Together, the five datasets cover different task types and thematic domains (agriculture, land cover, and flood mapping). Table 5 summarizes their key properties.

Table 5: Overview of datasets used in our experiments.

Dataset	Type	Spectral Bands	# Bands	Domain	Location
PASTIS	Time-series	B2–B8, B8A, B11, B12	10	Agriculture	France
Crop Type Mapping	Time-series	B2–B8, B8A, B11, B12	10	Agriculture	South Sudan
DynamicEarthNet	Time-series	B2–B5	4	Land Cover	Global
m-bigearthnet	Static	B1–B8, B8A, B9, B11, B12	12	Land Cover	Global
sen1floods11	Static	B1–B8, B8A, B9–B12	13	Flood	Global