Surya 2.0: Beyond 3 Million Tokens - A Multimodal Foundation Model for Full Resolution Heliophysics

Sujit Roy[1,2,*] **Johannes Schmude**[3,*] **Ata Akbari Asanjan**[4,2,*] **Thorsten Kurth**[5,*]

Rohit Lal[1,2,*] **Kshitiz Mandal**[1,2] **Vishal Gaur**[1,2] **Harris Abdul Majid**[6]

Nikolaos Dionelis[1,2] **Berkay Aydin**[7] **Himanshu Patil**[1,2] **Andrés Muñoz-Jaramillo**[8]

Campbell Watson[3] **Juan Bernabe-Moreno**[3] **Manil Maskey**[2] **Rahul Ramachandran**[2]

[1]Earth System Science Center, University of Alabama in Huntsville, AL, USA
[2]IMPACT, Office of Data Science and Informatics (ODSI)/NASA MSFC
[3]IBM Research [5]NVIDIA Corporation, Santa Clara, CA, USA
[4]Research Institute for Advanced Computer Science, Universities Space Research Association, USA
(work done with IMPACT, ODSI, NASA MSFC)
[6]The University of Edinburgh [7]Georgia State University, Atlanta, GA, USA [8]Southwest Research Institute, Boulder, CO, USA

[*]Equal contribution.
Correspondence: `sujit.roy@nasa.gov`

Abstract

Context length remains a fundamental bottleneck for vision foundation models operating on high-resolution imagery. Current architectures rarely exceed 1M tokens, forcing practitioners to downsample inputs at the cost of fine-grained spatial information. This limitation is particularly acute in heliophysics, where satellite instruments record full-disk solar observations at native 4096×4096 resolution across 13 channels, and downsampling discards the small-scale magnetic structures and localized dynamics critical to understanding solar phenomena. In this work, we present the first vision foundation model for heliophysics trained at native 4K resolution on approximately 15 years of multimodal solar data (257 TB) spanning eight extreme ultraviolet channels and five magnetic field and velocity products from the Solar Dynamics Observatory (SDO). We adapt MultiMAE with a dual-view formulation: 25% of tokens are observed directly, while the remaining 75% are replaced with fixed Gaussian random Fourier projections that act as structured non-invertible frequency views of the masked content. Using two-way parallelism with FP8 mixed precision, we scale the attended sequence to > 3 million tokens at 8×8 patch size, a $3\times$ advance over prior work, and demonstrate strong scaling efficiency up to 1024 GPUs across five hardware configurations (pretraining itself uses 128 GPUs), with an architecture capable of processing sequences exceeding 10M tokens. The learned representations achieve strong zero-shot reconstruction across masking ratios up to 90% and full missing-modality scenarios. On four SuryaBench downstream tasks (active region segmentation, flare forecasting, EVE irradiance, and solar wind) with a frozen encoder, one input channel and 10% of the labels, we achieve the best mean on every metric against existing baselines. Code available at `https://github.com/NASA-IMPACT/Surya-2.0`

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

1 Introduction

The Sun is a highly complex dynamical system whose behavior is driven by the continuous emergence and evolution of magnetic flux. Solar activity directly governs heliospheric dynamics and drives the space weather conditions that impact Earth and near-Earth space. Extreme events such as solar flares and coronal mass ejections (CMEs) originate from localized fine-grained magnetic anomalies but yield far-reaching global consequences, ranging from disrupting satellite avionics and communication networks to threatening terrestrial power grids [1]. Understanding and forecasting these phenomena requires modeling the Sun holistically. Therefore, models need to capture both the global context of the full solar disk and the fine-scale, cross-modal interactions among the various modalities representing multi-spectral observations [2].

Modern space-borne observatories, notably the Solar Dynamics Observatory (SDO) [3], capture these dynamics with unprecedented fidelity. SDO's Atmospheric Imaging Assembly (AIA) and Helioseismic and Magnetic Imager (HMI) instruments have accumulated more than 12 PB of synchronized multimodal imagery over past 15 years [2]. This unprecedented volume of data provides an ideal substrate for deep learning methods to automatically discover hidden patterns, extract subtle signals from noisy observations [4, 5], and transform scientific analysis. However, extracting actionable physics from this scale of data remains a formidable computational and scientific challenge.

While recent foundation models [2, 6, 7] have successfully demonstrated the viability of representation learning on solar data, they remain fundamentally constrained by sequence length bottlenecks. As established in computer vision, small patch sizes are crucial for capturing fine-grained spatial details [8–10]. In the context of SDO, a single AIA pixel represents a 435×435 km region on the solar surface. Consequently, even the 16×16 patches utilized in recent state-of-the-art models correspond to massive $6{,}960 \times 6{,}960$ km regions. This coarse spatial granularity physically limits a model's ability to represent the small-scale magnetic structures, fine-grained loops, and localized topological dynamics that act as precursors to major solar phenomena. Furthermore, to cope with memory limits, current models often conflate modalities by treating all 13 observational channels as a single stacked tensor. This prevents flexible per-channel downstream analysis and forces a monolithic inference process. The scientifically motivated approach is to treat each channel as an independent high-resolution modality to preserve exact spatial granularity. However, modeling 13 independent channels natively at an 8×8 patch size scales the required context window to over 3 million tokens, which is a sequence length regime previously inaccessible to vision architectures.

Furthermore, scaling the architecture is only half of the challenge; the model must also be equipped with a pretraining objective capable of extracting physically meaningful representations from such massive data. Standard masked autoencoding frameworks, such as MultiMAE [11], drop masked tokens entirely. While computationally efficient, this approach deprives the encoder of dense gradient signals and relies heavily on low-frequency spatial interpolation, which fails to capture the intricate cross-modal dependencies between magnetic fields and thermal plasma required in heliophysics.

To overcome these limitations, we introduce Random Fourier Projection (RFP) into the MultiMAE framework. Instead of dropping tokens, we replace 75% of them with a fixed non-invertible Gaussian Fourier projection that preserves coarse frequency structure, while impeding finer spatial detail. This creates a tighter and more physically meaningful information bottleneck that prevents trivial spatial interpolation, forcing the encoder to instead fuse heterogeneous views to learn robust cross-modal and cross-spatial representations.

In this work, we present the first vision foundation model for heliophysics trained at native 4096×4096 resolution with a context window exceeding 3 million tokens. Our main contributions are:

- **Integrating RFP in MultiMAE:** We integrate fixed random Gaussian Fourier projection into the MultiMAE framework. By replacing masked tokens with structured non-invertible frequency views, we force the encoder to learn cross-modal representations rather than relying on low-frequency interpolation. We validate this with an ablation study showing consistent gains over the standard pretraining MultiMAE architecture across all 13 modalities.

- **Context Scaling:** Utilizing two-way model parallelism (feature and sequence) combined with FP8 mixed precision, we scale the patch size down to 8×8 pixels on native 4096×4096 imagery. By

processing each of the 13 channels via independent adapters, our architecture successfully scales to large sequence lengths that exceed 3 million tokens.

- **Large-Scale Pretraining:** We train on approximately 15 years of multimodal SDO data (257 TB), representing the largest heliophysics foundation model training to date in terms of resolution, context length, and data volume. We demonstrate high-fidelity reconstruction of inherently noisy HMI magnetograms and high robustness to complete sensor dropout.
- **Label-Efficient Transfer on Four Downstream Tasks:** With a frozen encoder, a single AIA channel and 10% of the training labels, we evaluate on four SuryaBench tasks and baselines. Our model attains the best mean on every metric, raising flare-forecasting TSS from 0.397 to 0.557, lowering EVE irradiance MAE by 19.9%, and improving AR segmentation Dice by 1.3 points, while matching Surya on solar wind speed prediction.

2 Related Work

Task-Specific Architectures. Driven by the proliferation of high-resolution satellite observations, heliophysics has entered a heterogeneous big data era [2, 12–14]. Initial deep learning efforts in the field primarily focused on task-specific models tailored for localized problems such as solar image classification [15], event/object detection [16–19], image generation [20–23], and various space weather forecasting tasks [24–27]. However, managing the extreme memory requirements of full-resolution multi-modal solar rasters proved highly challenging for these conventional architectures. Consequently, many traditional approaches downsample images or restrict their analysis to small isolated patches of ARs [28, 29, 25, 30]. While effective for narrow tasks, this spatial reduction discards the crucial multiscale physics required to connect localized magnetic flux emergence with global coronal dynamics. By operating at native 4096×4096 resolution across the full solar disk, our approach preserves this essential multiscale physics without relying on spatial cropping.

Early Foundation Models. To address the limitations of task-specific training, recent research has shifted toward self-supervised learning and foundation models capable of learning generalized representations of solar behavior. Early models, such as SDO-FM [7] and Solaris [6], successfully demonstrated the viability of pretraining on solar observational data. These architectures typically rely on downscaled imagery, often operating at resolutions of 512×512 pixels, and restrict their inputs to a subset of AIA channels. As summarized in Appendix Table 6, this downsampling results in relatively short context lengths of under 10k tokens. While computationally efficient, operating at this reduced resolution prevents the model from resolving the fine-scale magnetic phenomena that govern broader solar activity, particularly when the critical vector magnetic fields provided by HMI are omitted. We bridge this gap by integrating all 13 (both AIA and HMI) channels as independent modalities, natively preserving the distinct physical pathways of each sensor.

Native-Resolution Pretraining and Its Limits. Advancing beyond downsampled architectures, Surya [2, 31] introduced the first native-resolution foundation model for heliophysics. Operating at the full 4096×4096 scale, Surya utilizes a spectral transformer to process SDO AIA and HMI data. Crucially, Surya operates as an auto-regressive model, optimizing for next-time-step prediction to learn temporal dynamics. While forecasting is a valuable downstream application, an auto-regressive objective encourages the network to prioritize temporal continuity, often at the expense of resolving the intricate instantaneous physical dependencies between the different solar layers. Furthermore, to manage the computational load, Surya concatenates all 13 observational channels into a single combined input tensor. This design conflates instrument-specific characteristics, forces a monolithic inference process, and relies on coarse 16×16 patch sizes that blur massive $6{,}960 \times 6{,}960$ km regions, obscuring the fine-grained magnetic topologies that drive and characterize solar flares and CMEs. To rectify this, we push the patch size down to 8×8 pixels via two-way model parallelism and replace auto-regressive forecasting with a non-invertible RFP method that explicitly forces the synthesis of high-fidelity cross-modal features. Ultimately, by scaling the context window beyond 3.4 million tokens, and also by using our frequency-based RFP, our architecture extracts the instantaneous multiscale physics of the Sun without compromise.

In summary, SDO-FM established SDO-scale pretraining at reduced resolution on AIA only, Surya reached native resolution with an autoregressive objective on a stacked 13-channel tensor, and our work adds native-resolution pretraining with each channel as an independent modality. Per-channel adapters are a usability requirement rather than a performance choice: most heliophysics analyses

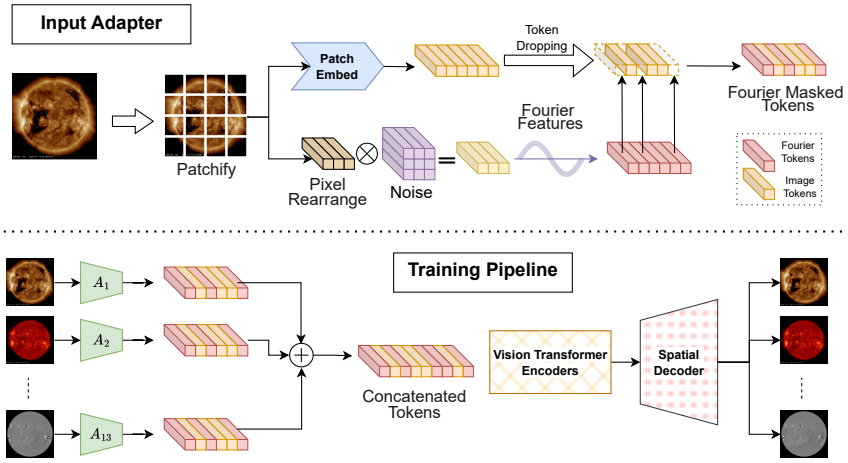

Figure 1: Architecture of our multi-modal heliophysics foundation model. The input adapter processes a single time step and separates it into the 13 sensor modalities. Each modality is independently patchified and mapped to token representations. A subset (25%) of the tokens is retained directly, while the remaining 75% are transformed using Gaussian random Fourier projection, resulting in tokens that are decorrelated from the original spatial content and act as structured noise views under our formulation. The visible and RFP-transformed tokens are then combined within each modality and jointly processed by a shared Vision Transformer (ViT) backbone. A unified decoder reconstructs all the 13 modalities, SDO AIA and HMI, simultaneously from the shared latent representation.

use one channel or a small subset, a stacked tensor makes channel availability all-or-nothing, and independent modalities enable the missing-modality evaluation of Sec. 4, which no prior heliophysics foundation model reports (Table 6).

3 Methodology

3.1 SDO data

We define a solar observation as a co-registered multi-channel snapshot from the two imaging instruments of SDO [3, 31]: (i) AIA, which captures extreme ultraviolet and ultraviolet photometric imagery of the solar atmosphere at distinct temperature regimes, and (ii) HMI, which measures photospheric vector magnetic fields and line-of-sight plasma velocities. Each observation is a tensor $X \in \mathbb{R}^{13\times4096\times4096}$, comprising 8 AIA channels (94, 131, 171, 193, 211, 304, 335, 1600 Å) and 5 HMI channels (B_{los}, B_x, B_y, B_z, V_{los}) at native $0.6''$/pixel resolution (from AIA) with a standardized 12-minute cadence. For further details on data preprocessing and normalization, see Appendix A.

3.2 Architecture

Model overview. Our multi-modal architecture in Figure 1 learns cross-modal interactions through a shared Vision Transformer (ViT) backbone and per-modality reconstruction. The proposed model uses Random Fourier Projection (Section 3.2), and treats each of the 13 SDO channels as an independent modality with per-channel input adapters, rather than combining them into a single tensor.

Random Fourier Projection (RFP). We train a model that learns spatial representations from single-frame solar imagery. The encoder is designed using a dual-view formulation, where each patch is observed through one of two complementary representations derived from the same underlying content. With reference to classical masked autoencoding (e.g., [11, 32, 33]), we replace token dropping with a frequency-based noise alternative. MultiMAE [11] is a multi-modal masked autoencoding framework that makes use of a reconstruction loss and learns cross-modal interactions among different modalities.

One of our main contributions is the utilization of RFP within this framework. Incorporating frequency-domain representations via spectral blocks or the Fourier transform [2, 34] applied to solar

rasters enables the extraction of localized high-frequency components, sharp edges, and fine-grained structural details, while also revealing spatial coherence, anisotropy, multi-scale organization, and subtle morphological differences that are not easily distinguishable in the spatial domain alone [35].

Problem setup. Let $x \in \mathbb{R}^{C\times H\times W}$ denote an input image. We partition x into non-overlapping patches and flatten each patch into a vector $\mathbf{z}_i \in \mathbb{R}^{P_d}$, where i indexes spatial location. P_d is the patch dimensionality. For each patch, we define two representations to map $\mathbf{z}_i$ into a shared embedding space $\mathbb{R}^D$. The first is a learnable linear embedding,

$$\mathbf{e}_i^{\text{vis}} = \mathbf{z}_i E, \qquad E \in \mathbb{R}^{P_d\times D}, \tag{1}$$

which preserves fine-grained local information and supports accurate reconstruction. The second is a fixed nonlinear representation obtained via RFP, intended to provide a compressed and non-invertible view of the same patch content.

To define the Fourier-based representation, we sample a random projection matrix,

$$B = [b_1, \ldots, b_m] \in \mathbb{R}^{P_d\times m}, \qquad b_j \sim \mathcal{N}(0, \sigma^2 I), \tag{2}$$

once at initialization and keep it fixed throughout training.

Given a patch vector $\mathbf{z}_i$, we compute:

$$\mathbf{u}_i = \mathbf{z}_i B \in \mathbb{R}^m, \tag{3}$$

and define the Random Fourier representation:

$$\mathbf{e}_i^{\text{rfp}} = \phi(\mathbf{z}_i) = [\sin(2\pi\mathbf{u}_i),\ \cos(2\pi\mathbf{u}_i)] \in \mathbb{R}^{2m}, \tag{4}$$

where $2m = D$ by construction. No learnable parameters are involved in this mapping, and the Fourier projection is applied directly to the raw patch vector $\mathbf{z}_i$, not to the output of any learnable transformation. In the expectation over the random draw of B, inner products in the RFP space approximate a shift-invariant kernel on the original patch space. This interpretation is not used explicitly in training but provides intuition for why the representation preserves coarse similarity structure while discarding fine detail. For additional implementation details, including ablation studies and an analysis of reconstruction fidelity, please refer to Appendix J.

The scale σ controls the frequency bandwidth of this kernel: $\langle\phi(z_i), \phi(z_j)\rangle \approx \exp\!\left(-2\pi^2\sigma^2 \left\|z_i - z_j\right\|^2\right)$. Larger σ places more spectral mass at high frequencies and makes the map harder to invert. We fix $\sigma = 10$, placing the projection in the maximally non-invertible regime where per-token spatial information cannot be recovered token-by-token. As shown in Appendix Figure 13, the log-spectral distance between the original and RFP radial power spectra increases monotonically with σ from ≈ 0.79 at $\sigma = 0.005$ to ≈ 2.42 at $\sigma = 5$ and saturates beyond $\sigma \approx 0.5$ across all 13 SDO channels; crucially, the RFP retains learnable structure even in this saturated regime, since spectral mass is redistributed from low to high radial frequency rather than collapsed to noise.

During training, exactly one of the two representations is selected for each patch:

$$\mathbf{e}_i = \begin{cases} \mathbf{e}_i^{\text{vis}}, & \text{with probability } \rho, \\ \mathbf{e}_i^{\text{rfp}}, & \text{with probability } 1-\rho, \end{cases} \qquad \rho = 0.25. \tag{5}$$

This stochastic view selection is performed independently across patches and training iterations. As a result, the encoder observes a mixture of high-fidelity tokens and frequency-compressed tokens, both anchored to the same underlying image but differing in their information content.

The Fourier-based representation is inherently lossy for three reasons: (1) it reduces dimensionality from P_d to D, (2) relies on a fixed random projection, and (3) uses periodic sinusoidal mappings that introduce global non-injectivity.

Thus, individual RFP-transformed tokens cannot be inverted to recover the patch exactly, while still retaining sufficient structure to guide representation learning when combined with the visible tokens and spatial context.

Both representations are augmented with a shared 2D Fourier positional encoding. Let $c_f^x = \cos(2\pi f x_i)$ and $s_f^x = \sin(2\pi f x_i)$, and similarly for y_i. The positional embedding is:

$$\mathbf{p}_i = \bigoplus_{f\in\mathcal{F}} \left[c_f^x,\ s_f^x,\ c_f^y,\ s_f^y\right] \in \mathbb{R}^D, \tag{6}$$

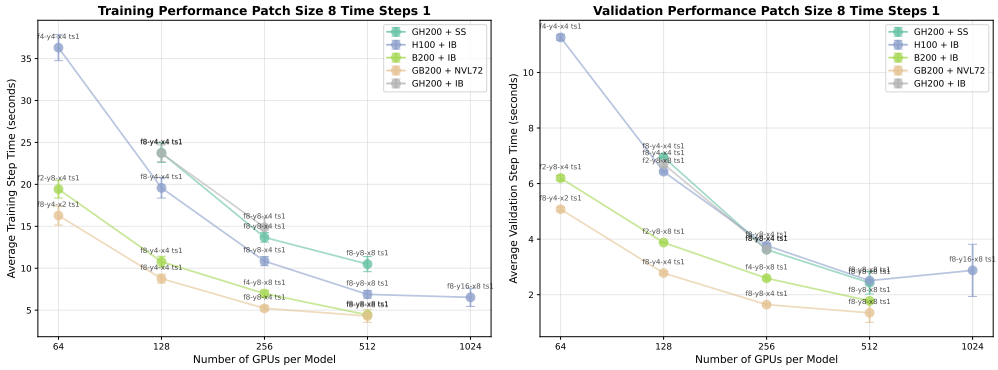

Figure 2: Single batch step time for training (left) and validation (right) vs number of GPUs for different GPU and network interconnect architectures with context tokens > 3M and patch size 8×8.

where we use the direct sum operator, and $\mathcal{F} = \{1, \ldots, G/2\}$ contains $D/4$ linearly spaced frequencies.

$G = H/P$ is the number of patches per spatial dimension. This deterministic encoding is added to both $\mathbf{e}_i^{\text{vis}}$ and $\mathbf{e}_i^{\text{rfp}}$, ensuring that the spatial location information is preserved. The encoder is trained jointly with a decoder to reconstruct the original image from the mixed-view token sequence, minimizing the Mean Squared Error (MSE) loss. Reconstruction depends on integrating information across heterogeneous representations and across space, rather than on per-token inversion. This training regime can be interpreted as a multi-view self-supervised objective, where different representations provide complementary, but individually insufficient, information about the same signal.

3.3 Scaling-Up Pretraining: Model Parallelism

The sequences processed by our model are long, leading to large memory requirements. This limits the size of the model which can be hosted on a single GPU. Both memory and computational complexity scale inverse-quadratically with patch size, e.g. leading to four times longer sequences when scaling from patch size 16×16 to patch size 8×8.

To mitigate both issues, we employ two-way model parallelism: feature parallelism and sequence parallelism, both of which are standard in contemporary LLM implementations (e.g., Megatron-LM [36, 37]). We discuss each in turn.

Feature parallelism: We split the embedding dimension across the attention and MLP layers in all Transformer blocks. Feature parallelism follows a fork-join pattern, enabled by the fact that both attention and MLP layers begin and end with a linear projection. In attention layers, these correspond to the QKV and output projections. The input and output of each layer remain serial in the feature dimension, while intermediate representations are projected into a feature-split space via column-wise parallel matrix multiplication. The final projection recombines the shards using row-wise parallel matrix multiplication. Importantly, gradients leaving the column-wise multiplication must be reduced across the feature-split communication group.

Feature parallelism is most beneficial if embedding dimensions are large. In our setting, the embedding dimension size is moderate, which limits the benefits of feature parallelism since smaller local matrix sizes underutilize the GPU. Nevertheless, we observe near-linear speedups with a feature-parallelization degree of 2 or 4 for patch sizes of 16×16. For smaller 8×8 patches, 8-way feature parallelism is more effective as shown in Figure 2. This is largely because attention layers are embarrassingly parallel in the feature dimension, as parallelism is performed across heads. Therefore, no communication in feature direction is performed in the attention step. For models with long sequences, where the attention layer dominates, the speedup is almost linear.

Sequence parallelism: This mode resembles classical domain decomposition used in HPC applications. We partition the spatial domain along x and y into equally sized chunks (N_x and N_y), which implicitly splits the Transformer sequence into $N_x \times N_y$ segments. Operations such as MLP layers act on a per-token basis and are thus trivially parallel in the sequence dimension. However, attention layers

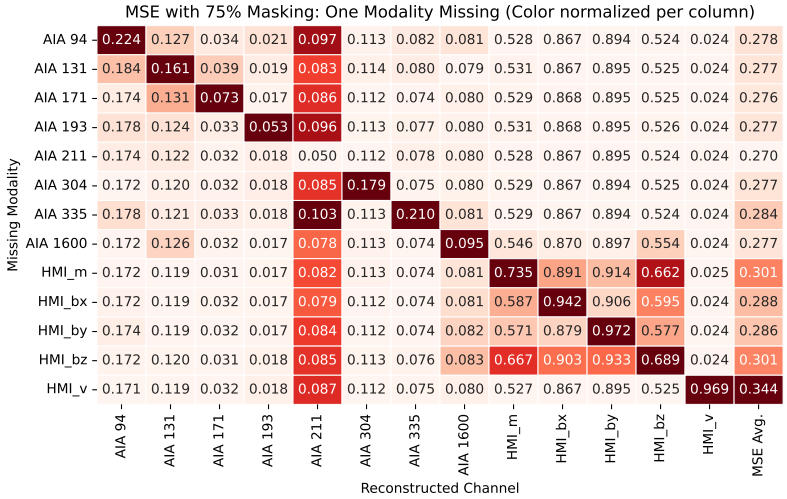

Figure 3: Reconstruction performance (MSE) with 75% masking when a modality is missing. Color normalized per column, per reconstruction modality. Rightmost column is mean reconstruction error.

couple every token to every other token in the global sequence, requiring a sequence-parallel attention formulation.[1]

To facilitate the aforementioned model parallelism efficiently, we use the distributed layers from the NVIDIA Transformer Engine (TE) library [38]. While TE supports the most performance critical layers of our model, such as attention, some of our layers still need custom parallelism implementations. Notably, those are the input and output adapters which employ masking. To optimize performance, we implement rank-local masking by processing tokens on a per-shard basis rather than performing global masking and reshuffling. While rank-local masking enforces a more uniform distribution of masked regions across shards compared to global masking, we observed no degradation in performance.

For MLP and Attention, feature parallelism is facilitated by the `tp_group` argument and related parameters. TE takes care of all relevant gradient reductions related to data gradients using all-reduce communication patterns. To facilitate context parallelism, we use the TE's `cp_group` attribute within attention layers, effectively implementing our sequence parallelism strategy. This is implemented by a distributed version of FlashAttention [39][2], using a round-robin subsequence exchange based on NCCL point-to-point (p2p) communication (sometimes referred to as send-receive communication pattern). For each round-robin step, all individual p2p communications are bundled into a single big batch to mitigate NCCL overhead and enforce lightweight synchronization between each exchange step. This significantly reduces performance variability at larger scales. Another benefit from using TE for our performance critical layers is that it supports FP8 (as well as FP4), precision arithmetic. This is enabled via autocasting and only eligible layers are converted. Most of our layers operate in BFLoat16 with the exception of MLP and Attention and some other linear layers, which can operate in FP8. We also experimented with FP4 precision arithmetic on NVIDIA Blackwell architectures, but those is currently limited to MLPs. Since our MLPs are not performance critical, we do not observe significant speedup from FP4 precision, and we thus employ BF16+FP8 mixed precision.

Figure 2 displays strong scaling performance of our model for patch size 8×8 with FP8 attention and MLP precision for a single global sample and varying GPU counts. We also tested different GPU architecture generations and network interconnects. We observe that the model heavily benefits from NVLink 72 (NVL72) system architectures, because p2p communication time is significantly reduced (attributed to improved all-to-all/p2p bandwidth and lower latency). This effect becomes more subtle when running the model across more NVLink islands, because in this case the IB connection between the islands becomes the bottleneck. Therefore, at around 512 GPUs the performance on the GB200 + NVL72 and B200 + IB system is comparable. In general, our code performs better for larger NVLink islands, as shown by the improved performance of H100 + Infiniband (IB) compared to GH200 +

[1] Note that in LLM jargon, sequence parallelism is commonly referred to as context parallelism.

[2] We effectively use a fused kernel variant which is offered by TE; the general principle is the same as that of FlashAttention.

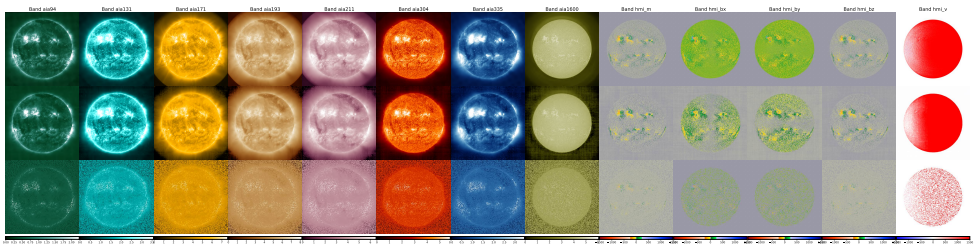

Figure 4: Zero-shot reconstruction of all modalities including AIA and HMI with 75% masking. Top row: Target prediction; Middle row: Generated reconstruction; Bottom row: Input to the model.

IB and GH200 + Slingshot (SS). The attention kernel of transformer engine is highly optimized, so the ring exchange step becomes more expensive than the local attention step around local sequence lengths of around 10K tokens. This limits the strong scalability in sequence dimension. However, scalability can be improved by increasing the number heads and thus employing a higher degree of feature parallelism as well as data parallelism.

Leveraging these architectural optimizations, we pretrain our multi-modal foundation model on 128 NVIDIA GH200 GPUs. Pretraining took 260 wall-clock hours on 128 GH200 GPUs (33,280 GPU-hours). The 1024-GPU runs in Figure 2 show how far the recipe scales; they are not required to train or use the model, and all downstream experiments run on a single GPU.

To accommodate the massive memory requirements, where even a coarser 16×16 patch size generates $851,968$ tokens ($65,536$ tokens for each of the 13 modalities), we actively employ the aforementioned feature and sequence parallelism (tensor parallelism) alongside mixed precision. The model is trained with a micro-batch size of 1 per GPU, achieving an effective global batch size of 128. Throughout pretraining, we optimize the network using AdamW paired with a cosine annealing learning rate schedule. For additional implementation details about model parallelism, please refer to Appendix C.

4 Evaluation and Results

In this section, we evaluate our foundation model capabilities and explore its learned representations. First, we analyze the effect of patch size scaling and compare our reconstruction results with existing methods. Second, we assess the model's ability to use a few modalities to infill entirely missing modalities. Third, we evaluate the transferability of the frozen latent space to four downstream tasks: Active Region (AR) segmentation, solar flare forecasting, solar wind speed prediction and Extreme Ultraviolet Variability Experiment (EVE) spectra prediction. Finally, we probe the latent space using multiple decoders to understand the optimization dynamics of large-scale multi-modal pretraining.

Patch-Size Scaling. To establish the usefulness of scaling the context window to 3 million tokens, we first evaluate the impact of patch size on model performance. Under identical training conditions, shifting the architecture from a 32×32 to a 16×16 patch size yields a 14.3% reduction in validation loss within 5 epochs (Appendix F). This improvement confirms that our architecture follows the vision patchification scaling law $\mathcal{L} \propto N_{\text{tok}}^{-\alpha}$ characterized by [8]. Moving to our final full-scale 8×8 model yields proportionally better loss.

Reconstruction Performance. We evaluate zero-shot reconstruction of high-resolution solar representations under a 75% masking rate. Direct comparisons with prior models are difficult due to differences in architectures, resolutions, pre-training objectives, alongside the lack of a public model weights release (e.g., Solaris [6]). To enable a head-to-head comparison with available open-source models, (1) we adapt the native-resolution Surya model [2, 31] for current-timestep prediction (rather than its default autoregressive forecasting) and (2) downsample our outputs to match the 512×512 resolution constraints of SDO-FM. As detailed in Table 1, our Random Fourier Projection (RFP) pretrained model significantly outperforms these baselines. It reduces the mean AIA reconstruction error by an order of magnitude compared to Surya (339.9 versus 11,537.5) and substantially improves upon SDO-FM (23,200.3). Ultimately, our method establishes a new reconstruction benchmark while effectively scaling to 4096×4096 resolution with joint AIA and HMI support.

Effect of Missing Modalities. In operational space weather forecasting, sensor dropouts and instrument calibrations frequently result in missing data. To emulate this, we performed an experiment with complete masking of few channels while maintaining the 75% spatial masking rate. The omission of individual AIA channels (e.g., 94, 131, or 171 Å) results in only marginal increases in the average MSE, indicating that the model successfully exploits the abundant cross-modal relationship. As seen in Figure 3, the architecture is highly sensitive to the loss of specific HMI components, particularly the line-of-sight velocity field (`hmi_v`). The removal of `hmi_v` pushes the average MSE to 0.344. This discrepancy aligns with the underlying solar physics: the velocity field represents a distinct kinematic modality with almost no direct visual proxies in the AIA thermal passbands. Complete degradation matrices are provided in Appendix L.

To test the practical utility of these representations, we evaluate the frozen encoder on four downstream tasks from SuryaBench [40] (Tables 2–5)

Downstream protocol. We re-ran every baseline ourselves under a shared protocol. All models use the published task-specific SuryaBench train/validation/test splits, which are temporally separated with leakage buffers, and the same fixed 10% subset of the training labels per task, probing the practically relevant low-label regime. Every model receives the same single AIA 94 Å channel and is trained for the same number of epochs. Our encoder is frozen and only a lightweight task head is trained; Surya [2] is adapted with LoRA [41], as in its original paper; SegFormer [42], U-Net [43], ResNet [44] and AlexNet [45] are trained end-to-end. We report mean ± standard deviation over 3 seeds, which vary initialization and data order.

Active Region Segmentation. With a two-layer convolutional head, our model reaches 61.12 Dice and 44.01 IoU (Table 2). Both transformers outperform the CNN by a wide margin (Surya and SegFormer each gain about 18 Dice points over U-Net), and our frozen encoder improves on the LoRA-adapted Surya by a further 1.3 Dice and 1.4 IoU points, a gap larger than the seed-to-seed spread of either model.

Solar Flare Forecasting. The task is binary: predict whether an M1.0-class or stronger flare occurs within the 24 h following the observation. Our model raises TSS from 0.397 (Surya) to 0.557 and recall from 0.424 to 0.644 (Table 3). It is also the most stable across seeds (TSS std 0.056 vs. 0.098–0.164 for the baselines), which matters for a rare-event task where only a fraction of the labelled examples are positive.

Solar Wind Speed Prediction. The model regresses the solar wind speed measured at L1 (OMNI) four days after the observation. Both foundation models are far ahead of the CNNs (over 40% lower MAE than the best CNN, ResNet-18), while our model and Surya are within one standard deviation of each other (MAE 0.125 vs. 0.127; Table 4). We therefore read this task as parity with Surya rather than an improvement.

EVE Spectra Prediction. We predict the full 1343-channel EVE irradiance spectrum [46] one hour ahead, using a Perceiver-style attention head [47]. Against Surya, our model reduces mean MAE by 19.9%, MAPE by 12.9% and RMSE by 10.3% (Table 5), with the tightest spread on RMSE and MAPE; Surya's spread is tighter on MAE. With three seeds we emphasise effect sizes and run-to-run variation rather than definitive significance claims.

Across all four tasks the frozen encoder, given a single channel and 10% of the labels, attains the best mean on every metric, indicating that native-resolution pretraining yields a feature space that transfers to dense prediction, rare-event classification and scalar and spectral regression even when downstream input modalities are severely restricted.

Latent Modality Entanglement. To probe the internal structure of our shared latent space, we designed a controlled decoder-specialization experiment (Appendix M). We froze the multi-modal encoder and trained two independent decoders from scratch, one for AIA and one for HMI. The two decoders behaved very differently. The AIA decoder converged smoothly, while the HMI decoder showed a sharp phase transition: as shown in Appendix Figure 17, its reconstruction loss stayed nearly flat for thousands of steps and then dropped abruptly. We interpret this asymmetry as evidence that the shared latent space is dominated by AIA-aligned directions, with HMI information concentrated in a smaller, nearly orthogonal subspace. The random HMI decoder is initially misaligned with this subspace and produces almost no useful gradient signal; training slowly rotates it into alignment, at which point reconstruction succeeds rapidly. This staged alignment-driven dynamic, where weaker directions in a structured representation are learned only after a long plateau, matches the analytical

Table 1: Per-channel and mean reconstruction error across AIA wavelengths (lower is better).

Experiment	aia131	aia1600	aia171	aia193	aia211	aia304	aia335	aia94	mean_aia
SDO-FM	62.071	312.539	31206.247	141385.411	11499.586	1088.171	44.245	4.499	23200.346
Surya	70.843	997.662	25111.951	53256.142	11014.614	1779.434	64.925	4.668	11537.530
Ours	**12.677**	**10.985**	**1119.857**	**1019.123**	**518.076**	**37.397**	**1.157**	**0.444**	**339.964**

Table 2: AR segmentation.

Model	IoU (%) ↑	Dice (%) ↑
U-Net	$26.10_{\pm 0.43}$	$41.40_{\pm 0.54}$
SegFormer	$41.98_{\pm 0.68}$	$59.13_{\pm 0.67}$
Surya (LoRA)	$42.65_{\pm 0.85}$	$59.79_{\pm 0.83}$
Ours (frozen)	$\mathbf{44.01_{\pm 0.73}}$	$\mathbf{61.12_{\pm 0.70}}$

Table 3: Solar flare forecasting (≥M1.0 within 24 h).

Model	TSS ↑	HSS ↑	F1 ↑	Recall ↑
AlexNet	$0.258_{\pm 0.155}$	$0.342_{\pm 0.170}$	$0.381_{\pm 0.172}$	$0.270_{\pm 0.159}$
ResNet-50	$0.267_{\pm 0.098}$	$0.335_{\pm 0.094}$	$0.389_{\pm 0.089}$	$0.294_{\pm 0.102}$
Surya (LoRA)	$0.397_{\pm 0.164}$	$0.466_{\pm 0.161}$	$0.514_{\pm 0.161}$	$0.424_{\pm 0.168}$
Ours (frozen)	$\mathbf{0.557_{\pm 0.056}}$	$\mathbf{0.511_{\pm 0.038}}$	$\mathbf{0.580_{\pm 0.032}}$	$\mathbf{0.644_{\pm 0.078}}$

Table 4: Solar wind speed.

Model	MAE ↓	RMSE ↓
ResNet-18	$0.220_{\pm 0.108}$	$0.253_{\pm 0.112}$
ResNet-50	$0.368_{\pm 0.019}$	$0.431_{\pm 0.058}$
Surya (LoRA)	$0.127_{\pm 0.007}$	$0.161_{\pm 0.008}$
Ours (frozen)	$\mathbf{0.125_{\pm 0.011}}$	$\mathbf{0.154_{\pm 0.018}}$

Table 5: EVE spectra prediction, one hour ahead.

Model	RMSE (10^{-2}) ↓	MAPE (%) ↓	MAE (10^{-2}) ↓
ResNet-50	$4.120_{\pm 2.820}$	$3.290_{\pm 1.460}$	$1.824_{\pm 0.923}$
AlexNet	$1.340_{\pm 0.176}$	$1.727_{\pm 0.250}$	$0.764_{\pm 0.180}$
Surya (LoRA)	$1.333_{\pm 0.140}$	$1.640_{\pm 0.130}$	$0.722_{\pm 0.044}$
Ours (frozen)	$\mathbf{1.196_{\pm 0.061}}$	$\mathbf{1.429_{\pm 0.127}}$	$\mathbf{0.578_{\pm 0.064}}$

Tables 2–5: single AIA 94 Å input, 10% of the SuryaBench training labels, mean ± std over 3 seeds. Ours: frozen encoder; Surya: LoRA; all other baselines trained end-to-end. Solar wind errors are on normalized speed.

picture of learning in deep linear networks given, for example, by [48]. Crucially, the fact that HMI is recovered *at all* shows that the encoder retains HMI information in the latent space rather than discarding it; the prolonged plateau shows that this information is stored in directions distinct from those carrying AIA information. Together, these observations indicate that the encoder organizes modality-specific physics into approximately orthogonal subspaces within the shared latent space.

5 Conclusion

We introduced the first native-resolution vision foundation model for heliophysics, scaling the context window beyond 3 million tokens to process 13 independent observations without spatial downsampling. By combining two-way model parallelism with a Random Fourier Projection (RFP) method, we enable the model to *efficiently* learn physically grounded cross-modal representations, and we validate this claim through reconstruction and downstream application experiments. Our frozen encoder extracts highly transferable features, achieving state-of-the-art results on all the four downstream task: dense prediction (AR segmentation downstream task), classification (Solar flare prediction, task), regression (solar wind prediction), and stringent spectral regression (EVE irradiance prediction downstream task).

Limitations and Future Works. While our RFP pretraining objective successfully captures complex multiscale spatial topologies, it focuses exclusively on instantaneous spatial representations and does not explicitly model temporal dynamics. Consequently, the current architecture cannot natively extrapolate solar evolution over time. Building upon the robust representations established here, future work will leverage this well-structured frozen spatial latent space as a foundation for temporal modeling. We aim to utilize advanced generative frameworks, such as continuous-time flow matching or autoregressive Transformers, operating directly within this latent space to accurately forecast the dynamic evolution of solar phenomena.

Acknowledgments This work is supported by NASA Grant 80MSFC25M0084. The Authors acknowledge the NSF NAIRR Pilot and NVIDIA for providing support under grant no. NAIRR240178. The authors would also like to thank NASA Advanced Supercomputing (NAS) Division for their compute support. We also acknowledge computational support from the NAIRR Deep Partnership with NVIDIA, and the Swiss National Supercomputing Centre (CSCS). The authors are grateful to the scientific and engineering team that developed Surya 1.0 for valuable insight and discussions.

References

[1] Carolus J. Schrijver. Socio-economic hazards and impacts of space weather: The important range between mild and extreme. *Space Weather*, 13(9):524–528, 2015. ISSN 1542-7390. doi: 10.1002/2015sw001252. URL http://dx.doi.org/10.1002/2015SW001252.

[2] Sujit Roy, Johannes Schmude, Rohit Lal, Vishal Gaur, Marcus Freitag, Julian Kuehnert, Theodore van Kessel, Dinesha V. Hegde, Andrés Muñoz-Jaramillo, Johannes Jakubik, Etienne Vos, Kshitiz Mandal, Ata Akbari Asanjan, Joao Lucas de Sousa Almeida, Amy Lin, Talwinder Singh, Kang Yang, Chetraj Pandey, Jinsu Hong, Berkay Aydin, Thorsten Kurth, Ryan McGranaghan, Spiridon Kasapis, Vishal Upendran, Shah Bahauddin, Daniel da Silva, Nikolai V. Pogorelov, Anne Spalding, Campbell Watson, Manil Maskey, Madhulika Guhathakurta, Juan Bernabe-Moreno, and Rahul Ramachandran. Surya: Foundation Model for Heliophysics. In *arXiv:2508.14112*, 2025.

[3] W Dean Pesnell, B J Thompson, and PC Chamberlin. The solar dynamics observatory (sdo). In *The solar dynamics observatory*, pages 3–15. Springer, 2012.

[4] Yukun Luo, Jie Jiang, and Ruihui Wang. The Sun's Magnetic Power Spectra over Two Solar Cycles. I. Calibration between SDO/HMI and SOHO/MDI Magnetograms. In *The Astrophysical Journal, vol. 954 (2)*, 2023.

[5] Eleanna Asvestari, Manuela Temmer, Ronald M. Caplan, Jon A. Linker, Stephan G. Heinemann, Rui F. Pinto, Carl J. Henney, Charles N. Arge, Mathew J. Owens, Maria S. Madjarska, Jens Pomoell, Stefan J. Hofmeister, Camilla Scolini, and Evangelia Samara. Coronal Models and Detection of the Open Magnetic Field. In *The Astrophysical Journal, 971:45 (24pp)*, 2024.

[6] Harris Abdul Majid, Pietro Sittoni, and Francesco Tudisco. Solaris: A Foundation Model of the Sun. In *arXiv:2411.16339*, 2025.

[7] James Walsh, Daniel G. Gass, Raul Ramos Pollan, Paul J. Wright, Richard Galvez, Noah Kasmanoff, Jason Naradowsky, Anne Spalding, James Parr, and Atılım Güneş Baydin. A Foundation Model for the Solar Dynamics Observatory. In *arXiv:2410.02530*, 2024.

[8] Feng Wang, Yaodong Yu, Guoyizhe Wei, Wei Shao, Yuyin Zhou, Alan Yuille, and Cihang Xie. Scaling Laws in Patchification: An Image Is Worth 50,176 Tokens And More. In *Proc. ICML*, 2025.

[9] Duy-Kien Nguyen, Mahmoud Assran, Unnat Jain, Martin R. Oswald, Cees G. M. Snoek, and Xinlei Chen. An Image is Worth More Than 16x16 Patches: Exploring Transformers on Individual Pixels. In *Proc. ICLR*, 2025.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proc. ICLR*, 2021.

[11] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. MultiMAE: Multi-modal Multi-task Masked Autoencoders. In *Proc. ECCV*, 2022.

[12] Gelu Nita, Manolis Georgoulis, Irina Kitiashvili, Viacheslav Sadykov, Enrico Camporeale, Alexander Kosovichev, Haimin Wang, Vincent Oria, Jason Wang, Rafal Angryk, Berkay Aydin, Azim Ahmadzadeh, Xiaoli Bai, Timothy Bastian, Soukaina Filali Boubrahimi, Bin Chen, Alisdair Davey, Sheldon Fereira, Gregory Fleishman, Dale Gary, Andrew Gerrard, Gregory Hellbourg, Katherine Herbert, Jack Ireland, Egor Illarionov, Natsuha Kuroda, Qin Li, Chang Liu, Yuexin Liu, Hyomin Kim, Dustin Kempton, Ruizhe Ma, Petrus Martens, Ryan McGranaghan, Edward Semones, John Stefan, Andrey Stejko, Yaireska Collado-Vega, Meiqi Wang, Yan Xu, and Sijie Yu. Machine learning in heliophysics and space weather forecasting: A white paper of findings and recommendations, 2020. URL https://arxiv.org/abs/2006.12224.

[13] Bo Shen, Marco Marena, Chenyang Li, Qin Li, Haodi Jiang, Mengnan Du, Jiajun Xu, and Haimin Wang. Deep Computer Vision for Solar Physics Big Data: Opportunities and Challenges. In *IEEE International Conference on Big Data (BigData), DOI: 10.1109/BigData62323.2024.10825648*, 2024.

[14] S. Riggi, P. Romano, A. Pilzer, and U. Becciani. Solar flare forecasting with foundational transformer models across image, video, and time-series modalities. In *Astronomy and Computing, vol. 55, 101042, Elsevier ScienceDirect*, 2026.

[15] John A. Armstrong and Lyndsay Fletcher. Fast solar image classification using deep learning and its importance for automation in solar physics. *Solar Physics*, 294(6), 2019. ISSN 1573-093X. doi: 10.1007/s11207-019-1473-z. URL http://dx.doi.org/10.1007/s11207-019-1473-z.

[16] Ahmet Kucuk, Berkay Aydin, and Rafal Angryk. Multi-wavelength solar event detection using faster r-cnn. In *2017 IEEE International Conference on Big Data (Big Data)*, page 2552–2558. IEEE, December 2017. doi: 10.1109/bigdata.2017.8258214. URL http://dx.doi.org/10.1109/BigData.2017.8258214.

[17] Ji-Hye Baek, Sujin Kim, Seonghwan Choi, Jongyeob Park, Jihun Kim, Wonkeun Jo, and Dongil Kim. Solar event detection using deep-learning-based object detection methods. *Solar Physics*, 296(11), November 2021. ISSN 1573-093X. doi: 10.1007/s11207-021-01902-5. URL http://dx.doi.org/10.1007/s11207-021-01902-5.

[18] Xulong Guo, Yunfei Yang, Song Feng, Xianyong Bai, Bo Liang, and Wei Dai. Solar-filament detection and classification based on deep learning. *Solar Physics*, 297(8), August 2022. ISSN 1573-093X. doi: 10.1007/s11207-022-02019-z. URL http://dx.doi.org/10.1007/s11207-022-02019-z.

[19] Long Gong, Yunfei Yang, Song Feng, Wei Dai, Bo Liang, and Jianping Xiong. Solar active regions detection and tracking based on deep learning. *Solar Physics*, 299(8), August 2024. ISSN 1573-093X. doi: 10.1007/s11207-024-02362-3. URL http://dx.doi.org/10.1007/s11207-024-02362-3.

[20] Sumiaya Rahman, Yong-Jae Moon, Eunsu Park, Ashraf Siddique, Il-Hyun Cho, and Daye Lim. Super-resolution of sdo/hmi magnetograms using novel deep learning methods. *The Astrophysical Journal Letters*, 897(2):L32, 2020. ISSN 2041-8213. doi: 10.3847/2041-8213/ab9d79. URL http://dx.doi.org/10.3847/2041-8213/ab9d79.

[21] Long Xu, Wen-Qing Sun, Yi-Hua Yan, and Wei-Qiang Zhang. Solar image deconvolution by generative adversarial network. *Research in Astronomy and Astrophysics*, 20(11):170, November 2020. ISSN 2397-6209. doi: 10.1088/1674-4527/20/11/170. URL http://dx.doi.org/10.1088/1674-4527/20/11/170.

[22] Ankan Dash, Junyi Ye, Guiling Wang, and Huiran Jin. High resolution solar image generation using generative adversarial networks. *Annals of Data Science*, 11(5):1545–1561, August 2022. ISSN 2198-5812. doi: 10.1007/s40745-022-00436-2. URL http://dx.doi.org/10.1007/s40745-022-00436-2.

[23] R. Jarolim, A. M. Veronig, W. Pötzi, and T. Podladchikova. A deep learning framework for instrument-to-instrument translation of solar observation data. *Nature Communications*, 16 (1), April 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-58391-4. URL http://dx.doi.org/10.1038/s41467-025-58391-4.

[24] Vishal Upendran, Mark C. M. Cheung, Shravan Hanasoge, and Ganapathy Krishnamurthi. Solar wind prediction using deep learning. *Space Weather*, 18(9), 2020. ISSN 1542-7390. doi: 10.1029/2020sw002478. URL http://dx.doi.org/10.1029/2020SW002478.

[25] Chetraj Pandey, Temitope Adeyeha, Jinsu Hong, Rafal A. Angryk, and Berkay Aydin. *Advancing Solar Flare Prediction Using Deep Learning with Active Region Patches*, page 50–65. Springer Nature Switzerland, 2024. ISBN 9783031703812. doi: 10.1007/978-3-031-70381-2_4. URL http://dx.doi.org/10.1007/978-3-031-70381-2_4.

[26] Zhenduo Wang, Yasser Abduallah, Jason T. L. Wang, Haimin Wang, Yan Xu, Vasyl Yurchyshyn, Vincent Oria, Khalid A. Alobaid, and Xiaoli Bai. Daily predictions of f10.7 and f30 solar indices with deep learning. *Journal of Geophysical Research: Space Physics*, 131(2), February 2026. ISSN 2169-9402. doi: 10.1029/2025ja034868. URL http://dx.doi.org/10.1029/2025JA034868.

[27] Yian Yu, Yang Chen, Lulu Zhao, Kathryn Whitman, Ward Manchester, and Tamas Gombosi. Solar energetic particle forecasting with multi-task deep learning: Sepnet, 2025. URL `https://arxiv.org/abs/2512.12786`.

[28] Richard Galvez, David F. Fouhey, Meng Jin, Alexandre Szenicer, Andrés Muñoz-Jaramillo, Mark C. M. Cheung, Paul J. Wright, Monica G. Bobra, Yang Liu, James Mason, and Rajat Thomas. A Machine Learning Dataset Prepared From the NASA Solar Dynamics Observatory Mission. In *The Astrophysical Journal Supplement Series*, 2019.

[29] Valentina Salvatelli, Souvik Bose, Brad Neuberg, Luiz F. G. dos Santos, Mark Cheung, Miho Janvier, Atilim Gunes Baydin, Yarin Gal, and Meng Jin. Using U-Nets to Create High-Fidelity Virtual Observations of the Solar Corona. In *NeurIPS Workshop ML4PS*, 2019.

[30] Marius Giger and André Csillaghy. Unsupervised Anomaly Detection With Variational Autoencoders Applied to Full-Disk Solar Images. In *Space Weather, 22, e2023SW003516*, 2023.

[31] Sujit Roy, Dinesha V. Hegde, Johannes Schmude, Amy Lin, Vishal Gaur, Rohit Lal, Kshitiz Mandal, Talwinder Singh, Andrés Muñoz-Jaramillo, Kang Yang, Chetraj Pandey, Jinsu Hong, Berkay Aydin, Ryan McGranaghan, Spiridon Kasapis, Vishal Upendran, Shah Bahauddin, Daniel da Silva, Marcus Freitag, Iksha Gurung, Nikolai Pogorelov, Campbell Watson, Manil Maskey, Juan Bernabe-Moreno, and Rahul Ramachandran. SuryaBench: Benchmark Dataset for Advancing Machine Learning in Heliophysics and Space Weather Prediction. In *arXiv:2508.14107*, 2025.

[32] Neelu Madan, Nicolae-Catalin Ristea, Kamal Nasrollahi, Thomas B. Moeslund, and Radu Tudor Ionescu. CL-MAE: Curriculum-Learned Masked Autoencoders. In *Proc. WACV*, 2024.

[33] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *Proc. CVPR*, 2022.

[34] Badri Narayana Patro, Vinay P. Namboodiri, and Vijay Srinivas Agneeswaran. SpectFormer: Frequency and Attention is what you need in a Vision Transformer. In *Proc. WAVC*, 2025.

[35] Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains. In *Proc. NeurIPS*, 2020.

[36] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism. *arXiv e-prints*, art. arXiv:1909.08053, September 2019. doi: 10.48550/arXiv.1909.08053.

[37] Vijay Korthikanti, Jared Casper, Sangkug Lym, Lawrence McAfee, Michael Andersch, Mohammad Shoeybi, and Bryan Catanzaro. Reducing Activation Recomputation in Large Transformer Models. *arXiv e-prints*, art. arXiv:2205.05198, May 2022. doi: 10.48550/arXiv.2205.05198.

[38] NVIDIA Corporation. Transformerengine. `https://github.com/NVIDIA/TransformerEngine`, 2022. Accessed: 2026-01-09.

[39] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. *arXiv e-prints*, art. arXiv:2205.14135, May 2022. doi: 10.48550/arXiv.2205.14135.

[40] Sujit Roy, Dinesha V. Hegde, Johannes Schmude, Amy Lin, Vishal Gaur, Rohit Lal, Kshitiz Mandal, Talwinder Singh, Andrés Muñoz-Jaramillo, Kang Yang, Chetraj Pandey, Jinsu Hong, Berkay Aydin, Ryan McGranaghan, Spiridon Kasapis, Vishal Upendran, Shah Bahauddin, Daniel da Silva, Marcus Freitag, Iksha Gurung, Nikolai Pogorelov, Campbell Watson, Manil Maskey, Juan Bernabe-Moreno, and Rahul Ramachandran. SuryaBench: Benchmark dataset for advancing machine learning in heliophysics and space weather prediction. *Scientific Data*, 2026. doi: 10.1038/s41597-026-06552-5.

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

[42] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, volume 34, pages 12077–12090, 2021.

[43] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[44] 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*, pages 770–778, 2016.

[45] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[46] T. N. Woods, et al. Extreme Ultraviolet Variability Experiment (EVE) on the Solar Dynamics Observatory (SDO): Overview of Science Objectives, Instrument Design, Data Products, and Model Developments. In *Sol. Phys. 275, 115–143, 10.1007/s11207-009-9487-6*, 2012.

[47] Andrew Jaegle, Felix Gimeno, Andy Brock, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver: General perception with iterative attention. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 4651–4664. PMLR, 2021.

[48] Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

[49] N. Nishizuka, K. Sugiura, Y. Kubo, M. Den, and M. Ishii. Deep Flare Net (DeFN) Model for Solar Flare Prediction. In *The Astrophysical Journal, vol. 858 (2)*, 2018.

[50] Akash Vinod Shirke, Sakshi Sheshrao Charde, and Balendra Pratap Singh. Analysis of Solar Flare and Sunspots on 4th Jan 2025 and Their Effects on Space Weather. In *arXiv:2601.05591*, 2026.

[51] Jingjing Wang, Pengyu Liang, Tingyu Wang, Ming Li, Yanmei Cui, Siwei Liu, Xin Huang, Xiang Li, Minghui Zhang, Yunshi Zeng, Zhu Cao, Jiekang Feng, Qinghua Hu, Bingxian Luo, and Bing Cao. Large Model Driven Solar Activity AI Forecaster: A Scalable Dual Data-Model Framework. In *arXiv:2508.06892*, 2025.

[52] Griffin T. Goodwin, Christoph Schirninger, Robert Jarolim, Angelos Vourlidas, and Lorien Pratt. FOXES: A Framework For Operational X-ray Emission Synthesis. In *NeurIPS Workshop ML4PS, arXiv:2510.22801*, 2025.

[53] Sean McGregor, Dattaraj Dhuri, Anamaria Berea, and Andrés Muñoz-Jaramillo. FlareNet: A Deep Learning Framework for Solar Phenomena Prediction. In *NeurIPS Workshop on Deep Learning for Physical Sciences*, 2017.

[54] Chetraj Pandey, Anli Ji, Rafal A. Angryk, Manolis K. Georgoulis, and Berkay Aydin. Towards coupling full-disk and active region-based flare prediction for operational space weather forecasting. In *Front. Astron. Space Sci., Space Physics, vol. 9*, 2022.

[55] Peter Butka, Martin Sarnovský, Viera Krešňáková, Šimon Mackovjak, and Lenka Kališková. Solar corona structures detection and segmentation using deep neural networks: A survey. *IEEE Access*, 2026. DOI: 10.1109/ACCESS.2026.3653555.

[56] Jonas Tirona, Sarang Patil, Spiridon Kasapis, Eren Dogan, John Stefan, Irina N. Kitiashvili, Alexander G. Kosovichev, and Mengjia Xu. Forecasting continuum intensity for solar active region emergence prediction using transformers, arXiv:2601.13144. 2026.

[57] Daniela Szwarcman, Sujit Roy, Paolo Fraccaro, Þorsteinn 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, Myscon Truong, Denys Godwin, Hyunho Lee, Chia-Yu Hsu, Ata Akbari Asanjan, Besart Mujeci, Disha Shidham, Trevor Keenan, Paulo Arevalo, Wenwen Li, Hamed Alemohammad, Pontus Olofsson, 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 Multi-Temporal Foundation Model for Earth Observation Applications. In *arXiv:2412.02732*, 2025.

[58] David Mizrahi, Roman Bachmann, Oguzhan Fatih Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4M: Massively Multimodal Masked Modeling. In *Proc. NeurIPS*, 2023.

[59] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive Image Generation without Vector Quantization. In *Proc. NeurIPS*, 2024.

[60] Roman Bachmann, Oguzhan Fatih Kar, David Mizrahi, Ali Garjani, Mingfei Gao, David Griffiths, Jiaming Hu, Afshin Dehghan, and Amir Zamir. M-21: An Any-to-Any Vision Model for Tens of Tasks and Modalities. In *Proc. NeurIPS*, 2024.

[61] 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 *Proc. NeurIPS*, 2022.

[62] 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 *Proc. ICCV*, 2025.

[63] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. Maskgit: Masked Generative Image Transformer. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11315-11325*, 2022.

[64] Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion Transformers with Representation Autoencoders. In *arXiv:2510.11690*, 2025.

[65] Xingzhuo Guo, Yu Zhang, Baixu Chen, Haoran Xu, Jianmin Wang, and Mingsheng Long. Dynamical Diffusion: Learning Temporal Dynamics with Diffusion Models. In *in Proceedings ICLR*, 2025.

[66] David Ruhe, Jonathan Heek, Tim Salimans, and Emiel Hoogeboom. Rolling Diffusion Models. In *Proc. ICML*, 2024.

[67] Salva Rühling Cachay, Bo Zhao, Hailey Joren, and Rose Yu. DYffusion: A Dynamics-informed Diffusion Model for Spatiotemporal Forecasting. In *Proc. NeurIPS*, 2023.

[68] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

[69] Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. In *Proceedings International Conference on Learning Representations (ICLR)*, 2020.

[70] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European Conference on Computer Vision (ECCV)*, volume 12356 of *Lecture Notes in Computer Science*, pages 776–794. Springer, 2020.

Appendix: Supplementary Material for 'Beyond 3 Million Tokens: A Multi-Modal Foundation Model for Full-Resolution Heliophysics'

The Appendix of this paper is organized as follows. First, in Appendix A, we describe the SDO data we use. This is in continuation of the description of the data in Section 3.1 in the main paper. Next, in Appendix B, we provide further information about the proposed method (i.e. following-up the analysis and presentation of our method in Section 3.2). In Appendix C, more information about the model parallelism, i.e. both feature parallelism and sequence parallelism, is provided (i.e. in continuation of the analysis and discussion in Section 3.3 in the main paper).

Furthermore, the impact of masking rate and the effect of missing modalities is examined in Appendix D and Appendix E, respectively. This is in continuation of the evaluation and results presented in Section 4 in the main paper. Moreover, in Appendix F, we examine and discuss the patch scaling ablation results for our model.

Downstream task of Active Region (AR) segmentation. The further evaluation results and analysis for our multi-modal heliophysics foundation model on the downstream task of Active Region (AR) segmentation is presented in Appendix G. This is following-up the evaluation and analysis of our main results in Section 4 in the main paper.

Downstream task of EVE prediction. In addition, the further results for the downstream task of EVE prediction are presented and examined in Appendix I.

Furthermore, the further results for the comparison of MAE versus Fourier-MAE on SDO AIA and HMI data are presented and discussed in Appendix J. Next, Appendix K examines the results for the visible token ratio ρ sensitivity analysis. In Appendix L, we present the detailed results for the masking rates and channel masking. This is in continuation of Appendix D (and E), as well as of the main results presented in Section 4 in the main paper. Finally, Appendix M presents and discusses the results of the experiment when the decoder is trained and the multi-modal encoders are frozen. Importantly, in this part of the Appendix, we also examine and analyze the phase-transition-like HMI loss drop during decoder-only finetuning in Figure 17.

A Preparation of SDO data

We follow the SDO preprocessing and normalization protocol used in Surya [2, 31], using eight AIA and five HMI channels at 4096×4096 resolution and a 12-minute cadence. We modify the temporal coverage and partitioning to include observations from 2010–2011, 2013–2021, and 2023–2024, including Solar Cycle 25. For each included year from 2011 onward, days 16–31 are assigned to validation and days 61 onward to training, leaving days 1–15 and 32–60 as temporal buffers. The 2010 training data begin on June 1, with no validation observations from that year. Neither partition includes 2012 or 2022.

B Further information about the proposed method

The main method of our proposed model is presented in Section 3. In addition, in Table 6, we compare our multi-modal heliophysics foundation model to other existing heliophysics foundation models (SDO-FM, Solaris, Surya). Strong solar flares events (e.g., X-class) can disrupt avionics and satellite-based navigation, affecting commercial aircrafts (i.e. see also Section 1). This motivates the need for accurate and early forecasting of solar activity [31, 2, 25, 49, 50]. The importance of this paper's research was underscored also by the November 2025 Airbus A320 incident in which Airbus issued a global precautionary fleet action after an aircraft experienced uncommanded maneuvers due to intense solar radiation from an X-class solar flare. This event in 2025 demonstrated that current space weather warning systems lack the precision to protect modern fly-by-wire avionics. In addition, as we also discussed in Sections 1 and A, solar flares, CMEs, and geomagnetic storms can damage power grids on Earth [31, 51–56], causing several-hour power outages and blackouts (for example, the Hydro-Québec event in 1989).

Furthermore, different from methods that use masking (e.g., [57], [58], and [59], as well as [60], [61], [62], and [63]), we utilize RFP in our model to improve performance. In addition, our multi-modal heliophysics foundation model learns generalizable representations and subsequent stages can focus

Table 6: Comparison of heliophysics foundation models. †Surya treats all 13 channels as a single combined input tensor, not as independent modalities with per-channel adapters. ‡SDO-FM uses the SDOML dataset downsampled to 512×512. ⋆Solaris uses only AIA channels at downsampled resolution. Context tokens are computed as $(\text{channels}) \times (H/P)^2$.

Property	SDO-FM	Solaris	Surya	Ours
Architecture	NVAE / MAE	3D Swin Transf.	Spectral + LS Attn.	ViT + RFP
Parameters	—	109M	366M	193M
Resolution	512×512‡	512×512⋆	4096×4096	4096×4096
Instruments	AIA (+ EVE)	AIA only⋆	AIA + HMI†	AIA + HMI
Channels used	9 AIA	7 AIA	13 (combined)†	13 (independent)
Modality treatment	Combined	Combined	Combined†	Per-modality adapters
Context tokens	~9k	~7k	~65k	~852k scaled to ~3.4M
Pretraining task	MAE reconstruction	12h forecasting	1h forecasting + rollout	RFP reconstruction
Temporal modeling	Single-frame	Temporal (12h)	Temporal (AR rollout)	Single-frame
Training data span	2010–2023	13 years	9 years	~15 years
HMI vector fields	✗	✗	✓	✓
Missing modality robustness	Not evaluated	Not evaluated	Not evaluated	✓ (Figure 3)

exclusively on learning dynamical evolution in an already structured latent space using our model's encoder frozen [64–67]. This separation ensures that spatial representation learning is decoupled from the temporal modeling to produce latent token sequences for accurate forecasting.

C Further information about model parallelism

We set up the feature and sequence parallel dimensions (labeled `f`, `y`, and `x`) using orthogonal NCCL communicators from the PyTorch `DeviceMesh` constructor. We create a common communicator for the joined `y-x` direction because it simplifies and speeds up communications related to sequence parallelism (the sequence is flattened into a 1D string of tokens). Since regular attention is permutation invariant, the strides of the splitting do not matter; thus, different model parallel grid decompositions are equivalent. Additionally, our code also supports traditional widely-used data-level parallelism.

TE does not reduce weight gradients across the respective dimensions. For example, all gradients for all MLP weights and biases have to be reduced across the sequence parallel dimension since all these weights are shared across this dimension. Those reductions are enabled using the post accumulation gradient hook logic in PyTorch. As soon as a gradient is fully computed on a given tensor, a registered post accumulation gradient hook is fired to compute averages or sums over model parallel dimensions. This setup enables modularity on layer level because each module is aware of which reductions need to happen in order to compute a full forward and backward pass. Furthermore, those hooks are completely overlapped with other computations in the backward pass. In fact, our profiles do not show exposed communications related to weight gradient reductions.

D Further discussion: The impact of masking rate

Effect of Masking Rate. To evaluate the model's generative capacity and structural accuracy, we present in Figure 5 an analysis involving a spectrum of masking rates (from 25% to 90%). We quantify reconstruction performance under varying levels of sparsity using MSE and SSIM. For each masking rate, we calculate a per-channel analysis covering AIA and HMI modalities, and a global average to demonstrate the model's consistency. The model exhibits stability at 25% masking rate with average MSE 0.195 and average SSIM 0.972, which suggests that moderate masking is an effective regularizer. Even at the severe 90%, the model maintains high-fidelity reconstructions.

In continuation of the discussion above about Figure 5 and the effect of the masking rate on the image reconstructions, AIA channels (and especially 171 and 193 Å bands) exhibit low error rates, while HMI vector components (B_x, B_y, B_z) are difficult to reconstruct. This disparity is expected given the abundant cross-spectral correlations inherent in AIA's multi-thermal EUV channels. In contrast, the HMI transverse magnetic field components are orthogonal, and the velocity component (`hmi_v`) is a distinct physical modality. This effect is further evidenced by the line-of-sight magnetic component

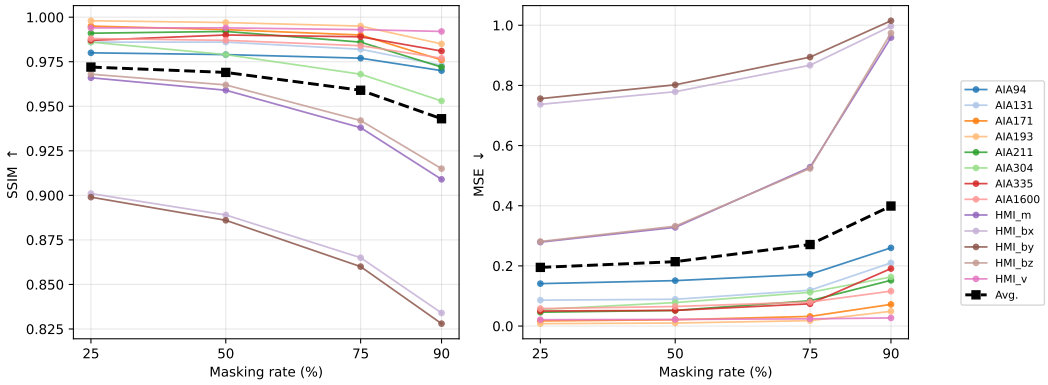

Figure 5: Reconstruction evaluation (SSIM, MSE) of our model across various masking rates.

(HMI_m), which, despite being a magnetic modality, benefits from stronger structural correlations with its vector counterparts.

E Further discussion on the impact of missing modalities

In continuation of the discussion in Section 4 about Figure 3 and the effect of missing modalities on the task of reconstruction, the removal of AIA 211Å results in the lowest mean MSE (0.2696), suggesting that its features are highly correlated with other EUV channels, that its absence does not hinder the global reconstruction task.

The difficulty in reconstructing HMI vector fields (Bx, By, and Bz), compared to AIA channels, stems from the fact that vector magnetic field components are largely orthogonal, and the velocity field represents a distinct physical modality with fewer spectral proxies. We also see a clear structural dependency between the line-of-sight and transverse magnetic field components (HMI_m and HMI_Bx). This relationship is physically expected, as these components are highly correlated, particularly in central disk locations where the geometry simplifies the relationship between the vector components.

F Further results: Patch scaling ablation

Patch-size scaling ablation. Under identical training conditions (same backbone, same RFP, same data), a 16×16 patch model achieves 14.3% lower validation loss than a 32×32 model after 5 epochs. This is consistent with the patchification scaling law $\mathcal{L} \propto N_{\text{tok}}^{-\alpha}$ in [8], where $N_{\text{tok}} = (H/P)^2$ and $\alpha > 0$. Moving from 32×32 to 16×16, quadruples the token count; the observed improvement after only 5 epochs indicates the model is on the favorable side of this curve. Our full-scale 8×8 model quadruples the count again (>3M tokens across 13 modalities), and the downstream task superiority over Surya's 16×16 model provides evidence that the additional compute invested in finer patches yields significant measurable representational gains.

In Figure 6, we show patch scaling ablation results for patch sizes 8 and 16.

G Active Region (AR) segmentation downstream task and Comparison with Prior Heliophysics Models

There are several differences in architecture, resolution, training data, and pretraining objective across the different heliophysics foundation models: SDO-FM operates at 512×512, Solaris uses only AIA channels, and Surya uses an autoregressive forecasting objective rather than masked reconstruction. On the downstream task of AR segmentation in Table 2 in the main text, our frozen encoder outperforms thr LoRA-Surya by +1.3 Dice and +1.4 IoU points under the shared 10%-label, 3-seed exepriments. Since Surya is the only prior model operating at native 4096×4096 with both AIA and HMI, it is the most fair comparison. In addition, in Table 2 compares against a transfromer

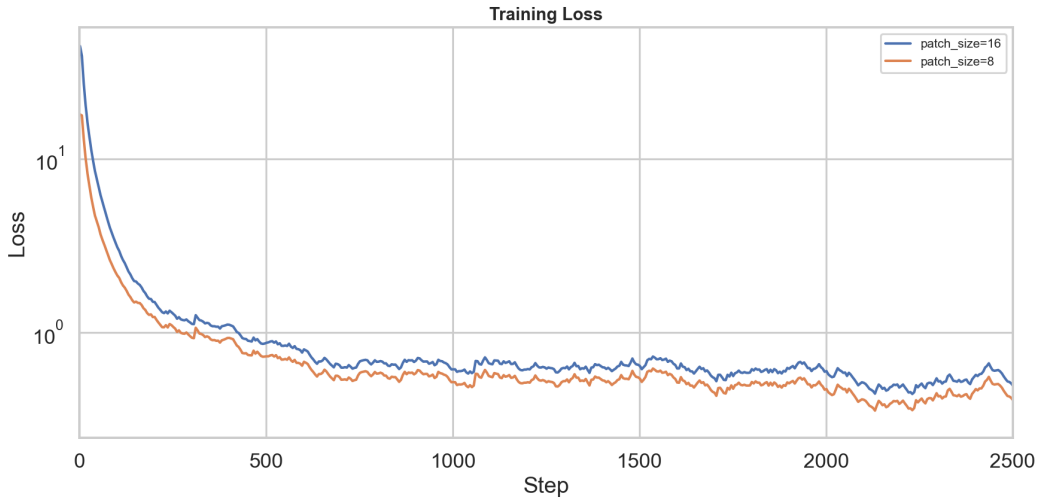

Figure 6: Training reconstruction loss over epochs for patch sizes 8 and 16.

segmentation model (SegFormer) and a CNN (U-Net), both trained end-to-end. We retrain U-Net deliberately, since CNN segmentation is widely used in heliophysics, so it gives the researchers a familiar refrence point.

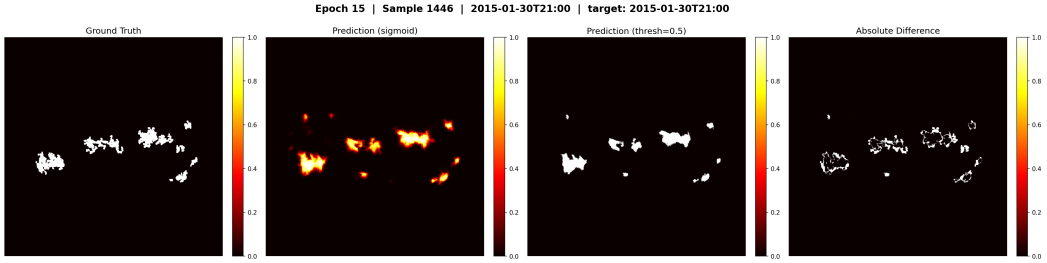

Figure 7: Reconstruction results by our heliophysics foundation model, at epoch 15.

In Figure 7, we examine the results of our model on the downstream task of AR segmentation. The qualitative results show that the proposed multi-modal heliophysics foundation model is effective and can detect and segment active regions of the Sun. In fact, for this result, we see that the prediction is actually better than the ground truth as the ground truth is patchy and not continuous.

Why RFP pretraining produces segmentation-ready features. The connection is grounded in the information-theoretic structure of RFP. Given a patch $\mathbf{z}_i$ and its RFP embedding $\phi(\mathbf{z}_i)$ (Eq. (5)), inner products in the projected space approximate a shift-invariant Gaussian kernel: $\langle \phi(\mathbf{z}_i), \phi(\mathbf{z}_j) \rangle \approx k(\mathbf{z}_i - \mathbf{z}_j)$ in expectation [68]. This means RFP tokens encode *local contrast and edge structure*, patches with dissimilar content map to distant embeddings while destroying pixel-level detail needed for trivial copying. To reconstruct from a mixture of 25% visible tokens and 75% such kernel-distance-preserving projections, the encoder must learn representations capturing: (a) high-frequency spatial boundaries (to delineate ARs from quiet Sun), (b) cross-modal correlations (to recover HMI magnetic structure from correlated AIA thermal emission), and (c) multi-scale morphological context. These are precisely the features AR segmentation requires.

H Representation-quality probes

Beyond task-specific evaluation, we perform two model-agnostic representation probes on the frozen encoder to characterize the *geometry* of the learned feature space.

(i) Effective rank. We compute the effective rank $r_{\text{eff}} = \exp\big(-\sum_i \bar{\sigma}_i \log \bar{\sigma}_i\big)$ (where $\bar{\sigma}_i = \sigma_i / \sum_j \sigma_j$ are the normalised singular values) of 50 K encoder tokens per modality. Table 7 reports the

results. The encoder uses ~120–160 of its 1024 dimensions meaningfully (12–15 %), indicating a well-structured, non-degenerate representation that is neither collapsed nor diffuse. Crucially, the per-modality variation is *physically consistent*: HMI magnetograms ($r_{\text{eff}} = 152$–158) require more effective dimensions than AIA EUV channels (121–147), reflecting the richer spatial complexity of vector magnetic-field data. The Dopplergram channel (`hmi_v`, 127) is notably lower, consistent with its smoother large-scale velocity patterns. The 1600 Å continuum (`aia1600`, 121) has the lowest rank, matching its less structured upper-photosphere emission. The global effective rank (144.5) lies *within* the per-modality range rather than exceeding it, demonstrating that different modalities are mapped into significantly overlapping subspaces - evidence of learned cross-modal structure rather than independent per-modality encoding.

Table 7: Effective rank of frozen encoder embeddings ($D = 1024$).

Modality	r_{eff}	**% D**	**Modality**	r_{eff}	**% D**
aia94	147.3	14.4	hmi_bx	153.5	15.0
aia131	138.2	13.5	hmi_by	152.9	14.9
aia171	133.5	13.0	hmi_bz	156.1	15.2
aia193	141.8	13.8	hmi_m	158.0	15.4
aia211	146.1	14.3	hmi_v	127.3	12.4
aia304	134.5	13.1			
aia336	134.0	13.1	**global**	**144.5**	**14.1**
aia1600	121.1	11.8			

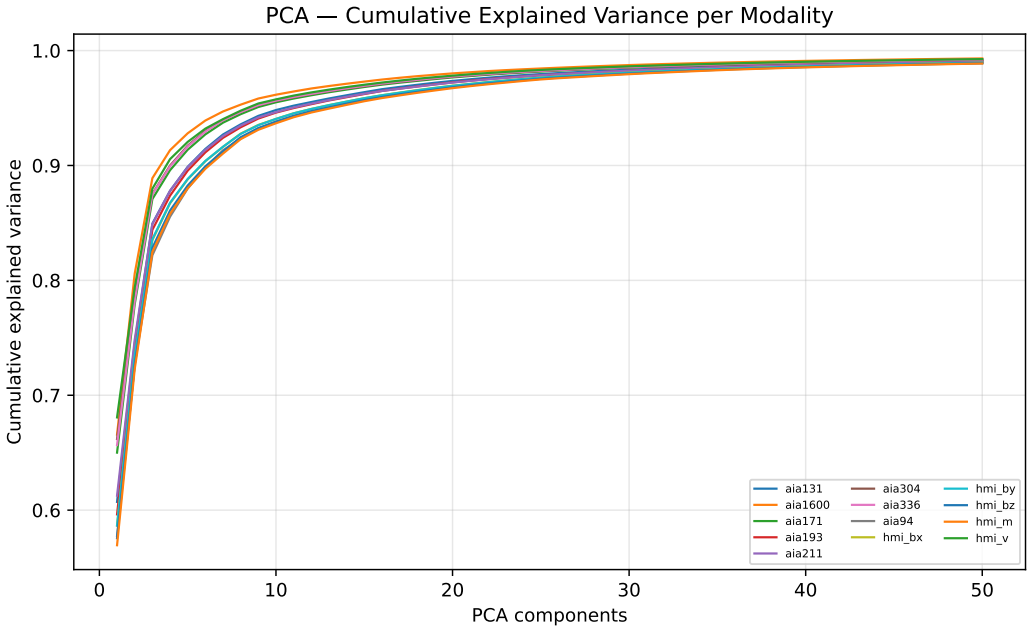

Figure 8: Explained variance as a function of the number of principal components.

(ii) PCA visualization. We probe the encoder's latent geometry by fitting a 3-component PCA on 500 K tokens pooled across all 13 modalities at zero masking, so the encoder sees every channel through full cross-modal attention. The resulting spatial maps in Figure 9 show that a single PCA basis, fitted jointly rather than per-channel, recovers physically meaningful spatial organization for both AIA and HMI. Although the encoder is shared across all modalities, it could in principle partition its embedding space into disjoint per-channel subspaces. The fact that a single 3-component PCA fitted jointly produces spatially coherent maps for channels measuring fundamentally different physics, i.e. coronal thermal emission, photospheric magnetic fields, and plasma velocity, shows that the encoder instead maps co-located physical structure into common latent directions.

The cumulative explained-variance curves in Figure 8 show a clear split: the first 10 components capture roughly 85–90 % of variance for AIA but only 75–80 % for HMI, reflecting the higher intrinsic dimensionality of vector magnetic-field data. This ordering matches the effective-rank analysis in Table 7 and provides independent confirmation that the representation allocates capacity according to the physical complexity of each modality.

Figure 10 shows the training dynamics of an earlier single-seed AR segmentation run. Our frozen encoder yields non-trivial Dice from the first evaluation, whereas U-Net produces near-zero predictions for the first ~6,000 steps, consistent with pretraining capturing structured spatial dependencies before any task supervision.

Taken together, the downstream segmentation (Table 2, as well as Figure 5, effective-rank analysis (Table 7), and PCA visualisations provide three complementary lines of evidence that the encoder learns structured, physically grounded, and task-transferable representations.

I Training Dynamics: EVE Single-Channel Probing

Figure 11 reports training loss and the three validation metrics (RMSE, MAPE, MAE) over 20 k optimizer steps for our model, ResNet-18 and AlexNet, all trained from the same single AIA 94 input. These curves come from a single-seed exploratory run that predates the multi-seed protocol of Table 5, so their absolute values are not comparable to that table; we include them for the qualitative training behaviour they reveal.

Three behaviours are worth highlighting:

Generalization gap. EUV_rn18 reaches the *lowest* train loss of all three models (3.19×10^{-4} vs. our 4.11×10^{-4}), yet its validation RMSE, MAPE, and MAE all *grow* after ~ 2.5 k steps and end up an order of magnitude worse than ours. This is a typical instance of overfitting: with the entire backbone trainable from scratch, the ResNet memorizes the training distribution and fails to generalize. Our encoder is frozen, so only the Perceiver-style head is fit, and the validation curves descend smoothly throughout training without any divergence.

Training stability. EUV_alexnet exhibits large periodic spikes in train loss (visible from ~ 5 k steps onward) and its validation metrics drift upward in the second half of training. The instability is consistent with end-to-end training of a small-capacity convolutional backbone on a high-dimensional (1343-d) regression target. Our model shows no such spikes – the head sees a fixed, well-conditioned latent space and behaves like a linear-probe-style optimization problem.

Headroom. At step 20 k our validation curves are still descending on every metric, while both baselines have either plateaued (EUV_alexnet) or are diverging (EUV_rn18). The reported numbers in Table 5 are therefore a conservative estimate of the gap between our model and the baselines.

Together, these curves support the central claim we present in Section 4: the gains over end-to-end CNN regressors come from the quality of the pretrained representation, not from extra capacity in the regression head, since the head is exposed to the same target with the same optimizer and schedule and simply has fewer parameters to fit.

Qualitative EVE Spectra Results. Figure 12 shows three representative one-hour-ahead predictions from the test split. Our model recovers both the *shape* and the *absolute level* of the spectrum, tracking the strong coronal line complex around 17–20 nm (Fe IX–XII), the Fe XV peak near 28.4 nm, and the dominant He II Lyman-α line at 30.4 nm, as well as the deep absorption-like minima beyond 32 nm. Per-sample MAE is 0.008–0.010, essentially indistinguishable from the ground truth at plotting resolution. The AlexNet baseline produces a systematically under-predicted smoothed envelope: it captures the broad shape but misses the fine line peaks and the absolute irradiance level, yielding 4–5× higher MAE on the same samples (0.044–0.046). This qualitative behaviour, i.e. the preservation of physically meaningful spectral features rather than a low-frequency fit, is what drives the quantitative gap reported in Table 5, and this is also consistent with the smooth monotonic validation curves in Figure 11.

J MAE vs. Fourier-MAE on SDO Data

Fourier-based MAE. To isolate the effect of replacing hard masking with frequency-based views, we implement a modified Masked Autoencoder (MAE), referred to as `FourierMAE`. Unlike standard MAE, which discards a large fraction of tokens to create an information bottleneck, FourierMAE replaces masked tokens with fixed Random Fourier Projection (RFP) features, preserving a structured but non-invertible view of the masked content. In Figure 15 (as well as in Figure 13), we present the RFP features on SDO data that our multi-modal heliophysics foundation model uses.

Given an input image divided into patches, we sample a random mask with ratio r. Visible patches are embedded using a standard linear projection, while masked patches are mapped to Fourier features. For a flattened patch vector z_i, the Fourier-based embedding is defined as

$$e_i^{\text{rfp}} = [\sin(2\pi z_i B), \cos(2\pi z_i B)], \tag{7}$$

where $B \sim \mathcal{N}(0, \sigma^2 I)$ is sampled once at initialization. This mapping removes exact pixel information while retaining frequency-dependent structure, preventing trivial reconstruction pathways.

The scale σ controls the frequency bandwidth of the projection. We fix $\sigma = 10$ in all experiments, yielding Fourier perturbations comparable in magnitude to the intrinsic variability of SDO observations. With this formulation, the model reconstructs masked regions by integrating visible spatial tokens with their corresponding frequency-domain views.

We compare Fourier-MAE and MAE on SDO AIA and HMI data. Both models share the same ViT backbone, patch size 16×16, 75% masking ratio, optimizer, and scheduler; the *only* difference is whether non-visible tokens are dropped (MultiMAE) or replaced with fixed Random Fourier Projections (RFT). Both are trained for 50k steps on the full 13-channel SDO dataset at native 4096×4096 resolution.

Large σ is a learnable bottleneck, not pure noise. Figure 13 quantifies how much of the original spatial structure survives the RFP map at six values of σ and exposes two distinct regimes. At $\sigma \leq 0.05$, the RFP-mean image carries the same low-frequency envelope as the original (LSD < 1, panel b) and the two radial spectra are nearly indistinguishable (panel a). For larger $\sigma \in \{0.5, 1, 5\}$, the LSD saturates at $\approx$ 2–2.4, but panel (a) shows the RFP spectrum is *not* flat: spectral mass migrates from low to high radial frequency rather than disappearing, producing a clean low/high-frequency crossover with the original. The information ordering $0 = \mathcal{I}_{\text{drop}} \ll \mathcal{I}_{\text{RFP}}(\sigma{=}10) \ll \mathcal{I}_{\text{vis}}$ is therefore strict: the encoder cannot invert each token in isolation, but the cross-token spectral structure is rich enough to drive fast convergence rather than stalling on random targets. HMI lies systematically above AIA in panel (b) at small σ, consistent with its richer high-frequency content (vector magnetic structure); the gap closes by $\sigma \geq 1$ where both modalities are in the saturated regime, which is precisely where we operate (σ=10) and which mirrors the AIA-then-HMI convergence ordering observed in Figure 14.

(1) RFT achieves lower final validation loss across all modalities. At convergence, RFT yields lower reconstruction error on total, AIA, and HMI losses. The loss slope remains steeper at 50k steps, indicating the gap would widen with additional compute.

(2) A harder pretext task produces richer representations. In MultiMAE, 75% of tokens are dropped entirely, the decoder must hallucinate content from spatial neighbors. In RFT, every masked token carries a non-invertible Fourier projection that preserves coarse frequency structure, while destroying fine spatial detail. If $\mathcal{I}_{\text{drop}}$ and $\mathcal{I}_{\text{RFP}}$ denote the mutual information between the masked representation and the original patch, then $0 = \mathcal{I}_{\text{drop}} < \mathcal{I}_{\text{RFP}} < \mathcal{I}_{\text{vis}}$. This places RFP in a regime where the encoder must fuse heterogeneous views rather than relying on spatial interpolation aligned with the multi-view information bottleneck principle [69, 70].

(3) Faster convergence from denser gradient signal. Every token, visible or Fourier-projected, contributes a gradient to the encoder in RFT, whereas MultiMAE's dropped tokens contribute nothing during the forward pass during training. At 75% masking rate, this effectively increases the useful training signal per step by $4\times$.

(4) Sequential modality convergence reveals physical structure. AIA converges first (center panel), followed by HMI (right panel), mirroring the underlying physics: AIA's 8 EUV/UV channels share strong thermal cross-correlations (i.e. see also Figure 3), while HMI vector magnetic components are

largely orthogonal. RFT *accelerates* HMI convergence, indicating that the Fourier tokens provide important cross-modal cues, especially valuable for harder modalities.

K Sensitivity Analysis for Visible Token Ratio

We have trained matched models at different Visible Token Ratio $\rho = 0.25$ (75% RFP tokens) and $\rho = 0.50$ (50% RFP tokens) under identical conditions.

Both settings converge stably, confirming the model is not brittle to this hyperparameter. The ρ=0.50 model achieves lower absolute reconstruction loss, which is expected: with twice as many visible tokens the pretext task is strictly easier. However, lower reconstruction loss does not imply better downstream representations. The central insight from self-supervised learning is that harder pretext tasks those imposing tighter information bottlenecks produce more transferable features [33, 70]. At $\rho = 0.25$, only one in four tokens retains full spatial fidelity, forcing the encoder to rely on cross-modal and cross-spatial integration for the remaining three-quarters.

Our downstream AR segmentation results, where the ρ=0.25 pretrained model achieves 61.1 Dice with a frozen backbone, validate this reasoning: the tighter bottleneck produces features that are more linearly separable for dense prediction. We also note that $\rho = 0.25$ is consistent with the masking ratio found optimal in MAE [33] and MultiMAE [11]. The transient spikes visible in both curves correspond to timesteps containing high solar activity (X-class flares), which produce large reconstruction errors; both models recover identically, indicating robustness.

L Detailed Analysis on Masking Rates and Channel Masking

In Tables 8 and 9, we present the SSIM and MSE evaluation results for our model. We use the Structural Similarity Index to measure reconstruction quality across different channels and masking scenarios. Table 8 shows the results when one channel is fully removed and the others are masked at a 75% rate, which adds to the findings in Figure 3. The omission of individual AIA channels (e.g., 94, 131, or 171Å) results in only marginal increases in the average MSE, often staying within the 0.276 to 0.278 range (compared to 0.271 with all modalities). The model effectively utilizes the high degree of physical redundancy inherent in multi-wavelength data.

Table 9 repeats this test without the spatial masking constraint to evaluate how the model captures cross-modality information. In these tables, columns represent specific channel values, while rows show which modality was removed.

The results show that the average SSIM stays relatively stable regardless of the masking rate. AIA channels demonstrate high consistency between 0% and 75% masking, which indicates that the model uses cross-modal correlations to fill in missing spatial data. In contrast, masking HMI channels more significantly hurts the reconstruction of the dropped channel. When compared to Figure 5, the MSE for individual HMI channels increases substantially when they are masked. Specifically, HMI_m increases from 0.528 to 0.735, HMI_bx from 0.867 to 0.941, HMI_by from 0.894 to 0.972, and HMI_bz from 0.524 to 0.689. The most significant change is in the velocity channel (HMI_v), where MSE rises from 0.0241 to 0.969. Similarly, we find that removing the HMI_v channel leads to the lowest average SSIM and highest MSE in both masking rates. However, its reconstruction stays resilient to spatial masking, with SSIM values of 0.6541 at 0% and 0.6565 at 75% masking. For the SDO AIA channels, the MSE increase is mostly below 0.05, except for AIA 335, which increases by about 0.12.

In addition, the results in the main text in Figure 3 indicate that for the omission of individual AIA channels (e.g., 94, 131, or 171Å), the remaining channels contain sufficient overlapping thermal information to reconstruct the missing channel. Conversely, the model is highly sensitive to the loss of HMI components, i.e. high error in HMI_v reconstruction.

These results are also qualitatively visible in Figure 16 (as well as in Figure 4), where we observe the reconstruction of all channels at different masking rates. In these figures, we observe that many of the relevant features are structurally visible even in 90% masking rates (as observed by over 85% SSIM), however, the scaling for HMI channels changes. We note that even when only 10% of each channel is provided as input, the model is able to decode with high structural fidelity showing that cross-modal interactions are well encoded.

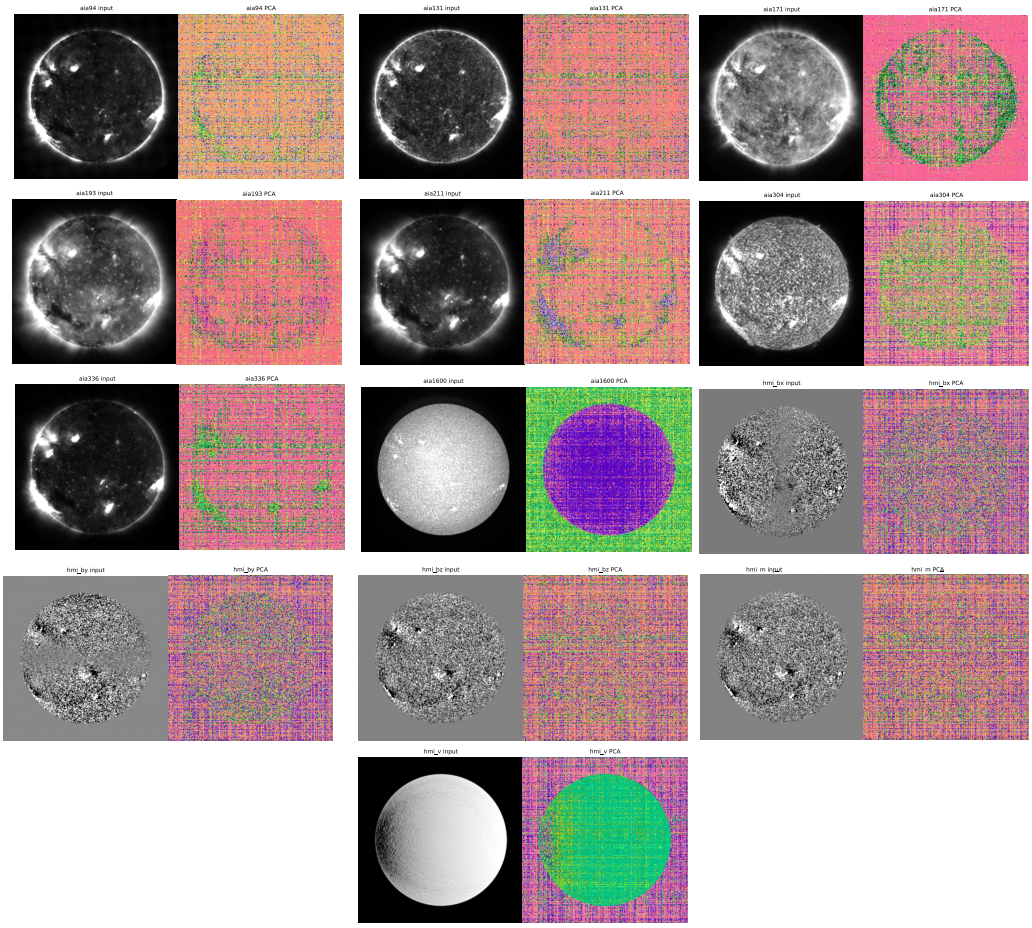

Figure 9: PCA visualization of frozen encoder embeddings. We fit a single 3-component PCA on 500k tokens pooled across all 13 modalities and map the three components to RGB at each spatial patch location. No training is involved; this is a zero-shot probe of the pretrained representation. Active regions and limb structure appear as spatially coherent color clusters across both AIA and HMI channels, indicating that the shared encoder aligns physically distinct measurements into a common latent geometry.

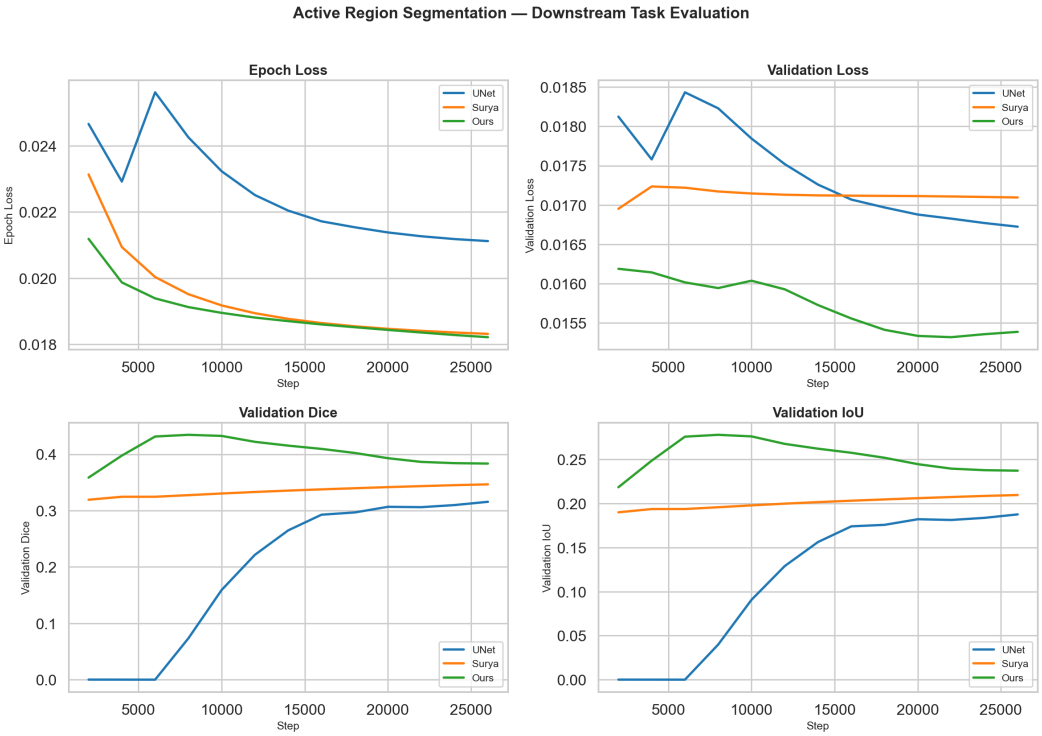

Figure 10: AR segmentation training and validation curves for U-Net, Surya and our model from a single-seed exploratory run that predates the multi-seed protocol of Table 2; absolute values are therefore not comparable to that table.

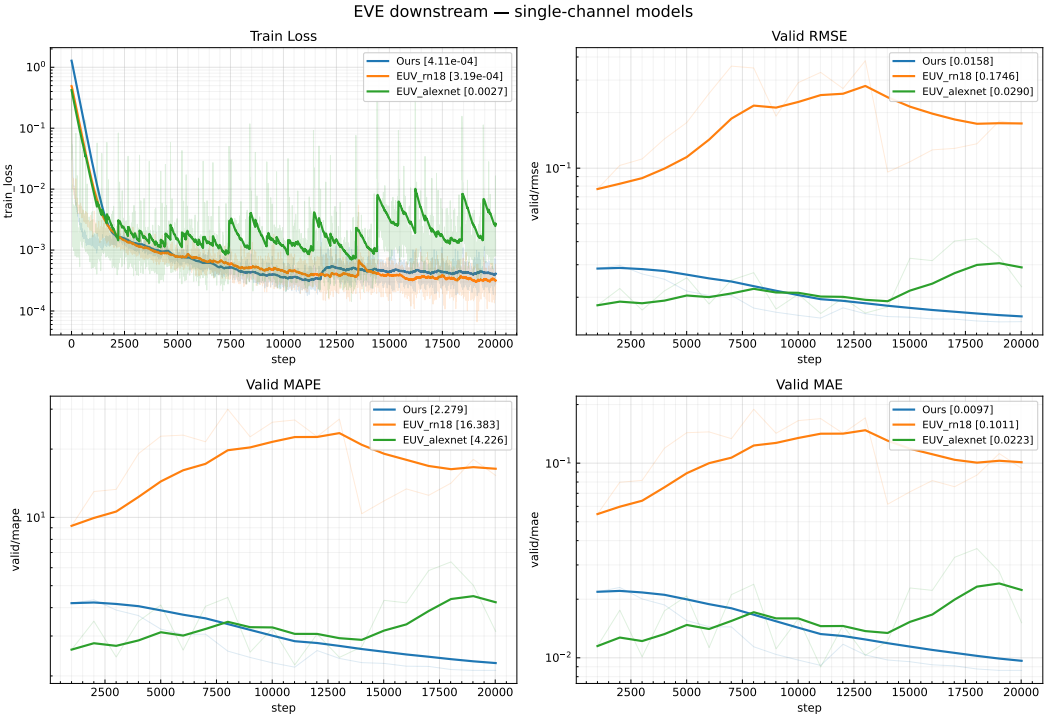

Figure 11: EVE prediction downstream task training curves (single-channel input). Faint lines are raw step-level values; bold lines are smoothed. The final values are shown in the legend. Our frozen-encoder + Perceiver-style head (blue) improves monotonically on every validation metric, whereas end-to-end CNNs (orange, green) trained on the same input degrade or plateau.

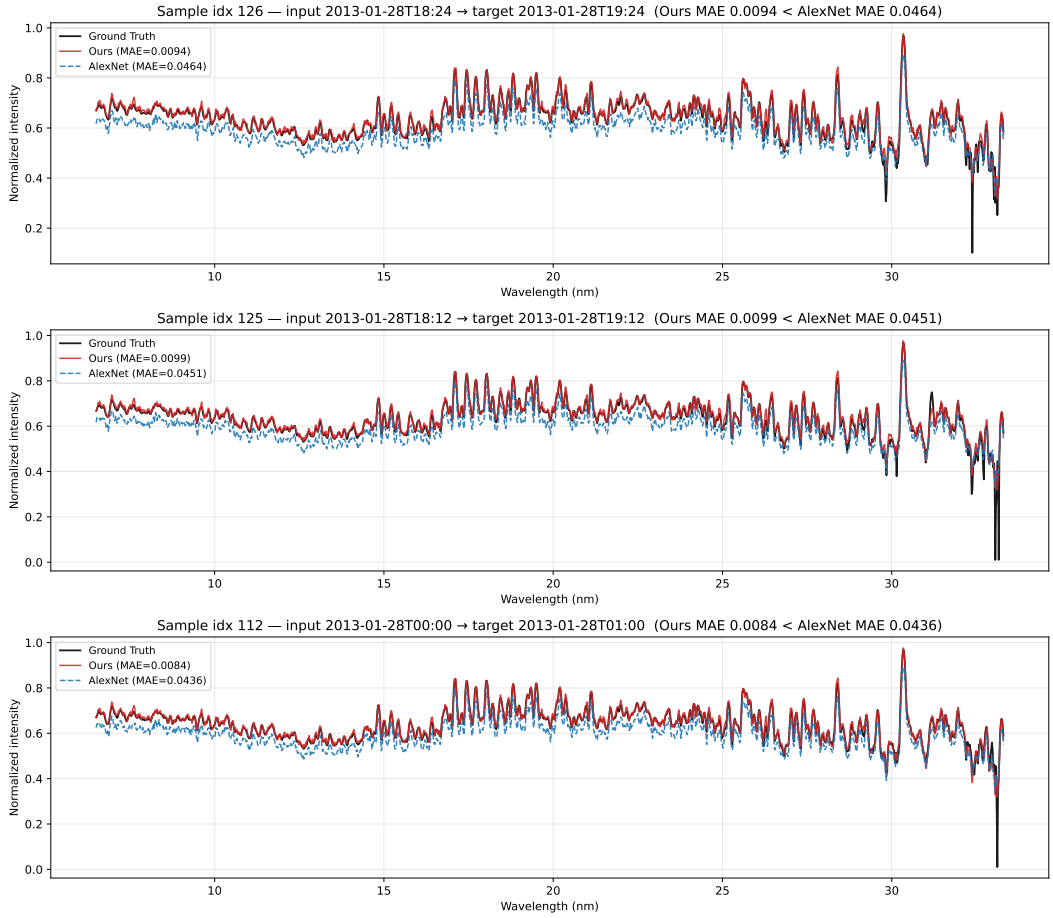

Figure 12: Predicted EVE spectra (∼ 6.5–33.3 nm) for three consecutive test samples on 2013-01-28, plotted against the ground truth. Our model (red) is overlaid with the AlexNet baseline (blue dashed) and the EVE measurement (black). The per-sample MAE is shown in the legend.

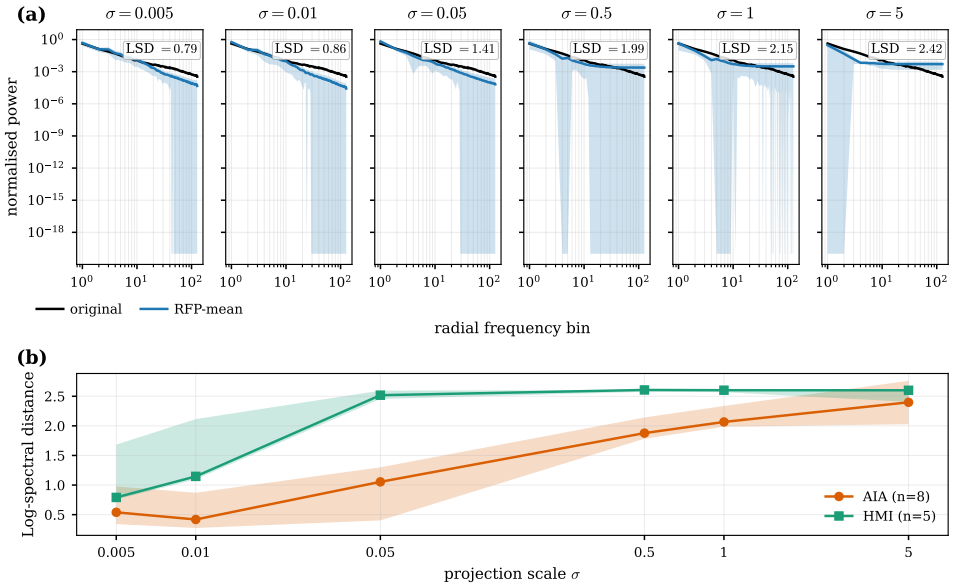

Figure 13: Spectral analysis of RFP representations across six projection scales $\sigma \in \{0.005, 0.01, 0.05, 0.5, 1, 5\}$ on the full 13-channel SDO dataset. **(a)** Top strip, one log–log subplot per σ. Each panel shows the L1-normalised radial 2D power spectral density (PSD) of the patch-mean image, averaged over all 8 samples × 13 channels = 104 (sample, channel) pairs. *Black* curve: original image PSD; *blue* curve: RFP-mean image PSD; *shaded blue band*: ±1 standard deviation of the RFP curve over the 104-curve population (it captures how much the RFP-mean spectrum varies across samples and channels at this σ, not a confidence interval). At $\sigma = 0.005$ the RFP curve tracks the original almost exactly (LSD ≈ 0.79). As σ grows, the per-patch mean of $[\sin(2\pi B z_i), \cos(2\pi B z_i)]$ collapses toward zero in expectation while higher-frequency components of B dominate the variance, so the RFP curve flattens. Because both curves are L1-normalised, this manifests as a crossover: the RFP spectrum lies *below* the original at low radial frequencies (less concentrated low-frequency mass) and *above* it at high frequencies (more diffuse high-frequency mass). Here, the visual story of the σ ablation is therefore not a transition from signal to noise but a redistribution of spectral mass from low to high radial frequency that the encoder can still exploit. **(b)** Log-spectral distance (LSD) between the original and RFP radial PSDs as a function of σ, aggregated by modality. Lines are the median across channels, shaded bands are the inter-quartile range across channels (*not* a confidence interval), separately for AIA ($n = 8$ channels) and HMI ($n = 5$ channels). LSD increases monotonically from ≈ 0.79 at $\sigma = 0.005$ to ≈ 2.42 at $\sigma = 5$ and saturates beyond $\sigma \approx 0.5$. The HMI band can appear one-sided at small σ because the per-channel LSD distribution is right-skewed: most components (`hmi_v`, `hmi_bx`, `hmi_by`) sit at low LSD while `hmi_m` pulls the upper quartile up, and with only five HMI channels the median can coincide with Q_1. This is a small-N skew of the channel-level summary, not a plotting artefact.

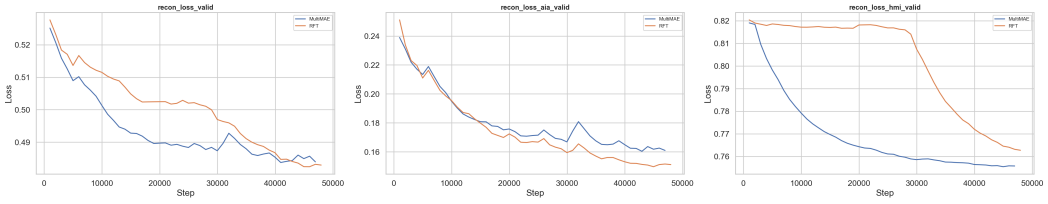

Figure 14: Comparison of RFT method and MultiMAE on AIA and HMI SDO data.

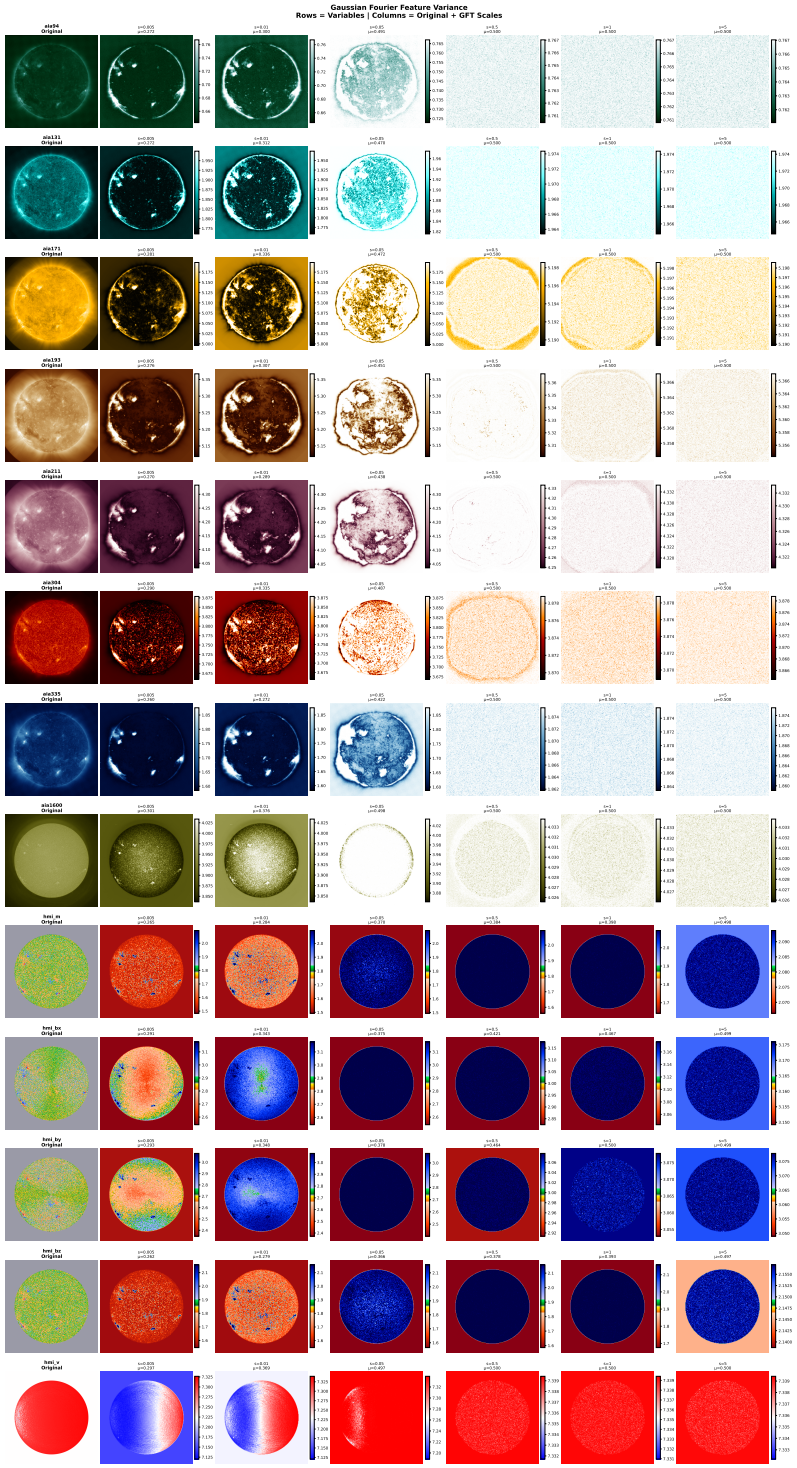

Figure 15: Visualization of Random Fourier Projection (RFP) features computed from SDO AIA and HMI data under varying projection scales $\sigma \in \{0.005, 0.01, 0.05, 0.5, 1, 5\}$, illustrating the effect of frequency bandwidth on the resulting representations.

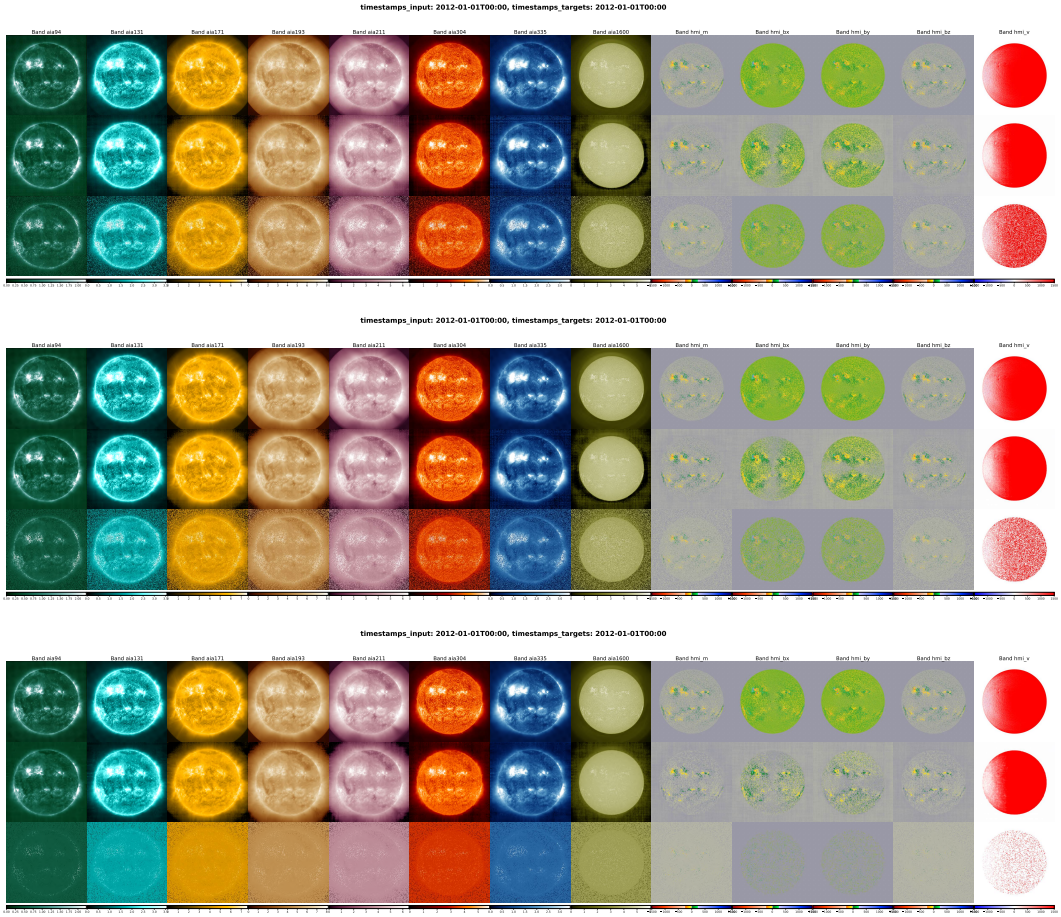

Figure 16: Zero-shot reconstruction with: 25% masking (top 3 rows), 50% masking (middle 3 rows), and 90% masking (bottom 3 rows). For each of these three groups of images: Zero-shot reconstruction of all modalities, where top row in each group is target prediction; middle row is generated reconstruction; bottom row is input to the model.

Table 8: Reconstruction performance evaluation results with SSIM for the reconstruction quality across channels under different masking conditions. Each column in the table represents a specific channel's reconstruction accuracy in terms of SSIM, while each row corresponds to a masking scenario where the indicated channel was removed, other channels are masked at 75%, complementing the results from Figure 3.

Missing	**AIA 94**	**AIA 131**	**AIA 171**	**AIA 193**	**AIA 211**	**AIA 304**	**AIA 335**	**AIA 1600**	**HMI_m**	**HMI_bx**	**HMI_by**	**HMI_bz**	**HMI_v**	**Avg.**
AIA 94	0.9728	0.9817	0.9887	0.9940	0.9846	0.9681	0.9888	0.9837	0.9376	0.8648	0.8605	0.9423	0.9927	0.9585
AIA 131	0.9775	0.9752	0.9873	0.9945	0.9858	0.9673	0.9892	0.9838	0.9374	0.8648	0.8604	0.9422	0.9926	0.9583
AIA 171	0.9764	0.9805	0.9704	0.9947	0.9852	0.9685	0.9896	0.9838	0.9377	0.8648	0.8604	0.9424	0.9926	0.9575
AIA 193	0.9756	0.9820	0.9892	0.9814	0.9848	0.9682	0.9891	0.9838	0.9375	0.8648	0.8604	0.9424	0.9926	0.9578
AIA 211	0.9767	0.9822	0.9893	0.9939	0.9847	0.9684	0.9889	0.9838	0.9376	0.8648	0.8604	0.9423	0.9926	0.9589
AIA 304	0.9766	0.9822	0.9894	0.9945	0.9854	0.9463	0.9893	0.9838	0.9368	0.8648	0.8604	0.9423	0.9926	0.9573
AIA 335	0.9766	0.9818	0.9892	0.9946	0.9837	0.9679	0.9749	0.9837	0.9374	0.8649	0.8604	0.9422	0.9927	0.9577
AIA 1600	0.9769	0.9819	0.9893	0.9949	0.9869	0.9677	0.9892	0.9792	0.9366	0.8644	0.8600	0.9404	0.9925	0.9585
HMI_bx	0.9769	0.9824	0.9894	0.9949	0.9868	0.9681	0.9895	0.9834	0.9319	0.8468	0.8572	0.9366	0.9926	0.9566
HMI_by	0.9768	0.9824	0.9895	0.9947	0.9854	0.9682	0.9893	0.9834	0.9334	0.8613	0.8406	0.9373	0.9927	0.9565
HMI_bz	0.9768	0.9824	0.9896	0.9946	0.9850	0.9680	0.9889	0.9832	0.9244	0.8551	0.8488	0.9286	0.9927	0.9552
HMI_m	0.9771	0.9823	0.9895	0.9948	0.9859	0.9679	0.9893	0.9836	0.9205	0.8581	0.8545	0.9312	0.9925	0.9559
HMI_v	0.9771	0.9824	0.9895	0.9948	0.9850	0.9681	0.9893	0.9838	0.9377	0.8648	0.8603	0.9425	0.6565	0.9332

Table 9: Reconstruction performance evaluation results using SSIM and MSE, when 0% masking is used, for the reconstruction quality across channels under different masking conditions. Each column in the table represents a specific channel's reconstruction accuracy in terms of SSIM and MSE in their respective tables, while each row corresponds to a masking scenario where the indicated channel was removed. Other channels are not masked.

SSIM

Missing	AIA 94	AIA 131	AIA 171	AIA 193	AIA 211	AIA 304	AIA 335	AIA 1600	HMI M	HMI Bx	HMI By	HMI Bz	HMI V	SSIM Avg.
AIA 094	0.9705	0.9792	0.9951	0.9970	0.9906	0.9929	0.9742	0.9909	0.9558	0.8713	0.8872	0.9601	0.9875	0.9656
AIA 131	0.9690	0.9732	0.9953	0.9976	0.9900	0.9928	0.9561	0.9901	0.9543	0.8716	0.8891	0.9588	0.9878	0.9635
AIA 171	0.9707	0.9787	0.9477	0.9976	0.9910	0.9929	0.9656	0.9905	0.9556	0.8702	0.8897	0.9597	0.9850	0.9611
AIA 193	0.9725	0.9803	0.9952	0.9809	0.9882	0.9926	0.9661	0.9905	0.9554	0.8703	0.8892	0.9594	0.9874	0.9637
AIA 211	0.9707	0.9794	0.9954	0.9974	0.9749	0.9925	0.9554	0.9905	0.9552	0.8703	0.8886	0.9592	0.9864	0.9628
AIA 304	0.9720	0.9789	0.9958	0.9976	0.9905	0.9297	0.9637	0.9902	0.9560	0.8702	0.8893	0.9600	0.9869	0.9601
AIA 336	0.9691	0.9764	0.9956	0.9975	0.9894	0.9930	0.9617	0.9901	0.9555	0.8706	0.8886	0.9595	0.9856	0.9640
AIA 1600	0.9746	0.9753	0.9942	0.9975	0.9887	0.9917	0.9635	0.9699	0.9558	0.8694	0.8890	0.9595	0.9864	0.9627
HMI M	0.9719	0.9770	0.9955	0.9979	0.9927	0.9928	0.9792	0.9898	0.9299	0.8662	0.8883	0.9574	0.9843	0.9633
HMI Bx	0.9712	0.9782	0.9948	0.9977	0.9912	0.9923	0.9728	0.9898	0.9541	0.8593	0.8878	0.9568	0.9854	0.9640
HMI By	0.9704	0.9793	0.9952	0.9978	0.9923	0.9924	0.9757	0.9901	0.9547	0.8705	0.7555	0.9583	0.9869	0.9553
HMI Bz	0.9714	0.9790	0.9956	0.9978	0.9919	0.9927	0.9766	0.9903	0.9500	0.8687	0.8983	0.9338	0.9869	0.9641
HMI V	0.9666	0.9803	0.9941	0.9937	0.9846	0.9911	0.9375	0.9910	0.9562	0.8705	0.8900	0.9602	0.6541	0.9361

MSE

Missing	AIA 94	AIA 131	AIA 171	AIA 193	AIA 211	AIA 304	AIA 335	AIA 1600	HMI M	HMI Bx	HMI By	HMI Bz	HMI V	MSE Avg.
AIA 094	0.2114	0.0891	0.0167	0.0092	0.0310	0.0355	0.0686	0.0493	0.3354	0.8509	0.8081	0.3462	0.0258	0.2213
AIA 131	0.1901	0.1414	0.0176	0.0082	0.0326	0.0364	0.0868	0.0519	0.3433	0.8496	0.8001	0.3469	0.0255	0.2254
AIA 171	0.1674	0.0890	0.1219	0.0084	0.0304	0.0359	0.0778	0.0504	0.3401	0.8555	0.7982	0.3458	0.0280	0.2268
AIA 193	0.1603	0.0879	0.0168	0.0434	0.0358	0.0370	0.0756	0.0505	0.3400	0.8550	0.8007	0.3430	0.0260	0.2209
AIA 211	0.1657	0.0874	0.0160	0.0081	0.0608	0.0370	0.0911	0.0504	0.3454	0.8552	0.8033	0.3487	0.0270	0.2228
AIA 304	0.1551	0.0910	0.0156	0.0083	0.0306	0.2094	0.0786	0.0509	0.3450	0.8556	0.8004	0.3511	0.0262	0.2321
AIA 336	0.1735	0.0915	0.0160	0.0082	0.0341	0.0376	0.1624	0.0511	0.3437	0.8543	0.8031	0.3473	0.0276	0.2270
AIA 1600	0.1448	0.0944	0.0186	0.0080	0.0400	0.0393	0.0819	0.1119	0.3452	0.8587	0.8028	0.3541	0.0275	0.2252
HMI M	0.1536	0.0934	0.0158	0.0075	0.0281	0.0359	0.0604	0.0523	0.4887	0.8692	0.8065	0.3843	0.0288	0.2326
HMI Bx	0.1664	0.0904	0.0173	0.0076	0.0317	0.0383	0.0712	0.0521	0.3695	0.9150	0.8080	0.4752	0.0278	0.2362
HMI By	0.1661	0.0889	0.0164	0.0075	0.0294	0.0377	0.0668	0.0515	0.3342	0.8557	1.2834	0.3421	0.0265	0.2543
HMI Bz	0.1616	0.0894	0.0155	0.0077	0.0305	0.0364	0.0638	0.0512	0.3548	0.8576	0.7579	0.5137	0.0263	0.2282
HMI V	0.1665	0.0885	0.0191	0.0152	0.0414	0.0421	0.1115	0.0493	0.3439	0.8542	0.7966	0.3509	0.9790	0.2968

M Decoder Specialization with Frozen Multi-Modal Encoders

Motivation. Large multi-modal foundation models trained across many heterogeneous channels often entangle modality-specific structure inside shared decoder pathways. While this improves global representation learning, it can obscure whether the learned encoder representations are intrinsically sufficient for individual modality reconstruction, or whether performance relies on decoder-side cross-modal coupling. To probe this question, we design a controlled decoder-specialization experiment where the encoder is frozen and the decoder is restructured to explicitly separate modality families.

Experimental Setup. We begin with a foundation model trained jointly on 13 solar modalities using a shared encoder-decoder architecture. In this experiment, the encoder is frozen and initialized with weights from the full 13-modality training regime. Unlike the original setup, the input consists of the complete, unmasked image with no channel dropping, ensuring that all available spatial information is presented to the encoder.

The original unified decoder is replaced by two independent decoders: D_1 and D_2. Decoder D_1 is tasked exclusively with reconstructing the 8 AIA channels and outputs tensors of shape $(8, 4096, 4096)$. Decoder D_2 reconstructs the 5 HMI channels and outputs tensors of shape $(5, 4096, 4096)$. Both decoders are trained from scratch while the encoder remains fixed.

Observed Training Dynamics. Figure 17 shows the training loss trajectories for the total loss, AIA reconstruction loss, and HMI reconstruction loss. A key observation is the highly non-smooth optimization behavior of the HMI decoder. For a substantial number of training steps, the HMI reconstruction loss remains nearly flat, indicating little apparent learning progress. This phase is followed by a sudden and sharp drop in loss, after which the HMI decoder converges rapidly.

In contrast, the AIA reconstruction loss decreases more gradually and smoothly throughout training, exhibiting standard optimization dynamics without abrupt transitions.

Interpretation of the Sudden HMI Loss Drop. The delayed yet abrupt improvement in HMI reconstruction is best understood as a representation alignment phenomenon rather than a numerical instability or optimizer artifact. Because the encoder is frozen, the decoder must learn to invert a fixed latent representation that was originally optimized jointly across all 13 modalities. During early training, the randomly initialized HMI decoder struggles to discover a projection that correctly isolates HMI-specific latent subspaces embedded within the shared representation.

Once the decoder parameters reach a critical alignment with the encoder's latent basis, gradients become informative and the decoder rapidly collapses to a much lower reconstruction error. This behavior is consistent with threshold-driven optimization dynamics often observed in overparameterized neural networks, where meaningful learning only begins after the model discovers an appropriate coordinate system in latent space.

Why This Experiment Matters. This experiment serves three purposes. First, it demonstrates that the frozen encoder learned during 13-modality training already contains sufficient information to reconstruct both AIA and HMI channels independently, without further encoder adaptation. Second, the starkly different learning dynamics between AIA and HMI highlight the asymmetric difficulty of

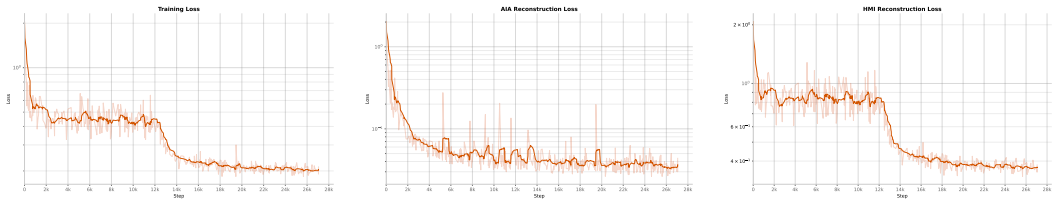

Figure 17: The plot shows the training loss when we have two encoders instead of thirteen for every channel. Here we input the complete image without masking and channel dropping. The encoder is fixed and initialized with the weights of 13 modality training. The decoder is swapped now with two decoders D1 and D2 where D1 produces 8 channels of AIA (that is, the D1 output shape is (8, 4096, 4096)) and D2 produces 5 channels of HMI (i.e., the D2 output shape is (5, 4096, 4906)).

decoding different physical modalities from a shared representation, with HMI requiring more precise latent disentanglement. Third, the results provide evidence that multi-modal foundation encoders implicitly organize modality-specific information in structured subspaces, even when trained with a unified decoder.

Implications for Multi-Modal Foundation Models. The observed phase-transition-like behavior in HMI decoding suggests that downstream task performance may be highly sensitive to decoder initialization and capacity when the encoder is frozen. This has direct implications for transfer learning and modular deployment of large multi-modal models, where specialized decoders may exhibit delayed convergence despite strong underlying representations. Importantly, the experiment confirms that specialization at the decoder level can be an effective strategy for isolating modality families without retraining expensive encoders, provided sufficient optimization time is allowed.

Further discussion of the observed phase-transition-like behavior in HMI decoding. The phase-transition-like behavior arises primarily from modality imbalance during pretraining interacting with the decoder's need to discover a specific latent subspace. The encoder is trained jointly on 8 AIA + 5 HMI channels, but AIA channels dominate both numerically (8 vs. 5) and in terms of cross-spectral redundancy. As a result, the shared latent representation is biased toward AIA-aligned structure.

When a fresh HMI decoder is trained on the frozen encoder, it must identify the lower-dimensional HMI-specific subspace embedded within an AIA-dominated representation. During early training, the randomly initialized decoder projects into subspaces that are largely orthogonal to the HMI-relevant directions, yielding flat loss. Once the parameters align with the correct latent basis a low-probability event in high-dimensional space that nonetheless occurs given sufficient optimization gradients become informative and the loss drops sharply. This is consistent with the "lottery ticket" dynamics observed in overparameterized networks, where meaningful learning begins only after the optimizer discovers a favorable coordinate system.

We also note that AIA's smooth convergence is expected under this interpretation: the AIA-dominated latent space is already approximately aligned with a random decoder initialization, so no phase transition is needed.

N Societal impact

The multi-modal heliophysics foundation model in this work has the potential to significantly advance solar astronomy and space weather science. By providing a better, fine-grained understanding of solar dynamics, our model has the potential to enable researchers to better characterize the physical processes governing the Sun's influence on the solar system. In addition, beyond scientific discovery, this research contributes to the field of Machine Learning (ML) by demonstrating the impact of processing a large set of context tokens on large-scale high-resolution image datasets. This technical advancement serves as a benchmark for future large-scale ML experiments involving complex multi-instrument data. The data used in this work is publicly and freely available.

The applications of this model can streamline the operationalization of space weather forecasting, which is critical for a wide spectrum of stakeholders. Accurate predictions are essential for space agencies, satellite operators, and power grid managers to protect infrastructure and ensure astronaut safety. Industries such as aviation, maritime navigation, telecommunications, agriculture, and even tourism rely on stable space weather conditions to maintain safety and economic continuity. With this model, we lay the foundations for a large-scale knowledge transfer experiment for efficient downstream application development, which is a critical need for both scientific and operational tasks related to space weather.

The development of this foundation model specifically targets three persistent challenges in space weather forecasting: data sparsity, class imbalance, and multi-modal integration. High-impact solar events, such as extreme solar flares, are statistically rare, leading to a lack of training data for traditional predictive models. The proposed model's ability to learn from diverse, multi-modal context tokens allows for more robust performance even when specific event instances are limited. While the benefits for disaster mitigation and technological resilience are clear, the deployment of such models must consider the uncertainties inherent in solar predictions. Future initiatives should focus on refining the interpretability of these large-scale models to ensure that automated space weather alerts are reliable and actionable for decision-makers worldwide.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer:[Yes]

   Justification: The Abstract and Introduction of our paper state our main contributions and these are also substantiated in the paper Section 3.2, 3.3. Our main contributions are in the paper and these are that we train and evaluate a foundation model for heliophysics on multi-modal full-resolution 4096 by 4096 SDO AIA and HMI data, as well as that we adapt MultiMAE by using a fixed Gaussian random Fourier projection (RFP). In addition, we also evaluate our model on the downstream task of Active Region (AR) segmentation, as well as on EVE prediction. The main text of the paper, as well as the Appendix, substantiates our main contributions.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We discuss the limitations of our work in Section 5 (as well as in Appendix B).

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [N/A]

   Justification: The paper does not include formal theoretical results; our contributions are empirical in Section 4, as well as Section 3, and the Appendix of the paper.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: Section 4, as well as Section 3, and the Appendix of the paper describe the datasets, model architectures, training procedures, hyperparameters, and evaluation procedure used. Together with the code which we will release and make publicly available at the camera-ready final paper, this is sufficient to reproduce our main results.

   Guidelines:

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

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Code: `https://github.com/NASA-IMPACT/Surya-2.0`
   Weights: `https://huggingface.co/nasa-ibm-ai4science/Surya-2.0`

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Section 4, as well as Section 3, and the Appendix of the paper describe the data splits, optimizer (i.e. AdamW), learning rate schedule, batch size, number of training steps, and evaluation protocol.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: All downstream results (Tables 2–5) report mean ± one standard deviation over 3 seeds, which vary model initialization and data order on a fixed 10% label subset and fixed SuryaBench splits (Section 4). Pretraining the foundation model on 257 TB of

SDO data (Appendix **??**) is too costly to repeat, so pretraining, reconstruction and scaling results come from a single run per configuration, evaluated with consistent settings across all methods.

Justification: Due to the computational cost of training large-scale models on full-scale benchmarks (i.e. heliophysics foundation model on SDO data of size 257 TB (i.e. Appendix **??**, as well as Section 1)), we report results from a single run per configuration. To ensure fair comparison between models, we use standard benchmarks with established evaluation protocols and apply consistent settings across all methods, as detailed in the Evaluation section and the Appendix.

Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Section 3.3, as well as Section 3 (and 4), and the Appendix of the paper report the hardware used (i.e. 128 NVIDIA GH200 GPUs), and total compute for the main experiments.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and confirm that our research conforms with its guidelines.

Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: Please see Appendix N.

    Guidelines:

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: We release pretrained weights for Surya, trained on publicly available solar observations without human-subject data. We do not foresee significant misuse risks arising from this release.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: All datasets, pretrained models, and code libraries used in this work are cited in Section 4, as well as Section 3, and the Appendix of the paper. We use each asset in accordance with its license.

    Guidelines:

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

13. **New assets**

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

    Answer: [Yes]

    Justification: We will release our code upon acceptance with a README documenting installation, dependencies, usage, and reproduction of the main results.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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

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

    Answer: [N/A]

    Justification: The paper does not involve research with human subjects, so IRB approval is not applicable.

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: LLMs are not used as a component of the core methodology of this work.

    Guidelines:

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