

SWFCast: Fusing Foundation Models and Diverse Data to Forecast Space Weather

Jeremy Bundgaard, PhD (*Booz Allen Hamilton*), Stephen Gerrells (*Booz Allen Hamilton*),
Derek Doyle, PhD (*Booz Allen Hamilton*), Matthew Shoupe (*Booz Allen Hamilton*)

ABSTRACT

Satellites are vulnerable to high radiation from solar storms, making it essential to monitor and predict solar flares and coronal mass ejections (CMEs) that damage satellite electronics, while solar radiation heats the atmosphere, increasing satellite drag leading to increased orbital uncertainties. AI and machine learning (AI/ML) advancements can significantly enhance empirical methods for solar forecasting by leveraging vast datasets and complex algorithms to dramatically improve accuracy. Despite these advancements, current models still face limitations, such as high computational demands and difficulties in predicting rare events. Indirect measurements of extreme ultraviolet (EUV) radiation often fail to capture the full spectrum of solar dynamics that impact the atmosphere. These methods typically rely on empirical models and approximations resulting in forecasts that may not fully account for the complex interactions between solar activity and atmospheric density. Our approach, leveraging direct observation of high-fidelity EUV imagery from NASA Solar Dynamics Observatory (SDO), offers a comprehensive and accurate representation of these solar events. This leads to more accurate predictions and a deeper understanding of how solar dynamics influence atmospheric conditions. At Booz Allen we are developing an MLOps pipeline called Sun2OD, that fuses multimodal foundation model ensembles of solar events, space weather, and orbit determination. The lowest latency information impacting the space environment is light from the sun, therefore the first component of the Sun2OD pipeline that we present is a deep learning-based solar weather foundation model trained on solar imagery, called SDOViT. To achieve this, we successfully developed this vision transformer architecture to extract an information-dense solar feature space from 12-channel SDO imagery. Secondly, we fuse the resultant SDOViT latent space with atmospheric model drivers, such as $F_{10.7}$, F_{30} , Dst, Kp, and Ap, augmenting historical model driver data to train a multimodal space weather forecast model called SWFCast. Preliminary SWFCast results showing 30% improvement in accuracy in F10.7, F30, Dst, Kp, Ap, Hp30, and Hp60 compared to the NOAA 27-day forecast over the last solar cycle; and the 27-day forecast accuracy is drastically improved within the 5-day horizon. The future direction of Sun2OD includes several key model enhancements. First, we will integrate additional datasets, such as radio and particle flux, geomagnetic indices, and coronagraphs, to enrich our models and expand coverage of learned solar dynamics. We will develop solar flare and CME prediction capabilities to provide early warnings to satellite operations. Finally, we will develop deep learning models to improve anomaly detection and maneuver planning based on high-fidelity satellite ephemerides, ensuring robust and adaptive space operations. Together, these advancements form a comprehensive AI-enabled space weather forecasting solution for use across civil, commercial, and defense applications.

1. INTRODUCTION

Space weather intermittently places critical space services at risk. Episodes of elevated solar radiation and geomagnetic disturbance can degrade satellite electronics, induce surface charging, and heat the thermosphere, thereby increasing drag and driving rapid growth in orbit prediction uncertainty. These effects propagate into operational decision making across conjunction assessment, maneuver planning, and custody maintenance. Despite decades of experience, the most widely used forecasting approaches are still primarily basic models and empirical relationships between a small set of scalar indices and subsequent solar or geomagnetic conditions. These methods are simple and easy to apply, but they cannot capture long range patterns, handle rare events with uncertainty, or take full advantage of the new datasets now available from recent instruments [31, 32, 36].

This work addresses these limitations by learning directly from high fidelity imagery and magnetograms. We introduce *Sun2OD*, a multimodal pipeline that combines a self-supervised vision foundation model (*SDOViT*) trained on SDO imagery with a time series forecaster (*SWFCast*) operating on solar drivers. [28, 22, 20, 24, 18, 14] The central hypothesis is that an information dense solar latent space learned from spatially and spectrally rich EUV and UV

observations together with vector magnetograms precursors that are either weakly expressed or altogether absent in scalar proxies; fusing this latent representation with standard drivers should therefore improve both the stability and the accuracy of forecasts at horizons relevant for operations. [37, 21] In the present study we focus on daily forecasts for radio fluxes, with extensions to geomagnetic indices underway, and we use the NOAA style 27-day persistence rule as a strong and understandable baseline for comparison. [6, 21, 5]

The contributions of this paper are threefold. First, we construct a masked autoencoder vision transformer tailored to multi channel solar imagery that learns a compact representation of the structure of the solar disk over an entire solar cycle (2010–2020). Second, we demonstrate a simple but effective fusion strategy in which forecasts produced from the learned image latents and from time series drivers are blended through a learnable coefficient; this design choice emphasizes interpretability and ease of operations while retaining the benefits of multimodality. Third, we evaluate the method against persistence across multi year windows and observe consistent error reductions at the forecast window of a 27-day horizon for radio flux targets, with even larger gains at shorter horizons. The intent is not merely to improve specific metrics, but to establish a principled path from image based representation learning to deployable forecasting services that can be audited, reproduced, and extended to additional modalities such as coronagraph imagery for CME aware prediction and geomagnetic indices forecasting.

2. RELATED WORK

The Frontier Development Lab (FDL) has shown that representation learning on SDO products can materially improve long horizon forecasting. In particular, they developed a variational autoencoder that compresses SDO AIA and HMI data into a learned latent and then used those features to drive a forecasting model, for 27-day forecasts of radio fluxes ($F_{10.7}$, F_{15} , F_{30}) and geomagnetic indices (K_p , a_p), achieving average RMS errors in the five to ten percent range [21, 37, 31]. Other lines of work have explored event focused prediction, such as the ability to forecast the occurrence of γ -class solar flares up to 72 hours in advance using localized HMI magnetic time series together with NCEI flare labels and report that Transformer based architectures outperform LSTM baselines on this task [15, 36].

In computer vision, the Transformer architecture originally developed for language modeling [35] and its pretraining strategies [33] have been adapted to images through patch tokenization and masked prediction objectives [28, 22, 20]. Hierarchical designs such as HiViT demonstrate improved sample efficiency and accuracy, with gains exceeding three percentage points over strong convolutional baselines in some regimes [19]. The cost of these gains is often higher data and memory demand [27], although a growing literature has reduced training requirements through architectural and optimization advances [18, 14, 24, 23, 19]. Our approach follows this trajectory by adopting masked image modeling while explicitly targeting multi-channel heliophysics imagery and by coupling the learned representation with classical drivers in a manner suitable for operationalization.

3. DATA

We treat the solar image stream as a multivariate spatiotemporal process $\{X_t\}_{t=1}^T$ with $X_t \in \mathbb{R}^{C \times H \times W}$, where $H = W = 512$ and C comprise of AIA EUV/UV channels stacked together with HMI vector components. A curated machine learning ready subset of SDO (SDOML) provides spatial and temporal alignment from 2010 to 2020 with downsampling, quality control, and orbit corrections, and includes code to apply the same preprocessing to live streams [32, 31, 9, 8]. In this study we use AIA channels at 6 min cadence and HMI vector magnetograms (Bx, By, Bz) at 12 min cadence. We align modalities to the slower cadence by nearest neighbor timestamp matching, selecting for each t in the HMI grid the AIA time $t' = \arg \min_{\tau} |\tau - t|$ and forming the concatenated tensor $X_t = [X_{t'}^{\text{AIA}} \parallel X_t^{\text{HMI}}]$. Over the decade, this yields on the order of 4.38×10^5 aligned observations; our training subset after alignment and storage constraints comprises approximately 3.3×10^5 multi-channel examples which equates to approximately 3 terabytes worth of data.

Channelwise normalizations act as measurable maps $T : \mathbb{R}^{C \times H \times W} \rightarrow \mathbb{R}^{C \times H \times W}$. For AIA we apply either a logarithmic transform $T_{\text{AIA}}(x) = \log(x - 1)$ or a median scaling $x \mapsto x / \text{median}(x_{\odot})$ computed over the solar disk; for HMI we use min–max scaling $T_{\text{HMI}}(x) = (x - \min x) / (\max x - \min x)$. We denote the transformed inputs by $\tilde{X}_t = T(X_t)$. No additional quality flags beyond SDOML defaults are used. We adopt full disk inputs without limb masking or center-to-limb correction to allow the model to learn such effects implicitly. The exact AIA\HMI set used in the primary results and the channel count in the input tensor is given by a target $C = 12$ via nine AIA bands plus three HMI

components: 94Å, 131Å, 171Å, 193Å, 211Å, 304Å, 335Å, 1600Å, 1700Å and B_x , B_y , B_z for AIA and HMI resp.

Scalar targets and drivers are assembled at daily cadence from public sources. NOAA provides GOES X-ray, electron, and proton flux, radio flux, and geomagnetic indices such as K_p and A_p [6], LISIRD aggregates radio fluxes ($F_{3.2}$, F_8 , $F_{10.7}$, F_{15} , F_{30}) via NRO and NRC/NRCan [3, 2, 7, 5], and AAVSO, LISIRD, and SIDC publishes daily sunspot numbers [1, 10, 4]. We standardize each scalar series within the training window and forward fill rare gaps when resampling to daily cadence for experiments that require higher frequency inputs.

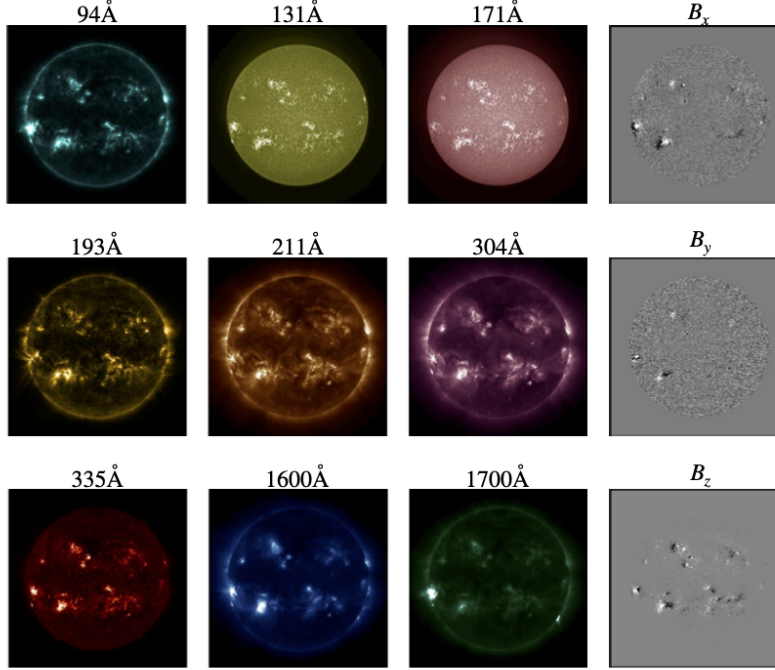


Fig. 1: Example SDOML image set showing AIA hyperspectral channels and HMI vector magnetograms.

4. METHODS

Our approach integrates self-supervised model of spatiotemporal imagery with downstream time series forecasting. The workflow consists of three components: (i) pretraining a hierarchical Vision Transformer (ViT) in the style of masked image modeling, (ii) extracting compact latent representations from the encoder, and (iii) forecasting solar flux indices using both exogenous drivers and imagery derived latents, combined through a multi horizon fusion mechanism. 2

4.1 Patch tokenization and masked pretraining

We employ a masked autoencoding strategy to learn compact representations of solar imagery. Each preprocessed image frame $\tilde{X}_t \in \mathbb{R}^{C \times H \times W}$ is divided into non overlapping square patches of size $P \times P$. This partition produces

$$N = \left(\frac{H}{P}\right) \left(\frac{W}{P}\right)$$

patches per frame and channel.

A linear unfold operator, denoted $\Pi_P : \mathbb{R}^{C \times H \times W} \rightarrow \mathbb{R}^{N \times (CP^2)}$, arranges the channels within each patch into vectors. These are then mapped into a d -dimensional embedding space by a learnable projection matrix $E \in \mathbb{R}^{(CP^2) \times d}$. To encode spatial context, we add two-dimensional positional encodings $p \in \mathbb{R}^{N \times d}$, forming the sequence of input tokens:

$$U_t = \Pi_P(\tilde{X}_t)E + p.$$

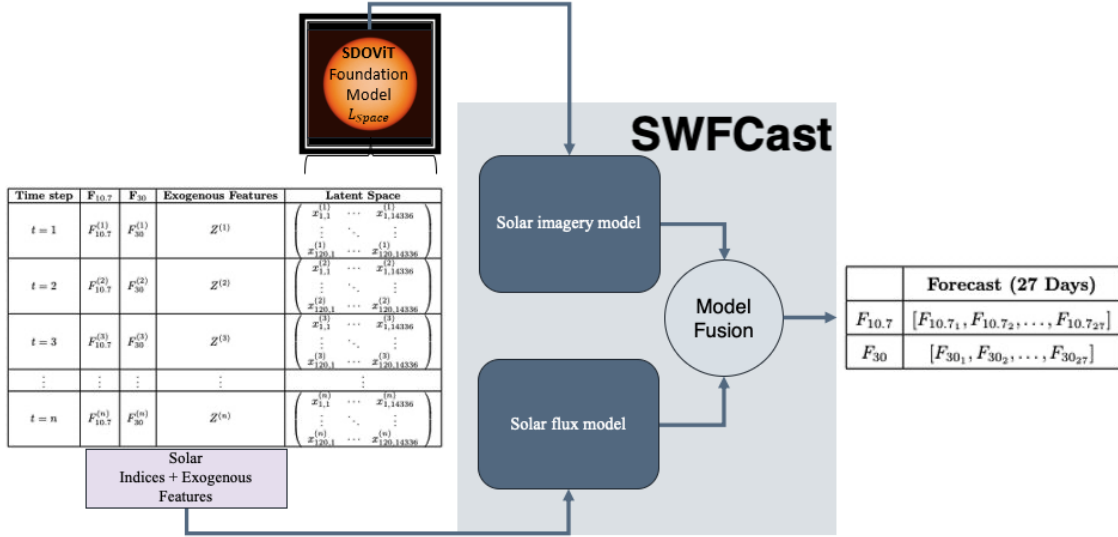


Fig. 2: SWFCast Architecture

A masking operator \mathcal{M} is applied to randomly hide a large fraction of patches. Specifically, a Bernoulli mask with ratio $m = 0.8$ selects the visible subset $\mathcal{V} \subseteq \{1, \dots, N\}$ and the masked subset $\mathcal{M} = \{1, \dots, N\} \setminus \mathcal{V}$.

The encoder, $\Phi : U_{t,\mathcal{V}} \mapsto H_t$, is a hierarchical Vision Transformer that consumes only the visible tokens, producing contextualized features $H_t \in \mathbb{R}^{|\mathcal{V}| \times d}$. A decoder, $\Psi : H_t \mapsto \widehat{\Pi}_P(\tilde{X}_t)$, reconstructs the masked patches in pixel space. Training minimizes the mean squared error between reconstructed and ground truth patches over the masked set:

$$\mathcal{L}_{\text{MAE}}(\Phi, \Psi) = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \|\Psi(H_t)^{(i)} - \Pi_P(\tilde{X}_t)^{(i)}\|_2^2.$$

This process forces the encoder to capture salient solar structures and features, like solar flares, active regions, etc. in its latent representation rather than memorizing raw pixel intensities or interpolating between pixels. Model checkpoints are selected using per-pixel reconstruction error on a held out validation set.

To understand the pretext behavior, Fig 3 shows the original, masked, and reconstructed inputs of an example a SDO image datapoint.

After pretraining, the decoder is discarded and we retain the encoder's latent embedding. Meaning that at the patch level, the token embeddings are mean pooled into a single vector $z_t \in \mathbb{R}^{d_z}$, which serves as a compact information dense representation of the solar disk at time t . In our configuration, with patch size $P = 16$ and embedding dimension $d = 2048$, this yields $d_z = 14,336$. A visualization of the learned latent space see 4

4.2 Forecasting Setup

The forecasting task is to predict daily scalar targets $y_t \in \mathbb{R}^K$, where $K = 2$ corresponds to the radio flux indices $F_{10.7}$ and F_{30} . At each time step, we also incorporate a set of external driver variables $x_t \in \mathbb{R}^P$ that represent traditional space weather indices in addition to exogenous features such as sunspot number.

Prior to training, each target series is standardized to zero mean and unit variance:

$$y_t^{(k)} \mapsto \frac{y_t^{(k)} - \mu_k}{\sigma_k},$$

where μ_k and σ_k denote the mean and standard deviation of the k th target within the training window. This normalization ensures numerical stability so we don't encounter gradient explosion or vanishing and to help the model learn.

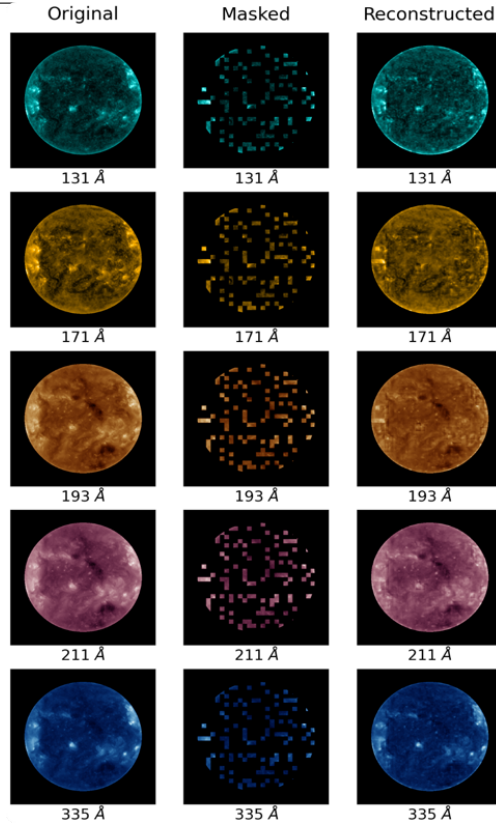


Fig. 3: Masked autoencoding behavior on SDOML

The forecasting model operates with a lookback horizon of $L = 108$ days, providing sufficient temporal context to capture both short term variability and rotation scale patterns. The model produces forecasts across the set of prediction horizons

$$\mathcal{H} = \{1, 2, \dots, 27\},$$

corresponding to daily lead times of up to 27 days.

Fundamentally, the goal is to approximate the conditional expectation of the future targets:

$$\hat{y}_{t+h} = \mathbb{E}[y_{t+h} \mid \mathcal{I}_t],$$

where the information set \mathcal{I}_t comprises both the external drivers and the image derived latent space,

$$\mathcal{I}_t = \{x_{t-L+1:t}, z_t\}.$$

Here, $x_{t-L+1:t}$ denotes the sequence of drivers over the past L days, and $z_t \in \mathbb{R}^{d_z}$ is the compact solar representation extracted from the pretrained encoder. Together, these inputs provide a multimodal basis for forecasting: solar indices capture long term empirical trends, while the latent features supply information about solar phenomenon that is absent from traditional drivers.

4.3 Radio flux and image forecasting.

To translate the inputs into multi horizon predictions, we employ two parallel forecasting modules based on long short term memory (LSTM) networks.

Radio flux based forecaster: The first forecaster operates on the external drivers, e.g. radio flux. We denote this mapping by

$$\Phi_{\text{ts}} : \mathbb{R}^{L \times p} \longrightarrow \mathbb{R}^{K \times |\mathcal{H}|},$$

where the input is the $L \times p$ lookback window of driver variables. The LSTM captures temporal dependencies through its recurrent gating mechanisms and a linear output head produces forecasts for all targets and horizons simultaneously.

Image based forecaster: In parallel, we construct a second LSTM forecaster that operates on the latent representation of the solar imagery. This mapping is defined as

$$\Phi_{\text{img}} : \mathbb{R}^{d_z} \longrightarrow \mathbb{R}^{K \times |\mathcal{H}|},$$

where $z_t \in \mathbb{R}^{d_z}$ is the compact solar embedding extracted during pretraining.

Both modules therefore yield forecasts with identical structure: a matrix of size $K \times |\mathcal{H}|$, where rows correspond to the two targets ($F_{10.7}$ and F_{30}) and columns correspond to the forecast horizons (1–27 days). The next stage of the pipeline combines these complementary predictions through a fusion mechanism, enabling the model to balance information from traditional drivers and image derived features.

4.4 Multi horizon Fusion

The forecasts from the radio flux based and image based modules are combined through a learned convex weighting mechanism that is both horizon and target specific. This fusion step allows the model to adaptively determine how much to rely on radio flux features versus image derived features depending on the forecast setting. What this means, is that for each target $k \in \{1, \dots, K\}$ and forecast horizon $h \in \mathcal{H}$, the fused prediction is

$$\hat{y}_{t+h}^{(k)} = \alpha_{k,h} \hat{y}_{t+h}^{\text{img},(k)} + (1 - \alpha_{k,h}) \hat{y}_{t+h}^{\text{ts},(k)},$$

where $\hat{y}_{t+h}^{\text{img},(k)}$ and $\hat{y}_{t+h}^{\text{ts},(k)}$ denote the outputs of the image based and driver based forecasters, respectively.

The fusion weights $\alpha_{k,h}$ are parameterized as

$$\alpha_{k,h} = \sigma(a_{k,h}) \in (0, 1),$$

with $a_{k,h}$ as learnable parameters and $\sigma(\cdot)$ the logistic function. This parameterization ensures valid convex weights, meaning that the fused forecast is always a weighted average of the two sources.

An important feature of this design is that $\alpha_{k,h}$ can be interpreted directly as an attribution score: values close to 1 indicate greater reliance on imagery derived information, while values near 0 indicate stronger reliance on radio flux inputs. This provides a transparent mechanism for understanding how the model balances multimodal information across different forecast horizons.

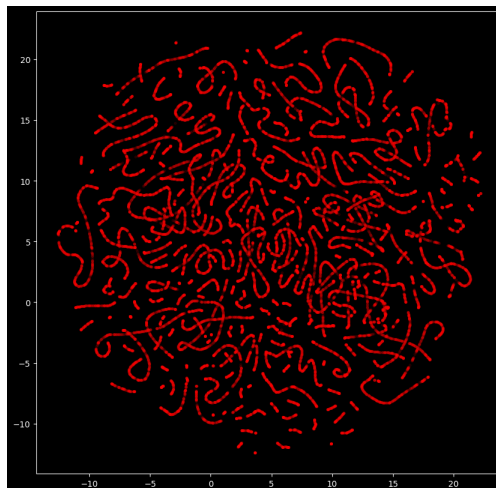


Fig. 4: Learned Latent Space

5. EXPERIMENTS AND RESULTS

All experiments are conducted at a daily cadence with strict temporal alignment between inputs and targets. As a baseline we adopt the NOAA 27-day persistence rule

$$\hat{y}_{t+h}^{\text{pers}} = y_{t+h-27},$$

which reflects rotation scale recurrence. The primary variables considered are the radio flux indices $F_{10.7}$ and F_{30} ; geomagnetic indices ($K_p, A_p, Dst, Hp_{30}, Hp_{60}$) are the subject of ongoing work.

5.1 Evaluation protocol and metrics

Forecast accuracy is reported as (i) root mean squared error (RMSE), (ii) mean absolute error (MAE), and (iii) a persistence normalized skill score. [30]. Unless stated otherwise, RMSE and MAE are computed on *relative* errors at each horizon, $\epsilon_{t,h} = \hat{y}_{t+h}/y_{t+h} - 1$, and expressed in percent by multiplying by 100. That is to say, for variable k ,

$$\text{RMSE}_k = \sqrt{\frac{1}{|\mathcal{T}_{\text{eval}}||\mathcal{H}|} \sum_{t \in \mathcal{T}_{\text{eval}}} \sum_{h \in \mathcal{H}} (\epsilon_{t,h}^{(k)})^2}, \quad \text{MAE}_k = \frac{1}{|\mathcal{T}_{\text{eval}}||\mathcal{H}|} \sum_{t,h} |\epsilon_{t,h}^{(k)}|. \quad (1)$$

The persistence skill is

$$\text{Skill}_k = 1 - \frac{\text{RMSE}_k}{\text{RMSE}_k^{\text{pers}}}, \quad (2)$$

reported per horizon by evaluating (1) at a fixed h . [30] When integrating across long spans with occasional gaps, we report a time weighted RMSE that down weights short missing segments.

5.2 Main Quantitative Results

Table 1 reports the window averaged forecast performance across horizons. For the $F_{10.7}$ radio flux index, our fused model reduces RMSE from 14.94% under persistence to 4.32% at a 1 day horizon (Skill = 71.1%). At a 5 day horizon, RMSE decreases from 14.94% to 9.75% (34.7% skill), and at 27 day from 14.91% to 12.95% (13.1% skill). Results for F_{30} follow a similar trend: 10.99% \rightarrow 2.91% at one day (73.5% skill), 10.98% \rightarrow 6.91% at 5 days (37.1%), and 10.96% \rightarrow 10.13% at twenty-seven days (7.6%).

Relative MAE exhibits consistent reductions across all horizons, with one-day errors falling from 11.56% \rightarrow 3.36% for $F_{10.7}$ and from 8.72% \rightarrow 2.34% for F_{30} . Pearson correlation coefficients likewise confirm strong shorthorizon agreement with observations, reaching $r = 0.984$ for $F_{10.7}$ and $r = 0.994$ for F_{30} at one day, with values decaying gradually as the forecast horizon extends.

Table 1: NOAA vs. SWFCast main results. RMSE shown in %; MAE in % (relative).

Var	Horizon	RMSE(NOAA)	RMSE(SWFCast)	Δ RMSE%	MAE(NOAA)	MAE(SWFCAST)	$r(\text{O})$
F10.7	1d	14.94	4.32	71.1	11.56	3.36	0.984
F10.7	5d	14.94	9.75	34.7	11.55	7.59	0.914
F10.7	27d	14.91	12.95	13.1	11.49	10.35	0.840
F30	1d	10.99	2.91	73.5	8.72	2.34	0.994
F30	5d	10.98	6.91	37.1	8.71	5.46	0.964
F30	27d	10.96	10.13	7.6	8.65	8.09	0.918

Overall, the results point to two clear performance regimes. At short forecast horizons ($\sim 1-5$ days), the model shows strong gains in forecast accuracy, possibly indicating that the image based latent features capture short time scale changing of solar structures. These localized and short lived patterns in the latent space directly reduce errors on smaller timescales.

At the longer 27-day forecast horizon, the gains are smaller but remain consistently above the baseline. In this case, the latent features appear to reflect slower, rotation scale patterns in solar activity. This looks to be effectively providing a stabilizing influence on top of the traditional driver inputs, limiting error growth over extended forecasts. Thus, combining image derived feature embeddings with timeseries drivers produces a representation that is sensitive to rapid changes while still able to capture long term solar downstream activity.

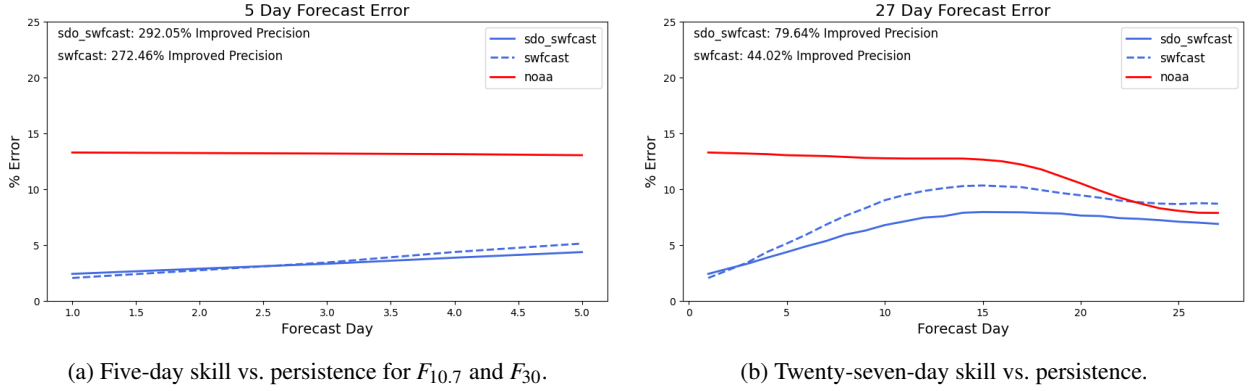


Fig. 5: Relative improvement at 5-day and 27-day horizons rendered at matched width for consistent visual scale.

5.3 Integrated Error over a Solar Cycle

To quantify the long term operational benefit of our approach, we move beyond pointwise accuracy metrics and instead compute an *integrated error ratio* across the full solar cycle from 2010–2020. This ratio compares the cumulative forecast error of our fused model to the NOAA 27-day persistence baseline, with values below unity indicating improvement.

Rather than summing raw squared errors directly, we employ a *windowed root mean squared error* (RMSE) procedure to obtain a smoothed and horizon aware estimate of forecast accuracy. For a given forecast error sequence $\{e_t\}$ and a window of length W , we compute

$$\text{RMSE}_{t:h}^{(W)} = \sqrt{\frac{1}{W} \sum_{i=t}^{t+W-1} e_{i,h}^2},$$

where $e_{i,h} = \hat{y}_{i,h} - y_{i,h}$ is the forecast error at horizon h . This produces a time series of horizon dependent RMSE values, which we then average across the evaluation period to yield a representative error profile. In addition to absolute errors, we also compute relative errors $\hat{y}/y - 1$ to ensure scale invariance when comparing indices of different magnitudes.

The integrated improvement score is then defined as

$$\text{Improvement} = \frac{\overline{\text{RMSE}}_{\text{baseline}}}{\overline{\text{RMSE}}_{\text{model}}} - 1,$$

where $\overline{\text{RMSE}}$ denotes the temporal mean of the windowed RMSE values. By construction, positive values of Improvement indicate that the proposed model reduces cumulative error relative to persistence.

Figures 6 and 7 show the resulting ten-year integrated error ratios for the $F_{10.7}$ and F_{30} indices, respectively. In both cases, our fused SDOViT+SWFCast model accumulates substantially less error than the persistence baseline, consistent with the per horizon skill improvements reported earlier.

6. ABLATIONS AND SENSITIVITY

We explored the design space along image resolution, patching, model capacity, masking, normalization, forecaster choice, and fusion design. To make a direct comparison, the foundation model (SDOViT) always ingests the full twelve channels (nine AIA EUV/UV bands and three HMI vector components), and all runs share the same preprocessing pipeline and optimizer schedule.

Resolution $(H, W) \in \{128, 256, 512\}^2$. The resolution modulates both the representational ceiling for fine coronal structure and the quadratic cost of attention. With patch size P , the token count scales as $N = (H/P)^2$, and under masked autoencoding with mask ratio m the encoder operates on $(1 - m)N$ visible tokens, with dominant cost $\mathcal{O}(((1 - m)N)^2 d)$ and memory $\mathcal{O}(((1 - m)N)d)$ for hidden width d . Empirically, $128^2 \rightarrow 256^2$ preserved most reconstruction behavior, while 512^2 captured higher spatial frequencies (e.g., loop bundles, moss, plage filamentation) that

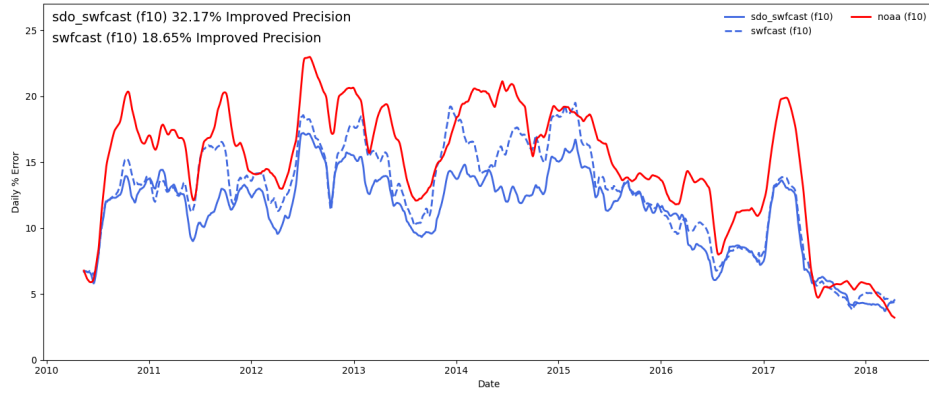


Fig. 6: Ten-year integrated error ratio for $F_{10.7}$ forecasts compared to persistence.

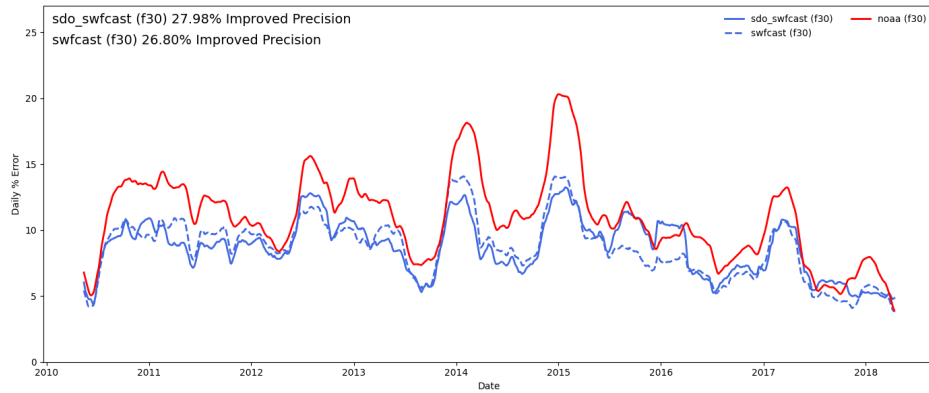


Fig. 7: Ten-year integrated error ratio for F_{30} forecasts compared to persistence.

are plausibly predictive of short horizon radio flux. We therefore favor 512^2 for the final model and accept the modest throughput trade.

Patch size $P \in \{32, 16, 8, 4\}$ and **hidden width** $d \in \{256, 512, 1024, 2048\}$. We investigated the influence of patch size and the dimension of the latent space on model performance and efficiency. Patch size directly controls how finely the input image is divided: smaller patches generate more tokens, enabling the model to capture subtle and finer local structures but also significantly increasing computational cost and training time. Larger patches provide a more global view of the feature space, and reduce the number of tokens and accelerate training, but at the expense of discarding fine grained spatial detail that may be predictive of solar activity.

In our experiments, a patch size of 16 provided the best balance, offering sufficient resolution to preserve small scale solar features while keeping training feasible. The dimension of the latent space, which governs the representational capacity of the transformer, showed a similar trade off. Narrower widths struggled to stably encode cross channel dependencies, while very wide settings imposed heavy memory demands. A width of 2048 dimensions yielded the most consistent results, capturing inter band relationships without introducing instability.

Although training at 512×512 resolution with these settings required roughly three times longer than with lower resolution or coarser patch configurations, the observed improvements in short horizon radio flux forecasting justified the additional cost. See Fig. 8 for the broader hyperparameter landscape

Mask ratio $m \in \{0.2, 0.4, 0.6, 0.8\}$. We evaluated a range of masking ratios during pretraining to understand how the level of occlusion influences model learning and downstream forecasting performance. The mask ratio essentially governs how much of the input image is hidden from the encoder during self-supervised training. Lower ratios expose

too much of the original image, which weakens the denoising pressure and leads to under regularization and interpolation. At the other extreme, very high ratios deprive the encoder of sufficient context to meaningfully reconstruct the missing regions, resulting in unstable training. Our experiments were consistent with prior findings in the masked autoencoder literature [22], showing that a mask ratio of eighty percent provided the best balance. At this setting, the model achieved stable pretraining dynamics, reliable downstream forecasts, and the additional benefit of reduced computational cost, since the encoder only processes a smaller set of visible tokens. As shown in Fig. 8 the most optimal masked ratio is $m = 0.8$.”

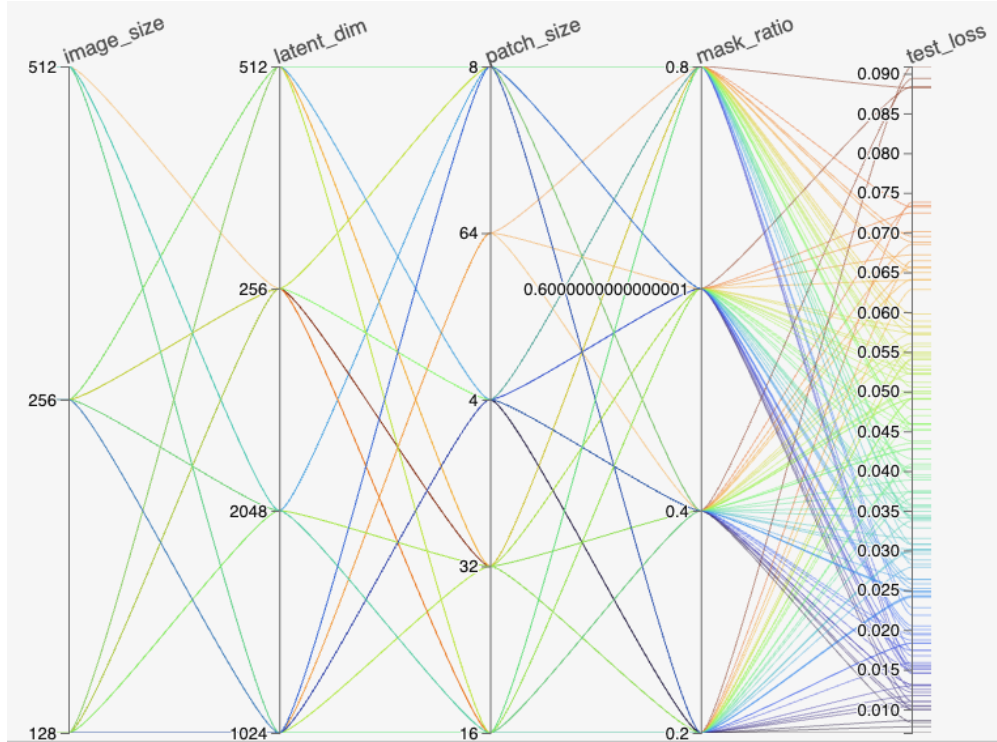


Fig. 8: HiPlot visualization of the SDOViT hyperparameter search across latent dimension, patch size, image resolution, and mask ratio.

Normalization. The raw input data from the Atmospheric Imaging Assembly (AIA) is characterized by an extremely skewed dynamic range, which, if left uncorrected, can lead to unstable optimization behavior, e.g. gradient explosion or gradient vanishing. To fix this issue we experimented with a range of normalization schemes. The most effective configuration combined a per channel logarithmic compression of the form $\log(x + 1)$ with a disk level median scaling for the AIA channels, a strategy that preserved relative intensity differences while reducing the impact of extreme values. For the Helioseismic and Magnetic Imager (HMI) data, we adopted min–max scaling to constrain values within a fixed interval, thereby preventing saturation during initialization.

In addition to these baselines, we evaluated several alternatives. A simple minimum shift transform ($x \rightarrow x - \min(x)$) was tested and then a combination of normalization schemes. These variants produced broadly similar asymptotic forecasting performance but exhibited differences in training stability and convergence rate. We also explored a more domain specific approach, applying a frequency domain filter to the imagery. This method involved transforming images into the Fourier domain, constructing a mask to suppress selected spatial frequencies intended to remove grid like artifacts that could have been from the SDO processing pipeline and or also just corrupted data then transforming back to the image domain. However, this spectral filtering approach yielded negligible impact on downstream results and was therefore not adopted in the final system.

The final candidate that was chosen was a logarithmic compression with median scaling for AIA and min–max scaling for HMI offered the best balance of stability and performance, while alternative methods mainly affected optimization

dynamics rather than long term forecast accuracy. This finding is consistent with the broader literature on vision transformers, which indicates that once inputs are brought into a reasonably stable range, the architecture can learn effective rescalings internally [28, 22, 27]

Modal content In order to preserve as much potentially useful information as possible, the final system was designed to ingest all twelve available channels without pruning. These include nine extreme ultraviolet and ultraviolet bands from AIA and three magnetic field components from HMI. While we did not conduct any ablation studies that isolates the contribution of AIA alone or any subset, our expectation is that they all contribute some amount and didn't have time to isolate prime channels.

Forecaster choice and hyperparameter search (SWFCast). Several forecasting architectures were considered, including teacher–student frameworks; however, a standard long short term memory (LSTM) network with a 108-day lookback horizon was ultimately selected as the most reliable baseline. The hyperparameter search space encompassed hidden dimensionality, network depth, dropout rate, learning rate, activation functions, optimizers, bias initialization, and regularization strength. This search was conducted systematically using the Ray framework for distributed orchestration, while PyTorch Lightning was employed to ensure training reproducibility and experimental control. Model configurations were evaluated under time blocked data splits to mitigate temporal leakage and ensure solar flux dynamics was fully captured, with the final selection based on highest validation performance. Although this initial study retained the classical LSTM for stability and interpretability, future work will investigate more advanced architectures, such as Transformer based forecasters, SimLSTM, and xLSTM [17, 16, 13].

Fusion design. The SDOViT latent representations are fused with driver (radio flux) only forecasts through a learned, horizon and variable dependent convex combination,

$$\hat{y}_{t+h}^{(k)} = \alpha_{k,h} \hat{y}_{\text{img},t+h}^{(k)} + (1 - \alpha_{k,h}) \hat{y}_{\text{drv},t+h}^{(k)}, \quad \alpha_{k,h} \in [0, 1].$$

This formulation ensures robustness by providing a fallback to driver (radio flux) based predictions in cases where imagery is unavailable, while also offering interpretability, as the learned coefficients $\alpha_{k,h}$ explicitly quantify the relative contribution of imagery versus drivers. Figure 9 shows the training dynamics of $\alpha_{k,h}$, where the model adaptively balances image derived and driver based forecasts. Preliminary experiments indicate that this simple fusion mechanism performs remarkably well, achieving a favorable balance between the solar flux LSTM and the solar imagery LSTM. Nonetheless, the approach remains relatively naive. A natural extension is the use of cross attention between image latents and driver embeddings, enabling the forecaster to condition on where relevant image evidence resides rather than relying solely on a pooled latent. We anticipate that such cross attention mechanisms will be particularly beneficial for geomagnetic targets and CME aware forecasting.

7. DISCUSSION

The results presented here suggest that there is some alignment with the underlying physics of solar phenomena. Extreme ultraviolet (EUV) imaging thus captures thermodynamic characteristics of the corona that are not accessible via scalar indices alone. Similarly, Helioseismic and Magnetic Imager (HMI) magnetograms trace the evolution of photospheric magnetic fields. When combined, these modalities encode precursors to fluctuations in radio flux scalar drivers alone cannot represent. The SDOViT foundation model is therefore able to learn a joint latent representation that preserves physically meaningful structure, enabling forecasts that outperform persistence and driver only baselines.

The convex fusion strategy further emphasizes operational interpretability. The horizon and variable dependent coefficients $\alpha_{k,h}$ serve as explicit attribution weights, quantifying the relative contribution of imagery and drivers at each forecast horizon. This design provides a principled way to understand when image derived information meaningfully improves predictive accuracy and when reliance on canonical drivers remains sufficient.

The observed performance gains reveal two distinct regimes. At short horizons (1–5 days), SDOViT's latent features deliver substantial improvements, suggesting that solar morphology encodes transient, rapidly decaying predictive

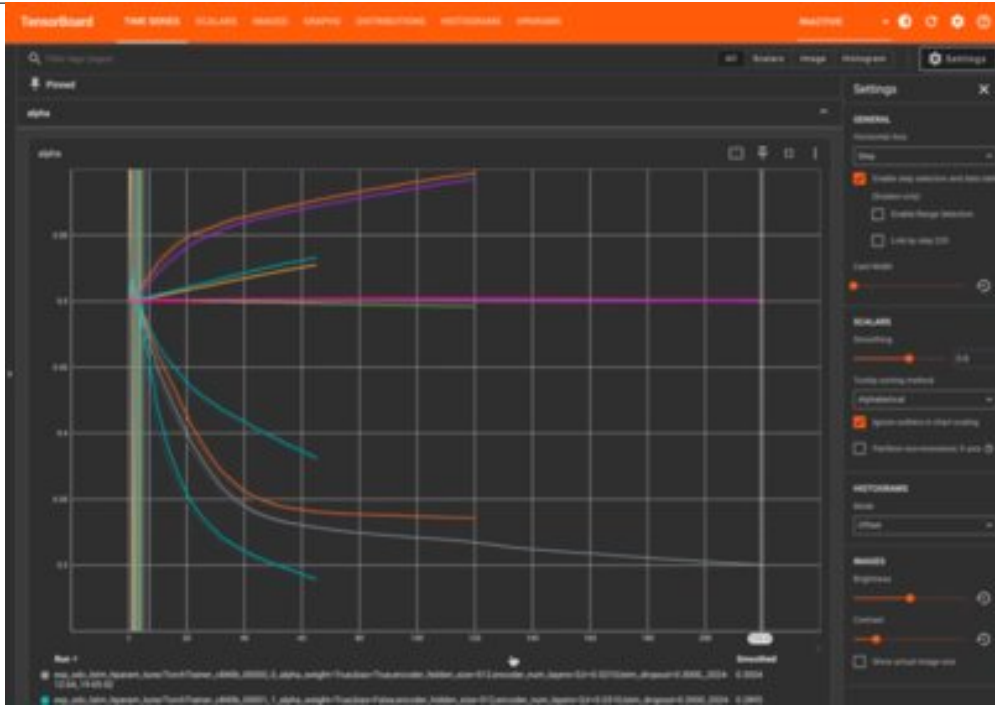


Fig. 9: TensorBoard trace of the learned fusion coefficients $\alpha_{k,h}$

signals that standard approaches are unable to capture. At longer horizons (e.g., 27 days), the improvements, though more modest, indicate that the learned latent space also captures rotation scale regularities, thereby stabilizing forecasts driven primarily by scalar proxies. This dual capability demonstrates the value of multimodal fusion, that is to say imagery improves responsiveness to short term dynamics while reinforcing stability at extended horizons.

8. LIMITATIONS

Three main limitations of the present study should be understood. First, the use of random splits introduces the risk of information leakage across seasonal or rotational timescales, potentially inflating the performance. To mitigate this concern, future evaluations will be conducted on post-2020 data and will adopt blocked time splits with a temporal buffer between the latest training instance and the earliest test instance, thereby reducing dependence on recurrent solar patterns.

Second, the forecasts presented here are strictly deterministic, providing point estimates without any measure of predictive uncertainty. While this is sufficient for demonstrating baseline improvements over standard models, it limits the operational value of the forecasts in risk sensitive situations. Incorporating probabilistic prediction heads and calibration strategies would enable the generation of calibrated uncertainty estimates, thereby supporting uncertainty aware decision making.

Third, the interpretability of the learned representations remains underexplored. While the SDOViT foundation model demonstrably improves predictive accuracy, the physical meaning encoded in its latent space has not yet been systematically analyzed. Future work will include embedding visualizations and dimensionality reduction techniques such as UMAP, with solar images superimposed to assess clustering behavior.

9. CONCLUSION & FUTURE WORK

This study introduced *Sun2OD*, a multimodal forecasting framework that combines a self-supervised foundation model trained on SDO imagery (SDOViT) with a driver based LSTM forecaster (SWFCast) through a learned fusion layer. Results across a full solar cycle demonstrate that this approach consistently outperforms persistence and standard

model baselines, with particularly strong gains at short horizons (1–5 days) and sustained improvements at the 27-day horizon. These findings highlight the operational value of fusing high fidelity solar imagery with canonical drivers, enabling forecasts that are both more accurate and more stable across forecasting regimes.

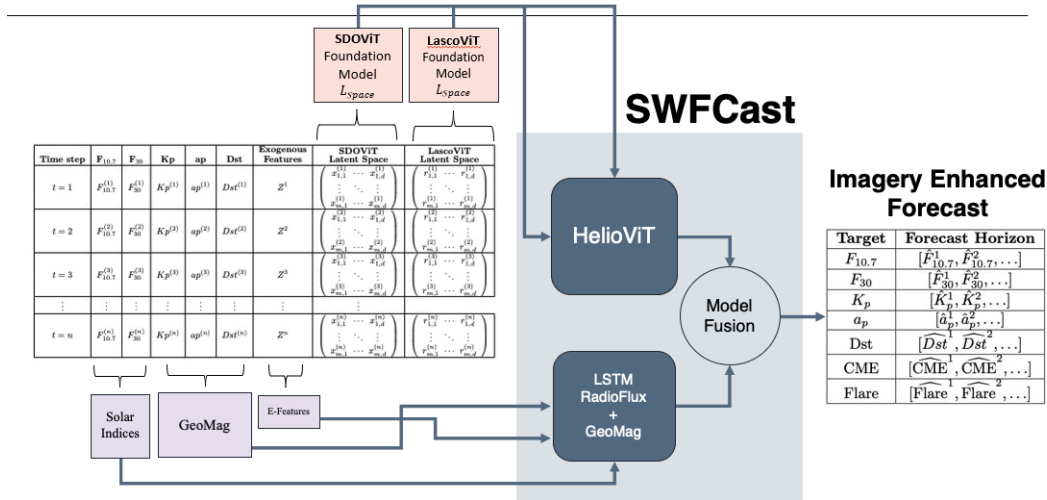


Fig. 10: Planned multimodal architecture extending Sun2OD.

Future work will extend this framework in several directions. First, while preliminary experiments have applied Sun2OD to geomagnetic indices, the results were limited, with forecasts collapsing toward mean values. This outcome is not unexpected, given the nonlinear and highly complex nature of magnetospheric dynamics. To address this, we have begun development of a complementary hierarchical foundation model trained on LASCO C2 and C3 coronagraph imagery, which we term LascoViT. The supporting infrastructure has been established and prototype models trained to validate feasibility prior to scaling to longer training runs. Our objective is to fuse LascoViT with SDOViT and driver based models in order to provide improved forecasts of geomagnetic indices, as well as to extend capabilities to solar flare and CME prediction. Cross attention fusion strategies will also be investigated to further enhance multimodal integration. This can be illustrated in Fig. 10, where SDOViT, LascoViT, and canonical drivers are fused to support geomagnetic forecasting and CME aware prediction.

Second, interpretability remains an important open direction. Embedding analyses using dimensionality reduction techniques such as UMAP, potentially augmented with topological data analysis, will allow us to examine whether the learned latent space organizes according to physically meaningful solar structures. In particular, recent work on Topological Autoencoders demonstrates how persistent homology can guide latent representations to retain multi scale connectivity patterns. Other studies apply TDA directly to model activations meaning that using persistent homology and persistence landscapes to summarize the evolving topology of network activations or constructing simplicial complexes to characterize latent transformations across model layers. Similarly, analysis of latent space “holes” in language models through persistent homology offers a lens into interpretability under adversarial conditions [25, 26, 11, 12, 29, 34].

Finally, while deterministic forecasts have been sufficient for demonstrating baseline improvements, future iterations will explore probabilistic extensions in order to provide calibrated uncertainty estimates. Such developments would further support operational readiness in risk sensitive environments.

ACKNOWLEDGMENTS

We would like to acknowledge the significant contributions of our colleague and coauthor, Jeremy Bundgaard, PhD, whose vision and dedication were central to the development of this work. Jeremy’s curiosity, insight, leadership, and generosity of spirit continue to inspire us, and this paper is dedicated to his memory.

REFERENCES

- [1] American Association of Variable Star Observers (AAVSO). *American Association of Variable Star Observers (AAVSO)*. <https://www.aavso.org/solar>.
- [2] Laboratory for Atmospheric and Space Physics (LASP). *Laboratory for Atmospheric and Space Physics (LASP)*. <https://lasp.colorado.edu/>.
- [3] LASP. *LASP Interactive Solar IRradiance Datacenter (LISIRD)*. <https://lasp.colorado.edu/lisird/>.
- [4] LASP, University of Colorado Boulder. *SILSO International Sunspot Number — LASP Interactive Solar IRradiance Datacenter (LISIRD)*. https://lasp.colorado.edu/lisird/data/international_sunspot_number/. Accessed [insert date you accessed it]; includes F10.7 cm, F30 cm solar radio flux, and sunspot number series.
- [5] National Research Council Canada (NRC) and Natural Resources Canada (NRCan). *Penticton Solar Radio Flux at 10.7 cm*. https://lasp.colorado.edu/lisird/data/penticton_radio_flux.
- [6] NOAA. *NOAA GOES X-Ray Flux*. <https://www.swpc.noaa.gov/products/goes-x-ray-flux>.
- [7] Nobeyama Radio Observatory. *Radio Polarimeters at the Nobeyama Radio Observatory (NRO)*. <http://solar.nro.nao.ac.jp/norp/>.
- [8] SDO. *Best Practices for getting SDO Browse Data*. <https://sdo.gsfc.nasa.gov/data/bestpractice.php>.
- [9] SDOMLv2. *A repository for getting and processing SDO/AIA, SDO/HMI, and SDO/EVE files*. <https://github.com/SDOML/SDOMLv2>.
- [10] WDC-SILSO, Royal Observatory of Belgium, Brussels. *Sunspot Number Data files*. <https://www.sidc.be/SILSO/datafiles>. Accessed [insert date you accessed it]; daily, monthly, and yearly sunspot number series (Version 2.0) available under CC BY-NC license.
- [11] Aideen Fay et al. *Holes in Latent Space: Topological Signatures Under Adversarial Influence*. 2025. arXiv: 2505.20435 [cs.LG]. URL: <https://arxiv.org/abs/2505.20435>.
- [12] Eduardo Paluzo-Hidalgo. *Latent Space Topology Evolution in Multilayer Perceptrons*. 2025. arXiv: 2506.01569 [cs.LG]. URL: <https://arxiv.org/abs/2506.01569>.
- [13] Maximilian Beck et al. *xLSTM: Extended Long Short-Term Memory*. 2024. arXiv: 2405.04517 [cs.LG]. URL: <https://arxiv.org/abs/2405.04517>.
- [14] L. Papa et al. “A survey of efficient vision transformers: algorithms, techniques, and performance benchmarking”. In: *arXiv preprint* (2024). arXiv: 2309.02031 [cs.CV]. URL: <https://arxiv.org/abs/2309.02031>.
- [15] Y. Abdullah et al. “Operational prediction of solar flares using a transformer-based framework”. In: *Scientific Reports* 13 (2023), p. 13665. DOI: 10.1038/s41598-023-40897-7.
- [16] Jiaxiang Dong et al. *SimMTM: A Simple Pre-Training Framework for Masked Time-Series Modeling*. 2023. arXiv: 2302.00861 [cs.LG]. URL: <https://arxiv.org/abs/2302.00861>.
- [17] Qingsong Wen et al. *Transformers in Time Series: A Survey*. 2023. arXiv: 2202.07125 [cs.LG]. URL: <https://arxiv.org/abs/2202.07125>.
- [18] B. Zhuang et al. “A Survey on Efficient Training of Transformers”. In: *arXiv preprint* (2023). arXiv: 2302.01107 [cs.LG]. URL: <https://arxiv.org/abs/2302.01107>.
- [19] X. Zhang et al. “HiViT: Hierarchical Vision Transformer Meets Masked Image Modeling”. In: *arXiv preprint* (2022). arXiv: 2205.14949 [cs.CV]. URL: <https://arxiv.org/abs/2205.14949>.
- [20] H. Bao et al. “BEiT: BERT Pre-Training of Image Transformers”. In: *arXiv preprint* (2021). arXiv: 2106.08254 [cs.CV]. URL: <https://arxiv.org/abs/2106.08254>.
- [21] B. Benson et al. “Simultaneous Multivariate Forecast of Space Weather Indices using Deep Neural Network Ensembles”. In: *arXiv preprint* (2021). arXiv: 2112.09051 [astro-ph.IM]. URL: <https://arxiv.org/abs/2112.09051>.

-
- [22] K. He et al. “Masked Autoencoders Are Scalable Vision Learners”. In: *arXiv preprint* (2021). arXiv: 2111.06377 [cs.CV]. URL: <https://arxiv.org/abs/2111.06377>.
- [23] Y. Liu et al. “Efficient Training of Visual Transformers with Small Datasets”. In: *arXiv preprint* (2021). arXiv: 2106.03746 [cs.CV]. URL: <https://arxiv.org/abs/2106.03746>.
- [24] Z. Liu et al. “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows”. In: *arXiv preprint* (2021). arXiv: 2103.14030 [cs.CV]. URL: <https://arxiv.org/abs/2103.14030>.
- [25] Michael Moor et al. *Topological Autoencoders*. 2021. arXiv: 1906.00722 [cs.LG]. URL: <https://arxiv.org/abs/1906.00722>.
- [26] Matthew Wheeler, Jose Bouza, and Peter Bubenik. “Activation Landscapes as a Topological Summary of Neural Network Performance”. In: *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, Dec. 2021, pp. 3865–3870. DOI: 10.1109/bigdata52589.2021.9671368. URL: <http://dx.doi.org/10.1109/BigData52589.2021.9671368>.
- [27] X. Zhai et al. “Scaling Vision Transformers”. In: *arXiv preprint* (2021). arXiv: 2106.04560 [cs.CV]. URL: <https://arxiv.org/abs/2106.04560>.
- [28] A. Dosovitskiy et al. “An Image is worth 16x16 Words: Transformers for Image Recognition at Scale”. In: *arXiv preprint* (2020). arXiv: 2010.11929 [cs.CV]. URL: <https://arxiv.org/abs/2010.11929>.
- [29] Leland McInnes, John Healy, and James Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. 2020. arXiv: 1802.03426 [stat.ML]. URL: <https://arxiv.org/abs/1802.03426>.
- [30] NOAA National Weather Service. *Glossary of Forecast Verification Metrics*. https://www.weather.gov/media/owp/oh/rfcdev/docs/Glossary_Forecast_Verification_Metrics.pdf. Accessed: 2025-08-28. 2020.
- [31] R. Galvez et al. “A Machine Learning Dataset Prepared From the NASA Solar Dynamics Observatory Mission”. In: *arXiv preprint* (2019). arXiv: 1903.04538 [astro-ph.IM]. URL: <https://arxiv.org/abs/1903.04538>.
- [32] R. Galvez et al. *A Machine Learning Dataset Prepared From the NASA Solar Dynamics Observatory Mission*. <https://arxiv.org/abs/1903.04538>. 2019.
- [33] J. Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *arXiv preprint* (2018). arXiv: 1810.04805 [cs.CL]. URL: <https://arxiv.org/abs/1810.04805>.
- [34] Leland McInnes et al. “UMAP: Uniform Manifold Approximation and Projection”. In: *The Journal of Open Source Software* 3.29 (2018), p. 861.
- [35] A. Vaswani et al. “Attention Is All You Need”. In: *arXiv preprint* (2017). arXiv: 1706.03762 [cs.CL]. URL: <https://arxiv.org/abs/1706.03762>.
- [36] M. Bobra et al. “The Helioseismic and Magnetic Imager (HMI) Vector Magnetic Field Pipeline: SHARP”. In: *arXiv preprint* (2014). arXiv: 1404.1879 [astro-ph.SR]. URL: <https://arxiv.org/abs/1404.1879>.
- [37] S. Hochreiter and J. Schmidhuber. “Long short-term memory”. In: *Neural Computation* 9.8 (1997), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.