SAREO-FM: DECOUPLED SEMANTIC SUPERVISION FOR SAR-EO FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Synthetic aperture radar (SAR) and electro-optical (EO) imagery provide complementary observations: SAR enables day-and-night, weather-resilient sensing, whereas EO provides rich appearance and fine-grained semantic cues. We introduce **SAREO-FM**, which avoids forcing a single token stream to serve two distinct roles: modality tokens preserve how each sensor observes the scene through masked reconstruction, while learnable semantic queries capture what the scene contains under guidance from a pretrained vision foundation model (VFM). By jointly encoding these queries with SAR and EO tokens, the queries acquire modality-grounded semantic context, while the modality-token outputs remain the explicit targets of masked reconstruction. This design assigns semantic and reconstruction supervision to separate token streams while preserving their interaction within the shared encoder. Pretrained on the million-scale SAR-1M corpus, SAREO-FM achieves strong unimodal transfer for both SAR-only and EO-only inputs, while delivering substantial gains from joint SAR–EO observations on tasks that benefit from complementary sensing.

1 INTRODUCTION

Synthetic aperture radar (SAR) and electro-optical (EO) imagery provide complementary views of the Earth. SAR measures microwave backscatter and enables day-and-night, weather-resilient observation, whereas EO imagery captures rich spectral appearance, texture, and fine-grained semantic cues. Their complementary information benefits numerous applications, including land-cover mapping, object detection, change and damage assessment, cross-modal retrieval, cloud removal, and SAR-to-EO translation Schmitt & Zhu (2016); Zhu et al. (2017); Do et al. (2026). So, a practical SAR–EO foundation model (FM) should learn strong representations from either modality alone while exploiting their complementary evidence when paired observations are available.

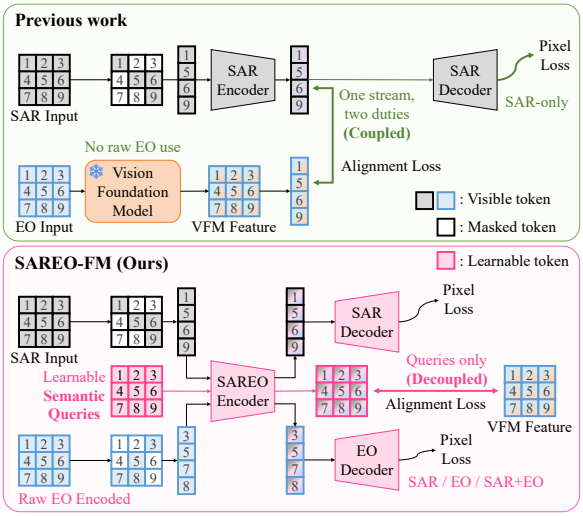

Figure 1: **Motivation of SAREO-FM.** Prior methods directly align image tokens with EO-derived VFM features, coupling semantic alignment with pixel reconstruction. SAREO-FM instead supervises learnable semantic queries while reconstructing SAR and EO from modality tokens, separating the destinations of the two objectives within a unified encoder.

Recent geospatial foundation models (GFMs) Cong et al. (2022); Wang et al. (2025); Yang et al. (2026) have demonstrated the value of pretraining across multiple sensors and data sources. However, simply combining modalities does not resolve a fundamental representation-learning challenge. SAR and EO capture the same geographic scene through substantially different sensing

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.

mechanisms and visual statistics. Their representations must preserve modality-specific details while also learning high-level semantics that transfer across modalities and downstream tasks.

Masked autoencoders (MAEs) He et al. (2022) provide a natural framework for preserving sensing-specific information through masked image reconstruction. Recent approaches further enhance MAE representations using semantic guidance from pretrained vision foundation models (VFMs) Yu et al. (2024); Yao et al. (2025); Liu et al. (2025). As illustrated in Figure 1(a), prior representative methods directly align SAR or EO image tokens with VFM features. Consequently, the same image tokens are explicitly optimized both to preserve modality-specific information for pixel reconstruction and to match high-level EO-derived semantics. To our best knowledge, we *firstly* address that this coupling raises a central question: *where should semantic supervision be applied so that shared semantic learning does not directly constrain modality-specific reconstruction?*

We address this question with **SAREO-FM**, a unified MAE-based FM with *decoupled semantic supervision*. As shown in Figure 1(b), our SAREO-FM processes SAR tokens, EO tokens, and learnable semantic queries using a shared and unified encoder. The semantic queries interact with all available image tokens, and are aligned patch-wise with features from a frozen pretrained VFM. In parallel, the modality-token outputs are passed to separate SAR and EO decoders for masked pixel reconstruction. Importantly, the semantic queries are excluded from the reconstruction decoders. Thus, the SAREO-FM separates the explicit destinations of semantic alignment and pixel reconstruction while retaining information exchange among all three streams within the shared and unified encoder. To support different input configurations, we further introduce a mixed modality masking strategy that combines independent masking, spatially shared masking, and complete dropping of either modality. This exposes the same encoder to joint SAR–EO, SAR-only, and EO-only observations during pretraining. At downstream transfer, semantic queries always participate in encoding, while task heads can selectively consume modality-token outputs, semantic-query outputs, or both according to specific tasks. We pretrain SAREO-FM on SAR–EO observations from the million-scale SAR-1M corpus Liu et al. (2025), treating both SAR and EO imagery as raw learning inputs rather than using EO solely to obtain teacher features. We evaluate the resulting representations on image classification and multi-label recognition, and further examine qualitative transfer to dense prediction under SAR-only, EO-only, and joint SAR–EO settings. Our SAREO-FM achieves strong unimodal transfer and substantial task-dependent improvements from joint observations. Extensive ablations further validate the effects of semantic-query supervision and its separation from reconstruction.

Our contributions are as follows:

- We introduce **SAREO-FM**, a unified MAE-based foundation model that learns from raw SAR and EO imagery using a shared encoder and supports SAR-only, EO-only, and joint SAR–EO transfer.
- We propose *decoupled semantic supervision*, which assigns EO-VFM alignment to learnable semantic queries and modality-specific reconstruction to image tokens, while allowing the three streams to interact within the shared encoder.
- We develop a mixed modality masking strategy that exposes the encoder to paired and missing-modality inputs. Comprehensive evaluations demonstrate strong unimodal transfer and substantial task-dependent gains from complementary SAR–EO observations.

2 Related Work

2.1 Masked Multimodal Modeling

MAE learns visual representations by reconstructing heavily masked image patches with an asymmetric encoder–decoder He et al. (2022). Subsequent work extends masked modeling beyond RGB imagery. MultiMAE Bachmann et al. (2022) jointly models RGB, depth, and semantic maps, while 4M Mizrahi et al. (2023) and 4M-21 Bachmann et al. (2024) scale masked modeling to a broad collection of input and output modalities, including images, geometry, semantics, and feature representations. These approaches demonstrate that masked modeling can effectively integrate heterogeneous signals. Our SAREO-FM follows this general framework but introduces separate token streams as the explicit destinations of pixel-reconstruction and frozen-VFM supervision.

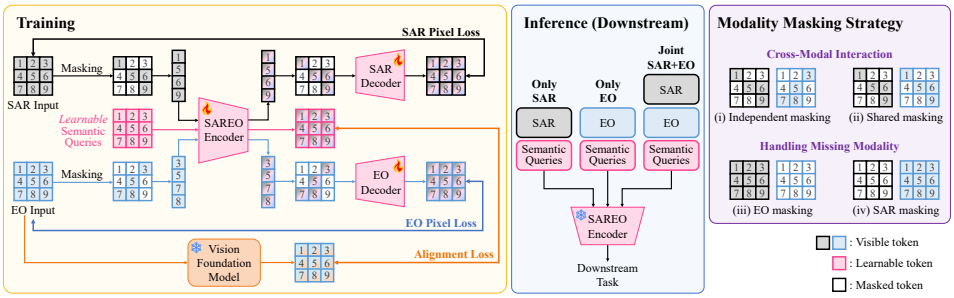

Figure 2: **Overview of SAREO-FM.** Visible SAR and EO tokens interact with learnable semantic queries through a unified encoder. The modality-token outputs are decoded independently for masked pixel reconstruction, whereas the semantic-query outputs are aligned patch-wise with features from a frozen pretrained VFM. Semantic queries are excluded from the reconstruction decoders, separating the destinations of pixel-level and semantic supervision. Modality masking exposes the encoder to SAR-only, EO-only, and joint SAR–EO inputs; during downstream transfer, task heads selectively consume modality-token outputs, semantic-query outputs, or both.

2.2 Geospatial Foundation Models (GFMs)

Self-supervised SAR representation learning has progressed from predictive and reconstruction-based approaches such as SAR-JEPA Li et al. (2023) to large-scale FMs such as SARATR-X Li et al. (2025) and SARMAE Liu et al. (2025). In parallel, multimodal Earth-observation pretraining has explored masked reconstruction, cross-modal alignment, and sensor-aware architectures. CROMA Fuller et al. (2023) combines contrastive radar–optical learning with masked autoencoding, while SatMAE Cong et al. (2022) and SatMAE++ Noman et al. (2024) address multispectral and multitemporal observations. Recent works such as MMEarth Nedungadi et al. (2024), msGFM Han et al. (2024), Copernicus-FM Wang et al. (2025), and MaRS Yang et al. (2026) further broaden the range of sensors, resolutions, and downstream tasks. DeCUR Wang et al. (2024) instead partitions modality-common and modality-specific embedding dimensions through redundancy reduction. Whereas these methods address sensor heterogeneity through alignment, representation partitioning, or architectural adaptation, we focus on separating the explicit supervisions assigned to *decoupled semantic and reconstruction representations* of SAR and EO modalities in a shared latent space by a unified encoder using *learnable semantic queries*.

2.3 Semantic Feature Alignment

Aligning learned representations with strong pretrained encoders has proven effective across representation and generative learning, as exemplified by REPA Yu et al. (2024) and VA-VAE Yao et al. (2025). In SAR pretraining, SARMAE Liu et al. (2025) introduces semantic guidance from a frozen DINOv3 Siméoni et al. (2025) teacher by directly aligning SAR encoder tokens with features extracted from paired EO imagery. The EO image therefore provides teacher-side supervision but is not encoded as an input modality by the student. CoDe-MAE Peng et al. (2026), the closest concurrent multimodal MAE, jointly handles SAR and EO through optical-token distillation, conditioned on (i) contrastive alignment and (ii) degraded cross-modal reconstruction. Our SAREO-FM differs in the destination of semantic supervision: frozen-VFM features supervise only the persistent semantic-query outputs, while SAR and EO token outputs are used for modality-specific reconstruction.

3 SAREO-FM

3.1 Overview

Our SAREO-FM learns modality-specific and cross-modal representations from SAR and EO imagery within a unified masked-autoencoding framework. Figure 2 illustrates the overall architecture of SAREO-FM. Our key design principle is to separate the destinations of pixel-level reconstruction and high-level semantic supervision. To this end, the SAREO-FM operates on three token streams:

SAR image tokens, EO image tokens, and learnable semantic queries, which interact through a shared SAREO Encoder. The semantic queries aggregate cross-modal context from the visible SAR and EO tokens and are aligned with patch-level features extracted by a frozen pretrained VFM. In parallel, the encoded SAR and EO tokens are passed to modality-specific decoders for masked image reconstruction, thereby preserving sensing-specific appearance and structure. The semantic queries are excluded from the decoding stage, preventing the reconstruction decoders from directly relying on the semantically supervised stream. After pretraining, the VFM teacher and reconstruction decoders are discarded, and the unified encoder serves as our SAREO-FM for SAR, EO, or joint SAR–EO observations.

3.2 Unified Three-Stream SAREO Encoder

Let $\mathbf{X}_s \in \mathbb{R}^{H\times W\times C_s}$ and $\mathbf{X}_e \in \mathbb{R}^{H\times W\times C_e}$ denote a spatially registered SAR–EO pair, where s and e index the SAR and EO modalities, respectively. Each image is divided into N non-overlapping patches. For modality $m \in \{s, e\}$, the i-th flattened patch $\mathbf{x}_{p,m}^i$ is projected into a shared D-dimensional embedding space using a modality-specific linear projection $\mathbf{E}_m$. The resulting image token is

$$\mathbf{y}_m^i = \mathbf{x}_{p,m}^i \mathbf{E}_m + \mathbf{p}^i + \mathbf{a}_m^{\text{enc}}, \qquad i = 1, \ldots, N, \tag{1}$$

where $\mathbf{p}^i$ is a shared 2D sine–cosine positional embedding and $\mathbf{a}_m^{\text{enc}}$ is a learnable modality embedding for the encoder. We denote the full token sequence by $\mathbf{Y}_m = [\mathbf{y}_m^1, \ldots, \mathbf{y}_m^N]$. For each modality, we sample a binary mask $\mathcal{M}_m \in \{0, 1\}^N$, where $\mathcal{M}_m[i] = 1$ or 0 indicates a masked or visible patch. We define the masked and visible index sets as

$$\Omega_m = \{i \mid \mathcal{M}_m[i] = 1\}, \quad \mathcal{V}_m = \{1, \ldots, N\} \setminus \Omega_m. \tag{2}$$

Following masked autoencoders, only the visible image tokens are provided to the encoder:

$$\mathbf{Y}_m^{\text{vis}} = [\mathbf{y}_m^i]_{i\in\mathcal{V}_m}. \tag{3}$$

In addition to the image streams, we introduce a spatially organized set of N learnable semantic queries, $\mathbf{Q} = \{\mathbf{q}^i\}_{i=1}^N$. Each query corresponds to one image-patch location and is augmented with the same positional embedding used for the image tokens, together with a dedicated query-stream embedding:

$$\widetilde{\mathbf{q}}^i = \mathbf{q}^i + \mathbf{p}^i + \mathbf{a}_q^{\text{enc}}, \qquad \widetilde{\mathbf{Q}} = [\widetilde{\mathbf{q}}^1, \ldots, \widetilde{\mathbf{q}}^N]. \tag{4}$$

Unlike the image tokens, the semantic queries are never masked. The visible image tokens and semantic queries are concatenated and jointly processed by the shared SAREO Encoder $\mathcal{E}$ that yields three latents for SAR, EO and query:

$$[\mathbf{Z}_s^{\text{vis}}; \mathbf{Z}_e^{\text{vis}}; \mathbf{Z}_q] = \mathcal{E}\left([\mathbf{Y}_s^{\text{vis}}; \mathbf{Y}_e^{\text{vis}}; \widetilde{\mathbf{Q}}]\right). \tag{5}$$

where $[\cdot;\cdot;\cdot]$ indicates a concatenation operator. When a modality is unavailable or intentionally withheld, its token sequence is omitted from Equation 5. Through joint self-attention, the query latent $\mathbf{Z}_q$ aggregates spatially aligned evidence from all available modalities. Meanwhile, the SAR and EO latents, $\mathbf{Z}_s^{\text{vis}}$ and $\mathbf{Z}_e^{\text{vis}}$, exchange cross-modal context through both self-attention within concatenated three-modal streams.

3.3 Decoupled Semantic Supervision

Masked reconstruction encourages image tokens to preserve low-level semantic representations such as local appearance, geometry, and sensing-specific characteristics. In contrast, a pretrained VFM operating on EO imagery encourages high-level semantic representations that are shared across sensing modalities. Directly applying both objectives to the same image tokens forces a single token stream to simultaneously satisfy these different representational demands. The SAREO-FM separates their explicit supervision destinations by applying the VFM target exclusively to the semantic-query stream. Given an EO image, a frozen VFM teacher f_θ processes the clean and unmasked observation, and extracts its final-layer patch tokens:

$$\mathbf{T} = f_\theta(\mathbf{X}_e) = [\mathbf{t}^1, \ldots, \mathbf{t}^N] \in \mathbb{R}^{N\times D_t}, \tag{6}$$

where $\mathbf{t}^i$ denotes the teacher feature corresponding to the i-th spatial location, and D_t is the teacher feature dimension. If the teacher produces a different spatial resolution, its patch-token grid is spatially resampled to match the N query locations. A lightweight semantic projector $h_\phi : \mathbb{R}^D \to \mathbb{R}^{D_t}$ maps the contextualized queries to the teacher feature space. We then impose patch-wise semantic alignment using cosine distance:

$$\mathcal{L}_{\text{align}} = \frac{1}{N} \sum_{i=1}^{N} \left(1 - \frac{h_\phi(\mathbf{z}_q^i)^\top \mathbf{t}^i}{\left\| h_\phi(\mathbf{z}_q^i) \right\|_2 \left\| \mathbf{t}^i \right\|_2} \right). \tag{7}$$

We refer to this design as *decoupled semantic supervision*. The modality tokens are not directly constrained to reproduce EO-VFM features, and can therefore retain information useful for sensing-specific reconstruction. Instead, the dedicated query stream serves as the explicit destination of high-level semantic supervision. Importantly, this decoupling applies to the supervision heads rather than to the encoder computation: all three streams continue to exchange information, and gradients from both objectives jointly optimize the SAREO Encoder $\mathcal{E}$.

3.4 Modality-Specific Masked Reconstruction

For pixel-level recovery, we employ a modality-specific reconstruction decoder $\mathcal{D}_m$ for $m \in \{s, e\}$. Consistent with our decoupled design, the semantic-query latents $\mathbf{Z}_q$ are discarded before decoding, and are not used as decoder conditions. For each modality, the encoded visible tokens are restored to their original spatial locations, and a shared learnable mask token $\mathbf{e}_m$ is inserted at every masked location. Specifically, we construct the ordered decoder input sequence

$$\mathbf{u}_m^i = \begin{cases} \mathbf{z}_m^{\rho_m(i)}, & i \in \mathcal{V}_m, \\ \mathbf{e}_m, & i \in \Omega_m, \end{cases} \tag{8}$$

where $\rho_m(i)$ maps the original spatial index i to its position in the packed visible-token sequence $\mathbf{Z}_m^{\text{vis}}$. The decoder input at location i is then

$$\mathbf{w}_m^i = \mathbf{u}_m^i + \mathbf{p}_{\text{dec}}^i + \mathbf{a}_m^{\text{dec}}, \quad \mathbf{W}_m = [\mathbf{w}_m^1; \ldots; \mathbf{w}_m^N], \tag{9}$$

where $\mathbf{p}_{\text{dec}}^i$ is a 2D sine–cosine decoder positional embedding and $\mathbf{a}_m^{\text{dec}}$ is a learnable decoder-specific modality embedding. The decoder predicts all image patches as:

$$[\hat{\mathbf{x}}_{p,m}^1; \ldots; \hat{\mathbf{x}}_{p,m}^N] = \mathcal{D}_m(\mathbf{W}_m). \tag{10}$$

Following standard masked reconstruction, the pixel loss is evaluated only at masked locations:

$$\mathcal{L}_{\text{pix}}^m = \frac{1}{|\Omega_m|} \sum_{i \in \Omega_m} \left\| \hat{\mathbf{x}}_{p,m}^i - \mathbf{x}_{p,m}^i \right\|_1. \tag{11}$$

where, $\mathcal{L}_{\text{pix}}^s$ and $\mathcal{L}_{\text{pix}}^e$ correspond to the SAR and EO pixel losses in Figure 2, respectively.

3.5 Modality Masking Strategy

To expose the unified SAREO encoder $\mathcal{E}$ to diverse downstream input configurations, we train the SAREO-FM using four complementary masking modes:

(i) Independent masking. SAR and EO masks are sampled independently. Visible patches from one modality can provide spatially aligned cues for locations masked in the other, encouraging fine-grained cross-modal reasoning.

(ii) Shared masking. The same spatial mask is applied to both modalities. Because the corresponding location is simultaneously hidden in SAR and EO, the model must rely on broader spatial and semantic context rather than direct local correspondence.

(iii) SAR-only input (EO masking). The EO token stream is entirely withheld from the encoder, while the SAR image is partially masked. The semantic queries must therefore extract semantic information solely from the available SAR observations.

(iv) EO-only input (SAR masking). The SAR token stream is symmetrically withheld, leaving only partially masked EO tokens and the semantic queries as encoder inputs.

The first two modes, (i) and (ii), promote cross-modal interaction, whereas the modality-only modes, (iii) and (iv), ensure that the same encoder $\mathcal{E}$ remains effective when only one modality is available.

3.6 PRETRAINING OBJECTIVE

Our pretraining objective combines modality-specific reconstruction and semantic alignment:

$$\mathcal{L} = \lambda_s \mathcal{L}^s_{\text{pix}} + \lambda_e \mathcal{L}^e_{\text{pix}} + \lambda_{\text{align}} \mathcal{L}_{\text{align}}, \tag{12}$$

where each coefficient is set to zero when its supervision is unavailable or excluded by the sampled input mode. SAR–EO inputs activate four cases: (i) SAR-only inputs with EO masked use $\mathcal{L}^s_{\text{pix}}$ and $\mathcal{L}_{\text{align}}$, where EO is retained only for the frozen teacher; (ii) EO-only inputs with SAR masked use $\mathcal{L}^e_{\text{pix}}$ and $\mathcal{L}_{\text{align}}$; (iii) Unpaired SAR samples with EO absent contribute only to $\mathcal{L}^s_{\text{pix}}$; and (iv) Unpaired EO samples also contribute to $\mathcal{L}^e_{\text{pix}}$ and $\mathcal{L}_{\text{align}}$.

3.7 DOWNSTREAM TRANSFER

During downstream transfer, the frozen VFM teacher f_θ, semantic projector h_ϕ, and reconstruction decoders $\{\mathcal{D}_m\}$ are discarded. The pretrained SAREO Encoder $\mathcal{E}$ receives all available, unmasked image tokens together with the semantic queries where a missing modality is represented by an empty token sequence, allowing $\mathcal{E}$ to process SAR, EO, or paired SAR–EO observations without modality-specific backbones or separate inference pipelines. $\mathcal{E}$ provides complementary representations for different downstream requirements. Modality-token outputs retain sensing-specific and spatially detailed information, whereas the semantic-query outputs provide contextually aggregated, semantically enriched features. Depending on the downstream tasks such as dense prediction or image-level recognition, their task-specific features can be selectively extracted from the modality tokens, the semantic queries, or their combination.

4 EXPERIMENTS

4.1 DATASETS AND DOWNSTREAM TASKS

Pretraining corpus. We pretrain SAREO-FM on SAR-1M Liu et al. (2025), which aggregates 18 public SAR datasets spanning classification, detection, and segmentation scenarios. It contains 1,312,902 SAR images, of which 1,042,156 are paired with geographically aligned EO observations. Overall, the corpus covers 57 target and scene categories, multiple SAR sensors and polarization modes, and ground resolutions ranging from 0.1 to 60 m.

Downstream benchmarks. Downstream heads use modality-token features (mod), semantic-query features (sem), or their concatenation (both). We evaluate SAR recognition on MSTAR Ross et al. (1998), FUSAR-Ship Hou et al. (2020), and SAR-ACD Sun et al. (2022). MSTAR and FUSAR-Ship measure in-corpus transfer, as they are included in SAR-1M, whereas SAR-ACD serves as our primary out-of-corpus benchmark. EO-only transfer is evaluated on EuroSAT Helber et al. (2019), NWPU-RESISC45 Cheng et al. (2017), and AID Xia et al. (2017) under few-shot and higher-supervision settings. For paired SAR–EO transfer, we evaluate single-label recognition on So2Sat LCZ42 Zhu et al. (2020) and multi-label recognition on BigEarthNet-MM Sumbul et al. (2021). We additionally examine qualitative dense transfer on BRIGHT Chen et al. (2025a), which requires building-damage assessment from paired pre-event EO and post-event SAR observations. We report the mean performance over three independent runs for end-to-end fine-tuning and five independent runs for linear probing.

4.2 IMPLEMENTATION DETAILS

Pretraining. SAREO-FM uses a ViT-B/16 encoder Dosovitskiy et al. (2020) trained from scratch on 256×256 inputs, yielding a 16×16 token grid. We accordingly introduce 256 ($N = 16 \times 16$) learnable semantic queries. The teacher f_θ is a frozen DINOv3-7B model Siméoni et al. (2025), and the semantic projector h_ϕ is a two-layer MLP. We empirically set the loss weights to $\lambda_{\text{s}} = 1$, $\lambda_{\text{e}} = 1$, and $\lambda_{\text{align}} = 0.5$. We sample independent masking, spatially shared masking, complete EO dropping, and complete SAR dropping with probabilities $0.35/0.35/0.15/0.15$. Each retained modality is masked by 75%, while all semantic queries remain active. We train for 300,000 steps using AdamW Loshchilov & Hutter (2017) with $\beta = (0.9, 0.95)$, weight decay 0.05, and an effective batch size of 2,560. The learning rate reaches 1.5×10^{-3} after 15,000 warm-up steps and then

Table 1: **SAR target classification with end-to-end fine-tuning.** The upper block reports published OA (%) under the source-specific protocol of each method and is included only as context. The lower block reports our controlled evaluation with identical data splits and downstream pipelines; mean and standard deviation over three seeds are shown. * denotes our re-evaluation of the official SARMAE checkpoint. **Bold** denotes the best controlled result.

Method	Venue	Backbone	FUSAR-Ship		MSTAR		SAR-ACD	
			40-shot	30%	40-shot	30%	40-shot	30%
ResNet-50 He et al. (2016)	CVPR'16	ResNet-50	–	58.41	–	89.94	–	59.70
Swin Transformer Liu et al. (2021)	ICCV'21	Swin-B	–	60.79	–	82.97	–	67.50
BEiT Bao et al. (2021)	ICLR'22	ViT-B	59.70	71.13	40.70	69.75	–	79.77
CROMA Fuller et al. (2023)	NeurIPS'23	ViT-B	83.71	–	–	–	–	88.99
SAR-JEPA Li et al. (2023)	ISPRS JPRS'24	ViT-B	85.80	–	91.60	–	75.50	–
SARATR-X Li et al. (2025)	TIP'25	HiViT-B	87.70	–	98.10	–	76.40	–
LoMaR Chen et al. (2025b)	WACV'25	ViT-B	82.70	–	77.00	–	67.40	–
SUMMIT Du et al. (2025)	IJAEOG'25	ViT-B	81.50	71.91	63.60	98.39	68.70	84.25
Copernicus-FM Wang et al. (2025)	ICCV'25	ViT-B	87.61	–	–	–	–	92.63
CoDe-MAE Peng et al. (2026)	arXiv'26	HiViT-B	89.40	–	98.70	–	78.30	–
MaRS Yang et al. (2026)	AAAI'26	SwinV2-B	77.70	–	75.50	–	68.40	–
SARMAE Liu et al. (2025)	CVPR'26	ViT-B	89.30	92.92	96.70	99.61	–	95.06
SARMAE Liu et al. (2025)	CVPR'26	ViT-L	90.86	92.80	97.24	98.92	–	95.63
Random initialization	–	ViT-B	60.31 ± 1.39	82.11 ± 0.52	34.71 ± 0.09	38.76 ± 8.35	33.14 ± 2.60	39.19 ± 10.27
SARMAE* Liu et al. (2025)	CVPR'26	ViT-B	90.09 ± 0.36	92.82 ± 0.21	94.54 ± 1.72	97.18 ± 0.32	68.50 ± 2.60	$\mathbf{95.01} \pm 0.85$
SAREO-FM (sem)	–	ViT-B	91.05 ± 0.16	$\mathbf{93.13} \pm 0.14$	$\mathbf{97.57} \pm 0.42$	99.13 ± 0.54	$\mathbf{78.40} \pm 1.24$	93.34 ± 0.55
SAREO-FM (both)	–	ViT-B	$\mathbf{91.41} \pm 0.23$	93.12 ± 0.06	96.66 ± 0.31	$\mathbf{99.36} \pm 0.47$	78.00 ± 1.00	92.84 ± 0.68

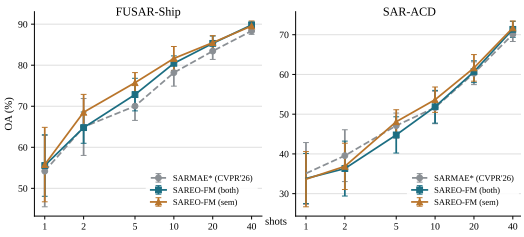

Figure 3: **Frozen few-shot SAR transfer.** We use fixed encoder features and train only the downstream classifier.

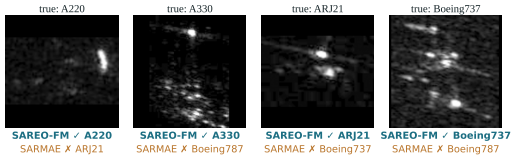

Figure 4: **Qualitative comparison on SAR-ACD.** Representative test images misclassified by SARMAE* but correctly classified by SAREO-FM.

follows cosine decay to zero. Training uses bfloat16 (bf16) precision, gradient clipping at 1.0, and four NVIDIA B200 GPUs, requiring approximately 390 B200 GPU-hours.

Controlled comparison. SARMAE Liu et al. (2025) serves as our primary controlled baseline. We denote by SARMAE* our reevaluation of its official pretrained ViT-B checkpoint under exactly the same downstream protocol as SAREO-FM, including dataset splits, sampled training subsets, input preprocessing, classifier heads, optimization settings, and evaluation code. Both methods employ ViT-B encoders pretrained on exactly the same SAR images from SAR-1M and leverage DINOv3 features Siméoni et al. (2025) extracted from corresponding EO observations. The key distinction is how the paired EO information is used: SARMAE uses EO-derived DINO features solely as semantic supervision for SAR representation learning, whereas our SAREO-FM additionally processes the raw EO observations as a second modality, and jointly learns SAR–EO reconstruction and cross-modal representations. SARMAE* therefore provides our closest controlled comparison, isolating the effect of incorporating raw paired EO observations and cross-modal learning.

4.3 SAR Representation Transfer

End-to-end fine-tuning. Table 1 first places our SAREO-FM among published SAR representation learners, and then provides a controlled comparison with SARMAE*. The SAREO-FM obtains the best accuracy in five of the six controlled settings. On FUSAR-Ship, it improves over SARMAE* by 1.32 points at 40 shots and 0.31 points with 30% supervision. The corresponding gains on MSTAR are 3.03 and 2.18 points. These results show that the benefit persists under substantially different supervision budgets, although both datasets overlap with the pretraining corpus. The strongest evidence comes from out-of-corpus SAR-ACD. With only 40 examples per class,

Table 2: **Frozen few-shot EO transfer.** We report accuracy (%) for SAR and EO inputs over five support draws. CoDe-MAE values use its separate 10-shot protocol and are included only as context.

Method / Feature	SAR				EO			
	1-shot	5-shot	10-shot	40-shot	1-shot	5-shot	10-shot	40-shot
DINOv3 (teacher)	38.17 ± 4.61	50.72 ± 2.68	56.01 ± 1.81	64.35 ± 1.56	59.51 ± 3.53	75.96 ± 3.15	83.34 ± 1.42	90.88 ± 0.65
SARMAE*	37.22 ± 5.71	48.64 ± 3.00	53.97 ± 1.39	64.72 ± 1.12	43.56 ± 6.57	59.57 ± 3.05	66.49 ± 1.97	78.36 ± 1.07
CoDe-MAE	–	–	$59.88^{\dagger}$	–	–	–	$81.18^{\dagger}$	–
SAREO-FM (both)	**39.07** ± 5.98	**53.54** ± 2.21	**59.36** ± 1.46	**69.24** ± 0.96	**64.41** ± 5.05	**81.09** ± 1.61	**86.76** ± 1.24	**92.58** ± 0.75

semantic queries improve accuracy from 68.50% to 78.40%, a 9.90-point gain. Under 30% supervision, however, SARMAE* remains 1.67 points stronger. The contrast indicates that our SAREO-FM primarily improves label-efficient transfer rather than performance after abundant task-specific adaptation. Across the 40-shot columns, sem is the strongest on MSTAR and SAR-ACD, while both is the strongest on FUSAR-Ship, supporting task-dependent use of the two output streams.

Frozen few-shot transfer. Figure 3 removes encoder adaptation, and tests whether the pretrained representations are directly accessible to a lightweight classifier. On FUSAR-Ship, our SAREO-FM becomes consistently stronger than SARMAE* from 2-shot onward, with the largest separation occurring at 5-shot. On SAR-ACD, the advantage emerges at 5-shot and persists through 40-shot. The similar trajectories of sem and both show that the learned semantic queries remain useful even without end-to-end fine-tuning, while concatenating modality tokens can provide a small task-dependent complement. Figure 4 complements the aggregate improvement with representative out-of-corpus examples. Our SAREO-FM correctly separates visually similar aircraft categories where the controlled SARMAE* baseline fails, consistent with the large 40-shot gain in Table 1.

4.4 EO Representation Transfer

Table 2 tests whether $\mathcal{E}$ transfers beyond SAR imagery. With EO input, our SAREO-FM consistently surpasses its frozen DINOv3 teacher, improving accuracy by 4.90, 5.13, 3.42, and 1.70 points from 1 to 40 shots. This result is notable because DINOv3 supplies the semantic target during pretraining: the SAREO-FM does not merely reproduce its features, but adapts them through joint SAR–EO context and reconstruction. The largest gains occur in the most label-scarce regimes, whereas the gap narrows as the number of labels increases. $\mathcal{E}$ also supports SAR-input EuroSAT recognition. Our SAREO-FM improves over the controlled SARMAE* baseline at every shot count, with gains ranging from 1.85 to 5.39 points, and reaches 69.24% at 40 shots. The separately reported CoDe-MAE numbers provide protocol-specific context rather than a direct comparison. Together, Table 2 shows that the shared representation retains useful scene semantics from either modality rather than specializing exclusively in SAR target recognition.

4.5 Joint SAR–EO Transfer

Table 3: **Transfer on paired SAR–EO benchmarks.** We report OA (%) on 5-shot So2Sat LCZ42 and micro-AP (%) on 5-shot BigEarthNet-MM.

Input	Method	Feature	Benchmark	
			So2Sat LCZ42	BigEarthNet-MM
SAR+EO	Random initialization	mod	40.56 ± 2.25	54.29 ± 2.30
EO	DINOv3 (teacher)	mod	54.25 ± 3.87	$\underline{64.73} \pm 1.43$
EO	**SAREO-FM**	mod	$\underline{58.14} \pm 2.90$	64.42 ± 1.44
SAR	SARMAE*	mod	28.96 ± 1.68	57.44 ± 0.69
SAR	**SAREO-FM**	mod	29.59 ± 2.64	60.47 ± 0.79
SAR+EO	**SAREO-FM**	both	**58.79** ± 4.19	**66.70** ± 0.58

Table 3 evaluates whether paired observations provide information beyond either sensor alone. On BigEarthNet-MM, joint SAR–EO input reaches 66.70 micro-AP, improving over the best SAREO-FM unimodal input by 2.28 points and the DINOv3 EO baseline by 1.97 points. The SAR branch itself improves over SARMAE* by 3.03 points, while the EO branch performs similarly to DINOv3. The joint gain therefore reflects complementary SAR information rather than a uniformly stronger EO representation. On So2Sat LCZ42, the EO branch improves over DINOv3 by 3.89 points, whereas adding SAR yields a smaller 0.65-point gain. Thus, multimodal fusion is beneficial when the downstream labels exploit complementary sensor cues, but is not guaranteed to improve every paired benchmark.

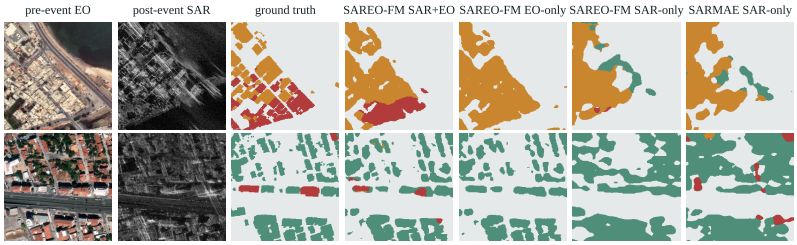

Figure 5: **Qualitative transfer on BRIGHT Chen et al. (2025a).** Given pre-event EO and post-event SAR observations, SAREO-FM fusion recovers destroyed buildings that the corresponding single-modality models predict as intact or background.

Dense qualitative transfer. Although our quantitative evaluation focuses on image-level recognition, Figure 5 illustrates how the unified encoder can also support spatially aligned dense prediction. On BRIGHT, combining pre-event EO with post-event SAR recovers destroyed-building regions missed by EO-only and SAR-only inference, including the SARMAE* baseline. These examples are qualitative evidence rather than a substitute for a full segmentation benchmark, but they are consistent with the task-dependent fusion gain observed on BigEarthNet-MM.

4.6 Ablation Study

We conduct a cumulative component ablation under matched pretraining and evaluation budgets. Each variant is pretrained from scratch for 30,000 steps with an effective batch size of 2,560. After pretraining, we freeze $\mathcal{E}$ and evaluate its SAR representation via 10-shot linear probing using the same 768-dimensional modality-feature readout. We report the mean and standard deviation over five seeds, using identical support sets for all variants. Table 4 shows a consistent improvement as each component is introduced. Plain masked autoencoding establishes a strong representation-learning baseline, outperforming random initialization by 34.25 points on average. Adding supervision from the frozen DINOv3 teacher further improves the average by 5.39 points, indicating that semantic information derived from the geographically corresponding EO observations transfers effectively to SAR representations.

Table 4: **Cumulative component ablation.** All variants use matched pretraining and evaluation protocols. We report 10-shot SAR linear-probing performance. Avg. denotes the arithmetic mean across the two benchmarks, while Δ measures the gain over the preceding configuration.

Pretraining configuration	MSTAR (OA↑)	FUSAR (OA↑)	Avg. (OA↑)	Δ
Random initialization	26.26 ± 1.86	47.86 ± 4.89	37.06	–
Plain MAE He et al. (2022)	66.16 ± 3.15	76.46 ± 2.52	71.31	+34.25
+ DINOv3 supervision	73.49 ± 2.31	79.90 ± 2.49	76.70	+5.39
+ Decoupled semantic supervision	$\underline{77.11} \pm 2.14$	$\underline{82.55} \pm 2.15$	$\underline{79.83}$	+3.13
+ Raw EO modality (**SAREO-FM**)	$\mathbf{78.41} \pm 2.14$	$\mathbf{84.00} \pm 2.88$	**81.21**	+1.38

Decoupling semantic prediction from modality reconstruction yields a further 3.13-point gain. In particular, assigning semantic alignment to dedicated queries avoids requiring the modality tokens to simultaneously serve semantic prediction and pixel reconstruction, leading to more transferable SAR features. Finally, exposing the encoder to the raw paired EO observations and enabling extra EO reconstruction provides an additional 1.38-point improvement. Overall, the complete SAREO-FM improves upon plain MAE by 9.90 points on average, and achieves the best result on both benchmarks. These results show that teacher-based semantic supervision, objective decoupling, and raw cross-modal learning provide cumulative and complementary benefits.

5 Conclusion

We presented SAREO-FM, a unified foundation model for SAR, EO, and joint SAR–EO observations. The SAREO-FM assigns EO-VFM supervision to semantic queries and masked reconstruction to modality tokens, separating their explicit supervision while preserving their interaction within a shared encoder. The learned representations transfer effectively across SAR-only, EO-only, and multimodal downstream tasks, with joint observations providing benefits when the sensors offer complementary task-relevant information. Future work will extend this framework to additional sensing modalities and dense prediction tasks.

AI USE STATEMENT

We used generative AI tools to assist with translation, English-language editing, and manuscript organization. The authors reviewed and revised all AI-assisted text, checking technical descriptions against the implemented method and empirical statements against the reported experimental results. The authors take full responsibility for the final submission, including all text, methodology, results, claims, and references.

ETHICS STATEMENT

This study uses existing synthetic aperture radar (SAR) and electro-optical (EO) datasets for representation learning and downstream evaluation. The source datasets remain subject to their original access, usage, and redistribution conditions. Even when explicit personal identifiers are absent, geospatial imagery may reveal sensitive information about locations, infrastructure, and human activities. Applications of SAREO-FM therefore require appropriate consideration of privacy, data governance, and potential misuse, including unauthorized monitoring. Performance on the evaluated benchmarks does not establish reliability or equitable performance across geographic regions, sensors, or acquisition conditions. Consequential applications, such as disaster assessment, require further validation under the intended deployment conditions.

REPRODUCIBILITY STATEMENT

The SAREO-FM architecture and learning objectives are described in Section 3, while the datasets, training configuration, and evaluation settings are presented in Section 4. The appendices provide additional details on masking modes and active losses (Appendix B), pretraining corpus composition and downstream data splits (Appendix C), architecture specifications (Appendix D.1), downstream feature extraction (Appendix E.1), and inference computation (Appendix F). Additional experimental results and the scope of the controlled comparisons are provided in Appendix A.

REFERENCES

Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *European conference on computer vision*, pp. 348–367. Springer, 2022.

Roman Bachmann, Oğuzhan F Kar, David Mizrahi, Ali Garjani, Mingfei Gao, David Griffiths, Jiaming Hu, Afshin Dehghan, and Amir Zamir. 4m-21: An any-to-any vision model for tens of tasks and modalities. *Advances in Neural Information Processing Systems*, 37:61872–61911, 2024.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Hongruixuan Chen, Jian Song, Olivier Dietrich, Clifford Broni-Bediako, Weihao Xuan, Junjue Wang, Xinlei Shao, Yimin Wei, Junshi Xia, Cuiling Lan, et al. Bright: A globally distributed multimodal building damage assessment dataset with very-high-resolution for all-weather disaster response. *Earth System Science Data*, 17(11):6217–6253, 2025a.

Jun Chen, Faizan Farooq Khan, Ming Hu, Ammar Sherif, Zongyuan Ge, Boyang Li, and Mohamed Elhoseiny. Local masked reconstruction for efficient self-supervised learning on high-resolution images. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 8046–8056. IEEE, 2025b.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

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 *Advances in neural information processing systems*, 2022.

Jeonghyeok Do, Jaehyup Lee, Seungchul Lee, and Munchurl Kim. C-diffset: Leveraging latent diffusion for sar-to-eo image translation with confidence-guided reliable object generation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Yuntao Du, Yushi Chen, Lingbo Huang, Yahu Yang, Pedram Ghamisi, and Qian Du. Summit: A sar foundation model with multiple auxiliary tasks enhanced intrinsic characteristics. *International Journal of Applied Earth Observation and Geoinformation*, 141:104624, 2025.

Anthony Fuller, Koreen Millard, and James Green. Croma: Remote sensing representations with contrastive radar-optical masked autoencoders. *Advances in Neural Information Processing Systems*, 36:5506–5538, 2023.

Boran Han, Shuai Zhang, Xingjian Shi, and Markus Reichstein. Bridging remote sensors with multisensor geospatial foundation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 27852–27862, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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*, pp. 16000–16009, 2022.

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

Xiyue Hou, Wei Ao, Qian Song, Jian Lai, Haipeng Wang, and Feng Xu. Fusar-ship: Building a high-resolution sar-ais matchup dataset of gaofen-3 for ship detection and recognition. *Science China Information Sciences*, 63(4):140303, 2020.

Weijie Li, Yang Wei, 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. *arXiv preprint arXiv:2311.15153*, 2023.

Weijie Li, Wei Yang, Yuenan Hou, Li Liu, Yongxiang Liu, and Xiang Li. Saratr-x: Toward building a foundation model for sar target recognition. *IEEE Transactions on Image Processing*, 34:869–884, 2025.

Danxu Liu, Di Wang, Hebaixu Wang, Haoyang Chen, Wentao Jiang, Yilin Cheng, Haonan Guo, Wei Cui, and Jing Zhang. Sarmae: Masked autoencoder for sar representation learning. *arXiv preprint arXiv:2512.16635*, 2025.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

David Mizrahi, Roman Bachmann, Oguzhan Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4m: Massively multimodal masked modeling. *Advances in Neural Information Processing Systems*, 36:58363–58408, 2023.

Vishal Nedungadi, Ankit Kariryaa, Stefan Oehmcke, Serge Belongie, Christian Igel, and Nico Lang. Mmearth: Exploring multi-modal pretext tasks for geospatial representation learning. In *European Conference on Computer Vision*, pp. 164–182. Springer, 2024.

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

Bowen Peng, Yongxiang Liu, Jie Zhou, Xiaodong Chen, Tianpeng Liu, Xiaogang Yu, and Li Liu. Better with less: Tackling heterogeneous multi-modal image joint pretraining via conditioned and degraded masked autoencoder. *arXiv preprint arXiv:2604.16952*, 2026.

Timothy D Ross, Steven W Worrell, Vincent J Velten, John C Mossing, and Michael Lee Bryant. Standard sar atr evaluation experiments using the mstar public release data set. In *Algorithms for synthetic aperture radar imagery V*, volume 3370, pp. 566–573. SPIE, 1998.

Michael Schmitt and Xiao Xiang Zhu. Data fusion and remote sensing: An ever-growing relationship. *IEEE Geoscience and Remote Sensing Magazine*, 4(4):6–23, 2016.

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

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 [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 9(3):174–180, 2021.

Xian Sun, Yixuan Lv, Zhirui Wang, and Kun Fu. Scan: Scattering characteristics analysis network for few-shot aircraft classification in high-resolution sar images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–17, 2022.

Yi Wang, Conrad M Albrecht, Nassim Ait Ali Braham, Chenying Liu, Zhitong Xiong, and Xiao Xiang Zhu. Decoupling common and unique representations for multimodal self-supervised learning. In *18th European conference on computer vision, ECCV 2024*, volume 15087, pp. 286–303, 2024.

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

Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, 2017.

Ruoyu Yang, Yinhe Liu, Heng Yan, Yiheng Zhou, Yihan Fu, Han Luo, and Yanfei Zhong. Mars: A multi-modality very-high-resolution remote sensing foundation model with cross-granularity meta-modality learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11685–11693, 2026.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15703–15712, 2025.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE geoscience and remote sensing magazine*, 5(4):8–36, 2017.

Xiao Xiang Zhu, Jingliang Hu, Chunping Qiu, Yilei Shi, Jian Kang, Lichao Mou, Hossein Bagheri, Matthias Haberle, Yuansheng Hua, Rong Huang, et al. So2sat lcz42: A benchmark data set for the classification of global local climate zones [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 8(3):76–89, 2020.

APPENDIX

This *Supplementary Material* first anchors the additional analyses to the principal results in the main paper, and then provides the implementation and evaluation details needed to interpret them. Section A reports the exact frozen few-shot SAR results and defines the scope of the controlled SARMAE* Liu et al. (2025) comparison. Section B makes the four masking modes, executed sampling probabilities, encoder sequence lengths, and active losses explicit. Section C documents corpus membership, downstream splits, and the interpretation of in-corpus and out-of-corpus transfer. Section D records the architecture actually used by the reported checkpoint. Section E defines the feature readouts and presents checkpoint-only readout and fusion controls. Finally, Sections F and G discuss deployment cost, reproducibility, and the appropriate scope of the empirical claims.

Table 5: Overview of the *Supplementary Material*.

Section	Contents
Section A	Main-paper results and additional controls
Section B	Masking modes and training objectives
Section C	Pre-training corpus, datasets, and evaluation scope
Section D	Architecture and pre-training details
Section E	Feature readouts and downstream protocols
Section F	Inference cost
Section G	Limitations

A ADDITIONAL RESULTS AND CONTROLS

A.1 EXACT FROZEN FEW-SHOT SAR RESULTS

Table 6 provides the numerical values underlying the frozen-transfer trends in the main paper. On the in-corpus FUSAR-Ship benchmark, the strongest SAREO-FM readout is numerically above SARMAE* at every evaluated supervision budget, with gains of 1.60, 3.59, 5.71, 3.47, 2.06, and 1.46 points from 1 to 40 shots. The largest separation occurs at 5 shots. On the out-of-corpus SAR-ACD benchmark, SARMAE* remains stronger in the extremely sparse 1- and 2-shot settings. The semantic-query readout crosses over at 5 shots and improves the baseline by 0.98, 1.97, 1.35, and 1.61 points at 5, 10, 20, and 40 shots, respectively. Across the twelve dataset–shot combinations, the semantic-query readout is the stronger of the two SAREO-FM readouts in ten cases. This pattern indicates that semantic queries usually provide the most direct frozen representation for classification, while modality-token features offer a task-dependent complement. These are numerical comparisons; we do not infer statistical significance from Table 6 alone.

B MASKING AND OBJECTIVES

B.1 ENCODER SEQUENCE COMPOSITION

At 256×256 resolution, each modality is represented by a 16×16 grid of 256 image tokens. A 75% mask leaves 64 visible tokens for each retained modality. All 256 semantic queries remain active in every training and inference configuration. Table 7 reports the resulting encoder sequence lengths. The token counts exclude the frozen teacher and the decoder-side mask tokens, neither of which is part of the shared encoder sequence.

B.2 MASKING MODES AND ACTIVE LOSSES

Table 8 distinguishes patch masking from complete modality dropping. "Dropped" means that the corresponding image-token stream is omitted entirely from the student encoder. It does not mean that the model is trained to reconstruct a completely absent modality. When EO is dropped from a paired student input, the clean EO observation remains available only to the frozen teacher and supplies semantic targets for the query stream.

Table 6: **Exact frozen few-shot SAR classification results corresponding to Fig. 3 of the main paper.** We freeze the pretrained encoder and optimize only a linear downstream classifier. We report mean OA (%) and sample standard deviation over five matched support draws. sem uses semantic-query features, whereas both concatenates modality-token and semantic-query features. SARMAE* denotes the official checkpoint evaluated with the same downstream protocol. **Bold** and underline denote the best and second-best values within each dataset and shot count, respectively; shaded rows are ours.

Dataset	Method / Feature	1-shot	2-shot	5-shot	10-shot	20-shot	40-shot
FUSAR-Ship	SARMAE*	54.19 ± 8.69	64.95 ± 6.93	70.03 ± 3.48	78.18 ± 3.28	83.44 ± 2.03	88.35 ± 0.83
	SAREO-FM (both)	55.56 ± 7.50	64.79 ± 3.81	72.84 ± 3.97	80.46 ± 1.81	85.32 ± 1.79	$\mathbf{89.81} \pm 0.70$
	SAREO-FM (sem)	$\mathbf{55.79} \pm 9.07$	$\mathbf{68.54} \pm 4.35$	$\mathbf{75.74} \pm 2.46$	$\mathbf{81.65} \pm 2.91$	$\mathbf{85.50} \pm 1.71$	89.56 ± 1.22
SAR-ACD	SARMAE*	$\mathbf{35.09} \pm 7.77$	$\mathbf{39.56} \pm 6.52$	47.12 ± 3.16	51.68 ± 4.08	60.30 ± 2.86	70.11 ± 1.82
	SAREO-FM (both)	33.81 ± 6.30	36.31 ± 6.91	44.78 ± 4.56	51.87 ± 4.12	60.69 ± 2.75	71.35 ± 2.01
	SAREO-FM (sem)	33.64 ± 6.95	36.87 ± 5.83	$\mathbf{48.10} \pm 3.02$	$\mathbf{53.65} \pm 3.19$	$\mathbf{61.65} \pm 3.35$	$\mathbf{71.72} \pm 1.74$

Table 7: **Shared-encoder sequence composition.** Training counts use the 75% patch-mask ratio; downstream inference uses all available image patches.

Configuration	SAR	EO	Queries	Total
Independent/shared masking	64	64	256	384
Complete EO dropping	64	0	256	320
Complete SAR dropping	0	64	256	320
Unpaired SAR training	64	0	256	320
SAR-only inference	256	0	256	512
EO-only inference	0	256	256	512
Joint SAR–EO inference	256	256	256	768

The values in Table 8 are the code- and log-verified sampling probabilities used for the reported checkpoint. Across 357,422 recorded batches, the realized frequencies were $0.3506/0.3506/0.1493/0.1494$, closely matching the configured $0.35/0.35/0.15/0.15$ distribution.

The general formulation can accommodate an unpaired EO corpus, but the SAR-1M Liu et al. (2025) run used in this work consists of paired SAR–EO observations and unpaired SAR observations. We therefore do not present unpaired EO as a sample type instantiated in the reported training run. Likewise, no pixel loss is applied to an image stream that is completely absent from the student encoder.

B.3 One Pretraining Iteration

For completeness, one optimization step can be summarized as follows.

1. Draw either a paired SAR–EO record or an unpaired SAR record from the pretraining corpus.
2. For a paired record, sample one of the four modes in Table 8; for every retained modality, sample an exact 75% patch mask.
3. Concatenate the visible image tokens and all 256 semantic queries, and process them with the shared encoder.
4. For each retained target modality, restore its visible outputs to the spatial grid, insert decoder mask tokens, and predict its image patches with the corresponding modality-specific decoder.
5. If a paired EO observation is available, process its clean, unmasked teacher view with the frozen DINOv3 model, resample the teacher patch grid to 16×16 when needed, and align it patch-wise with the projected query outputs.

Table 8: **Student inputs, teacher availability, and active losses.** Retained image streams are masked by 75%, and all semantic queries remain active. The loss coefficients are $\lambda_s = 1$, $\lambda_e = 1$, and $\lambda_{\text{align}} = 0.5$.

Sample / mode	Prob.	Student encoder image streams	Teacher input	Active objective
Paired, independent masks	0.35	Visible SAR + visible EO	Clean EO	$\mathcal{L}^s_{\text{pix}} + \mathcal{L}^e_{\text{pix}} + 0.5\mathcal{L}_{\text{align}}$
Paired, shared mask	0.35	Visible SAR + visible EO	Clean EO	$\mathcal{L}^s_{\text{pix}} + \mathcal{L}^e_{\text{pix}} + 0.5\mathcal{L}_{\text{align}}$
Paired, complete EO dropping	0.15	Visible SAR only	Clean EO, teacher only	$\mathcal{L}^s_{\text{pix}} + 0.5\mathcal{L}_{\text{align}}$
Paired, complete SAR dropping	0.15	Visible EO only	Clean EO	$\mathcal{L}^e_{\text{pix}} + 0.5\mathcal{L}_{\text{align}}$
Unpaired SAR	–	Visible SAR only	Unavailable	$\mathcal{L}^s_{\text{pix}}$
Unpaired EO	–	Not sampled in this work	–	–

Table 9: **Downstream datasets and interpretation.** OA denotes overall accuracy; micro-AP denotes micro-averaged average precision. The paired EuroSAT protocol uses geospatially matched SAR observations in addition to standard EO imagery; standard EuroSAT itself is EO-only.

Dataset	Input	Classes	Metric	Protocol	SAR-1M relation	Role in this work
MSTAR	SAR	10	OA	1–40-shot / 30%	In corpus	In-corpus target transfer
FUSAR-Ship	SAR	10	OA	1–40-shot / 30%	In corpus	In-corpus ship transfer
SAR-ACD	SAR	5	OA	1–40-shot / 30%	Out of corpus	Primary out-of-corpus SAR test
EuroSAT protocol	EO or SAR	10	OA	1/5/10/40-shot	No out-of-corpus claim	Unimodal scene transfer
NWPU-RESISC45	EO	45	OA	Few-shot / higher supervision	No out-of-corpus claim	EO scene transfer
AID	EO	30	OA	Few-shot / higher supervision	No out-of-corpus claim	EO scene transfer
So2Sat LCZ42	SAR, EO, or joint	17	OA	5-shot, city-disjoint	No out-of-corpus claim	Paired single-label transfer
BigEarthNet-MM	SAR, EO, or joint	19	micro-AP	5-shot, official split	No out-of-corpus claim	Paired multi-label transfer
BRIGHT	Pre-EO + post-SAR	4	Dense labels	Qualitative only	In-corpus source	In-corpus dense diagnostic

6. Sum only the losses activated by Table 8 and update the shared encoder, queries, patch embeddings, decoders, and semantic projector. The DINOv3 Siméoni et al. (2025) teacher remains frozen.

C Data and Evaluation Scope

C.1 Pretraining Corpus Composition

SAREO-FM is pretrained on SAR-1M Liu et al. (2025), which contains 1,312,902 SAR observations aggregated from 18 public datasets. Among them, 1,042,156 have geographically corresponding EO observations and the remaining 270,746 are unpaired SAR observations. The corpus covers 57 target and scene categories, multiple SAR sensors and polarization modes, and ground resolutions from 0.1 to 60 m. Paired EO is used in two distinct ways: as a raw student input when the sampled mode retains EO, and as the clean input to the frozen semantic teacher. Unpaired SAR samples contribute only SAR masked reconstruction.

C.2 Downstream Benchmark Summary

Table 9 summarizes the role of each downstream benchmark. "In corpus" describes image-source membership in SAR-1M, not supervised label leakage: no downstream labels are used during pretraining. We reserve the term "out of corpus" for benchmarks whose source observations are absent from SAR-1M. Benchmarks for which we make no out-of-corpus claim are reported simply as downstream transfer tasks.

C.3 Corpus Membership and Geographic Overlap

MSTAR and FUSAR-Ship are constituents of SAR-1M and are explicitly interpreted as in-corpus representation transfer. SAR-ACD is absent from SAR-1M and provides the primary out-of-corpus SAR classification result. BRIGHT also contributes source imagery to SAR-1M; consequently, Figure 5 of the main paper is an in-corpus qualitative dense probe rather than evidence of unseen-event or unseen-scene generalization. Its purpose is to illustrate that the unified encoder can be

Table 10: **Architecture summary.**

Component	Specification
Input resolution	256×256
Patch size / grid	16×16 / 16×16
SAR / EO tokens	256 per available modality
Shared encoder	ViT-B/16, 12 blocks, $D = 768$
Attention / MLP	12 heads / ratio 4
Semantic queries	256 learnable, spatially indexed
Position encoding	Shared 2D sine–cosine grid
Stream encoding	Learnable SAR / EO / query embeddings
Teacher	Frozen DINOv3-7B patch features
Semantic projector	Two-layer MLP
Reconstruction	Separate SAR/EO decoders

attached to a dense prediction head and can use complementary pre-event EO and post-event SAR evidence.

For paired remote-sensing corpora, exact duplicate detection alone is not sufficient to characterize overlap because resized crops can differ at the byte level while originating from the same geographic tile, parent scene, or event. A complete audit therefore uses source identifiers, geographic coordinates, parent-scene identifiers, and event identifiers in addition to file hashes. We do not use EuroSAT, So2Sat LCZ42, BigEarthNet-MM, or BRIGHT to support an out-of-corpus claim unless such a source-level exclusion is established.

C.4 Paired-Dataset Splits

So2Sat LCZ42 is evaluated using a city-disjoint split so that geographic regions do not cross the downstream train and test partitions. For each few-shot draw, the same class-balanced support identifiers are used by all compared input modes and feature readouts. BigEarthNet-MM contains 519,284 paired observations in the protocol used here, partitioned into 269,695/123,723/125,866 train/validation/test samples, respectively. We map the labels to 19 classes and report micro-AP. Support sampling is performed only within the training partition; validation is used for downstream model selection, and the test partition remains fixed.

The "SAR input" rows associated with EuroSAT use a geospatially paired SAR companion rather than the EO-only EuroSAT release by itself. The accompanying dataset manifest records the companion release, VV/VH channel ordering, geographic matching key, and paired-location split. These identifiers are treated as part of the evaluation protocol rather than as an implementation detail.

D Architecture and Pretraining

D.1 Architecture Specification

The shared encoder follows ViT-B/16 Dosovitskiy et al. (2020): 12 Transformer blocks, hidden dimension $D = 768$, 12 attention heads, and an multi-layer perceptron (MLP) expansion ratio of 4. SAR and EO use separate linear patch projections but share the same 16×16 two-dimensional sine–cosine positional grid. Learnable stream embeddings distinguish SAR tokens, EO tokens, and semantic queries. A spatially organized bank of 256 learnable queries uses the same positional grid as the image tokens, establishing one query per patch location.

The frozen DINOv3-7B Siméoni et al. (2025) teacher receives a clean EO view and supplies final-layer patch features. If its native token grid differs from 16×16, the patch features are spatially resampled to the query grid before patch-wise cosine alignment. A two-layer MLP projects each query output from the encoder dimension to the teacher-feature dimension. SAR and EO have separate reconstruction decoders and mask tokens.

Table 11: **Lifecycle of SAREO-FM components.**

Component	Pretraining	Downstream
Shared SAREO encoder	Trainable	Used
Semantic queries	Trainable	Used
SAR/EO patch embeddings	Trainable	Used as available
SAR/EO reconstruction decoders	Trainable	Discarded
Semantic projector	Trainable	Discarded
DINOv3-7B teacher	Frozen	Discarded

Table 12: **Image-level readout dimensions.**

Available input	mod	sem	both
SAR or EO	768	768	1,536
SAR + EO	1,536	768	2,304

D.2 Component Lifecycle

Table 11 separates pretraining-only modules from the deployed foundation model. The DINOv3 teacher, semantic projector, and both reconstruction decoders are removed after pretraining. Therefore, the large teacher contributes neither parameters nor computation to downstream inference.

D.3 Pixel and Semantic Targets

For each retained modality, the reconstruction decoder predicts all 256 patches, but the ℓ_1 pixel objective is evaluated only on the masked locations. The SAR and EO losses are not compared numerically across modalities because their target spaces and normalization differ. For semantic alignment, the projected query at each spatial location is matched to the corresponding teacher patch by cosine distance. All 256 queries participate in semantic alignment, including locations whose SAR or EO image patches are hidden from the student. This design encourages the query grid to infer spatial semantics from the visible evidence and context rather than copy a teacher feature from an identically visible patch.

E Downstream Protocol

E.1 Feature Readouts

During downstream transfer, all available image patches and all semantic queries are passed to the shared encoder without masking. Let $\mathcal{A}$ denote the set of available modalities and let $Z_m = \{z_m^i\}_{i=1}^N$ and $Z_q = \{z_q^i\}_{i=1}^N$ be the corresponding encoder outputs. We first pool within each modality and then concatenate the available modality streams:

$$\bar{z}_m = \frac{1}{N}\sum_{i=1}^{N} z_m^i, \tag{13}$$

$$g_{\mathrm{mod}} = \operatorname*{concat}_{m\in\mathcal{A}} \bar{z}_m, \tag{14}$$

$$g_{\mathrm{sem}} = \frac{1}{N}\sum_{i=1}^{N} z_q^i, \tag{15}$$

$$g_{\mathrm{both}} = [g_{\mathrm{mod}}; g_{\mathrm{sem}}]. \tag{16}$$

Table 12 gives the resulting dimensions. In particular, mod is not an average across modalities: it is 768-D for unimodal input and 1,536-D for paired input. The semantic readout remains 768-D, making both 1,536-D or 2,304-D, respectively.

For K classes, the corresponding linear heads contain $K(768|\mathcal{A}|+1)$, $K(768+1)$, and $K(768(|\mathcal{A}|+1)+1)$ parameters for mod, sem, and both, respectively. We therefore treat both as a complementary representation and do not attribute every difference to feature content alone.

F Inference Cost

The deployed model contains the shared encoder, the two patch embeddings, the persistent semantic queries, and the selected downstream head. Teacher features are never computed at transfer time, and neither the semantic projector nor the reconstruction decoders are loaded for normal downstream inference. The exact shared-encoder sequence length is 512 tokens for either unimodal input and 768 tokens for joint input (Table 7). These sequence lengths explain why unimodal and joint inference should be profiled separately.

Measured latency, throughput, peak memory, and FLOPs depend on hardware, precision, batch size, and kernel implementation. We therefore report them only under one explicitly fixed profiling protocol rather than extrapolating from token counts. The checkpoint-only workflow uses 256×256 inputs, evaluation mode, inference-only execution, fixed warm-up and measurement iterations, explicit GPU synchronization, and identical precision across modes. It reports the deployed encoder separately from pretraining-only parameters.

G Limitations

Our quantitative evaluation focuses on image-level classification and multi-label recognition. The BRIGHT examples demonstrate compatibility with a dense head but are qualitative, in-corpus diagnostics and should not be interpreted as a complete segmentation benchmark. The model also assumes that paired SAR and EO observations are sufficiently registered for patch-wise interaction and teacher alignment. Geographic misregistration, acquisition-time differences, clouds, SAR layover, and sensor-dependent resolution can weaken local correspondence. Finally, the semantic target is obtained from an EO-pretrained VFM; although SAR reconstruction preserves sensing-specific information, semantics that are poorly represented by the EO teacher may be transferred less effectively.