Solar flare forecasting with foundational transformer models across image, video, and time-series modalities

S. Riggi[a,*], P. Romano[a], A. Pilzer[b], U. Becciani[a]

[a] *INAF - Osservatorio Astrofisico di Catania, Via Santa Sofia 78, 95123 Catania, Italy*
[b] *NVIDIA AI Technology Center, Italy*

Abstract

We present a comparative study of transformer-based architectures for solar flare forecasting using heterogeneous data modalities, including images, video sequences, and time-series observations. Our analysis evaluates three recent foundational models – *SigLIP2* for image encoding, *VideoMAE* for spatio-temporal video representation, and *Moirai2* for multivariate time-series forecasting – applied to publicly available datasets of solar magnetograms from the SDO/HMI mission and soft X-ray fluxes acquired by GOES satellites. All models are trained and validated under consistent data splits and evaluation criteria, with the goal of assessing the strengths and limitations of transformer backbones across spatial and temporal representations of solar activity. We investigate multiple loss formulations (weighted BCE, focal, and score-oriented) and training balance strategies to mitigate class imbalance typical of flare datasets. Results show that while both SigLIP2 and VideoMAE achieve typical performance on image and video data True Skill Statistic TSS~0.60-0.65), the time-series model Moirai2 reaches superior forecasting skill (TSS~0.74) using irradiance-based temporal evolution alone. These findings highlight the potential of pretrained transformer architectures and cross-modal learning for advancing operational space weather forecasting, paving the way toward unified multimodal models that integrate visual and temporal information.

Keywords: Sun: general, Sun: flares, space weather, event forecasting, deep learning, transformers, foundational models

1. Introduction

Solar flares are among the most energetic manifestations of magnetic energy release in the solar atmosphere. They are powered by magnetic reconnection — the rapid topological reconfiguration of coronal magnetic field lines — which converts magnetic free energy accumulated in active regions into plasma heating, particle acceleration, and broadband electromagnetic emission, from hard X-rays to radio wavelengths (Shibata & Magara, 2011; Benz, 2017). The reconnection process typically occurs in the corona above magnetic polarity inversion lines where strong shear and electric currents develop (Priest & Forbes, 2002; Toriumi & Wang, 2019). The released energy propagates along magnetic loops toward the lower atmosphere, producing impulsive chromospheric brightenings and heating of dense plasma that fills coronal loops through chromospheric evaporation.

This magnetic energy build-up is driven by the continuous emergence, twisting, and shearing of photospheric magnetic flux. Observations and nonlinear force-free field extrapolations have shown that flare-productive active regions are characterized by high magnetic complexity and helicity, strong shear along polarity inversion lines, and the presence of magnetic flux ropes or δ-sunspots (e.g., Romano et al., 2014, 2018, 2024). These magnetic configurations, identifiable through their footprints in photospheric magnetograms, favour the storage of large amounts of free magnetic energy, which can be rapidly released through reconnection once critical thresholds of current density or magnetic stress are exceeded. The onset of a flare is often accompanied by coronal restructuring and, in many cases, by the ejection of magnetized plasma in the form of a coronal mass ejection (CME), whose propagation through the heliosphere can drive interplanetary shocks and strong space-weather disturbances (Zhang et al., 2001; Temmer, 2021).

The impact of these eruptive phenomena on Earth and near-Earth space can affect both technological systems and human health. Major consequences include disruption of radio communications, damage to satellites and spacecraft, increased radiation exposure for astronauts and high-altitude flights, geomagnetically induced currents (GICs) in power grids, and disturbances in navigation and communication systems (Pulkkinen et al., 2017; Kilpua et al., 2017).

Solar flares are commonly categorized by thresholds of the soft X-ray maximum peak flux (MPF) in the 0.1–0.8 nm band, where M-class (10^{-5} < MPF < 10^{-4} W m^{-2}) and X-class (MPF > 10^{-4} W m^{-2}) flares indicate stronger events that are often, but not always, associated with the most geoeffective CMEs and particle storms. Accurate forecasting of solar flares and CMEs is therefore a key objective in space-weather research, enabling mitigation of their potentially adverse effects.

Over the past decades, flare forecasting has evolved through several methodological stages, progressing from empirical and statistical approaches to physics-based modeling and, more

*Corresponding author
Email address: simone.riggi@inaf.it (S. Riggi)

recently, to machine learning techniques. The advent of high–spatial and high–temporal resolution multi-wavelength observations from modern solar observatories – such as NASA's Solar Dynamics Observatory (SDO, Pesnell et al. 2012), ESA/-NASA's Solar Orbiter (Müller et al., 2020), NASA's Parker Solar Probe (PSP, Raouafi et al. 2023), ESA's Proba-3 (Zhukov et al., 2025), the Advanced Space-based Solar Observatory (ASO-S, Gan et al. 2023), ISRO's Aditya-L1 (Tripathi et al., 2022), and NOAA's Geostationary Operational Environmental Satellites (GOES) – has enabled the development of deep learning models capable of forecasting directly from images, sequences, videos, and multimodal data. These approaches minimize the need for manual feature engineering and can potentially uncover complex spatio-temporal patterns associated with solar activity.

In this context, the goal of this work is to explore how recent state-of-the-art transformer-based foundational architectures, designed for different data modalities (images, videos, time series), perform on the solar flare forecasting problem. To support future research and reproducibility within the community, we carried out our analysis on a publicly available dataset and released both the developed software and fine-tuned models as open resources [1,2]. While transformer-based video models have only recently been applied to flare forecasting (Li et al., 2025), this study represents, to the best of our knowledge, one of the first explorations of foundational models for time-series in this domain.

The paper is organized as follows. Section 2 reviews related work, with emphasis on recent advances in deep learning architectures. Section 3 describes the datasets used for all data modalities and the procedures adopted for generating the training, validation, and test splits. Section 4 provides an overview of the foundational models considered and describes how they were adapted to the solar flare forecasting task, along with details of the training and evaluation strategies. The obtained results are presented and discussed in Section 5, and the conclusions and future directions are summarized in Section 6.

2. Related work

Traditional solar flare forecasting has relied on handcrafted physical and empirical predictors, such as magnetic field features, historical flare activity, and statistical models. The recent availability of high-resolution solar imagery and magnetogram data (e.g., from SDO/HMI[3]) has facilitated the adoption of deep learning techniques for this task. In this section, we review the main deep learning approaches proposed to date, organized by data modality, and summarize their forecasting performance in terms of the *True Skill Statistic* (TSS) and the *Heidke Skill Score* (HSS) metrics (see Section 4.2 and Appendix B).

2.1. Image-based approaches

Image-based methods have long represented the most direct approach to solar flare forecasting, exploiting the rich spatial information encoded in photospheric and coronal images. Most studies rely on SDO/HMI line-of-sight (LOS) or vector magnetograms, provided either as full-disk maps or as cropped active region (AR) patches. The general goal is to determine whether an AR will produce a flare above a given class threshold (e.g., ≥C, M, or X) within a specified prediction window, typically 24 hours.

Early work in this domain established the foundation for deep-learning-based flare prediction through convolutional neural networks (CNNs), which demonstrated the capacity to extract meaningful magnetic field features from magnetograms. For instance, Huang et al. (2018) designed a custom CNN trained on LOS magnetograms from SOHO/MDI and SDO/HMI to investigate flare predictability across multiple thresholds and forecast horizons, achieving a TSS of 0.66 and an HSS of 0.14 for M-class and above (M+) forecasting over a 24-hour horizon. Building upon this, Zheng et al. (2019) introduced a hybrid architecture of three binary CNN classifiers, inspired by VGGNet, for multi-class flare prediction. Using SDO/HMI magnetograms, they reported mean TSS scores ranging from 0.53 to 0.55 across flare classes, highlighting the discriminative potential of CNN features for different flare intensity levels.

Subsequent studies began addressing key challenges such as class imbalance and solar-cycle variability. Deng et al. (2021) incorporated a Generative Adversarial Network (GAN) to generate synthetic magnetograms, thereby augmenting the training data and reducing imbalance between flare and non-flare samples. Their hybrid CNN, trained separately for the rising and declining phases of the solar cycle, achieved mean TSS values of 0.65 (C), 0.65 (M), and 0.76 (X), marking a notable improvement over earlier CNN-based methods.

With the success of the attention mechanism in computer vision (Vaswani et al., 2017), subsequent models integrated spatial attention blocks or adopted Vision Transformer (ViT) (Dosovitskiy et al., 2020) backbones to enhance feature extraction. Pandey et al. (2023) introduced an attention-augmented CNN trained directly on full-disk SDO/HMI magnetograms for ≥M1.0-class flare prediction, achieving average scores of TSS = 0.54 ± 0.03 and HSS = 0.37 ± 0.07 – significantly outperforming standard CNNs without attention. In a systematic evaluation, Yan et al. (2024) compared conventional CNNs, attention-enhanced variants (SE, CBAM, and ECA blocks), and a pre-trained ViT Base/16 architecture. Their analysis showed that attention modules did not lead to statistically significant gains over baseline CNNs across all metrics, although the CNN-SE variant yielded the best overall results, with TSS = 0.68 ± 0.13 and HSS = 0.69 ± 0.13 for C+ forecasting, and TSS = 0.56 ± 0.10 and HSS = 0.32 ± 0.07 for M+ forecasting.

The exploration of image modalities beyond magnetograms has also expanded the scope of flare forecasting. Sun et al. (2023) applied CNN-based architectures (ResNet18, AlexNet, SqueezeNet) to SDO/AIA extreme ultraviolet (EUV) images at six wavelengths (94, 131, 171, 193, 211, and 335 Å), finding that the 94 Å channel provided the strongest single-wavelength performance (TSS = 0.92), while multi-wavelength fusion achieved TSS ∼ 0.75. More recently, Francisco et al. (2025) proposed a patch-distributed CNN (P-CNN, based on EfficientNetV2-S)

[1] https://github.com/inaf-oact-ai/solar-flare-forecaster/
[2] https://huggingface.co/inaf-oact-ai
[3] Helioseismic and Magnetic Imager

to forecast C+ and M+ flares from full-disk SDO/HMI and SDO/AIA images. Their EUV-based models achieved slightly higher scores (TSS = 0.61, HSS = 0.42) than the magnetogram-based ones, while the regional patch configuration (TSS = 0.43, HSS = 0.22) demonstrated potential for localizing flare-productive regions without explicit AR labeling.

Other studies, such as Legnaro et al. (2025), explored the classification of sunspot magnetic type in solar active-region cutouts using ResNet18 and DeiT (Data-Efficient Image Transformer) architectures, demonstrating that combining magnetogram and continuum images can significantly enhance model performance.

Together, these studies outline a steady evolution of image-based flare forecasting – from early CNN models to hybrid and transformer-augmented architectures – and reveal both the promise and the limits of spatial-only approaches. Despite meaningful progress, performance remains sensitive to dataset balance, magnetic complexity representation, and the inclusion (or lack) of temporal and multi-wavelength information, motivating the shift toward video-based and multimodal forecasting strategies.

2.2. *Spatio-temporal and video-based forecasting approaches*

Building upon the advances of image-based flare forecasting, recent research has shifted toward modeling the AR temporal evolution through spatio-temporal or video-based approaches. These methods aim to capture how magnetic structures evolve and accumulate energy prior to flare onset, integrating both spatial morphology and temporal dynamics. By combining convolutional and sequential processing – through CNN–recurrent network hybrids, 3D convolutional networks, or transformer architectures – they attempt to represent the buildup of magnetic complexity that static models cannot fully resolve.

An early demonstration of this paradigm was provided by Guastavino et al. (2022), who applied a Long-term Recurrent Convolutional Network (LRCN) to 24-hour SDO/HMI magnetogram sequences (40 frames at a 36-minute cadence). The model combined CNN-based spatial encoders with an Long short-term memory (LSTM) module to learn temporal dependencies, achieving mean TSS scores of 0.55 ± 0.05 and 0.68 ± 0.09 for C+ and M+ forecasting, respectively.

Sun et al. (2022) employed 3D CNNs to model 24-hour SHARP magnetogram sequences (15 frames at a 96-minute cadence) for C+ and M+ flare prediction within the next 24 hours. To address class imbalance, they adopted a focal loss function and systematically explored the impact of the focal parameter γ. Their best configuration, with $\gamma = 2$, achieved TSS = 0.826 and HSS = 0.667 for M+ forecasting.

Subsequent efforts explored architectures capable of capturing more complex temporal correlations and multi-wavelength context. Francisco et al. (2024) proposed VideoLENS, an architecture integrating a C3D block with multiple 3D convolutional layers and a Local-TimeSeries module combining multi-head attention and LSTM layers. Trained on 24-hour sequences (13 frames at a 2-hour cadence) of full-disk SDO/AIA EUV images at 193, 211, and 94 Å, VideoLENS achieved TSS = 0.65 ± 0.01 and HSS = 0.50 ± 0.01 for M+ forecasting – an improvement of about 5% relative to single-channel models. Their results established that incorporating temporal evolution, even over relatively short windows, can enhance predictive performance compared to static image classifiers.

Further refining temporal modeling, Xu et al. (2025) developed a hybrid CNN–TCN framework that combined spatial encoding with Temporal Convolutional Networks (TCN) to model the progressive evolution of SDO/HMI magnetograms, supplemented by magnetic field parameters. Their model achieved TSS = 0.85 ± 0.07 and HSS = 0.75 ± 0.10 for 24-hour C+ and M+ predictions – among the highest scores reported to date – demonstrating the effectiveness of TCNs in representing medium-term dependencies in solar magnetic activity.

The most recent advances introduced transformer-based video models capable of learning long-range spatio-temporal dependencies. Li et al. (2025) employed a Multiscale Video Transformer (MViT) trained on 24-hour magnetogram sequences (40 frames at 36-minute cadence) from SDO/SHARP, SDO/HMI, and FMG/ASO-S instruments. Their analysis compared single- and multi-AR samples, showing that transformer models consistently outperformed CNN baselines. Performance was notably higher for single-AR sequences (TSS = 0.74 ± 0.06, HSS = 0.61 ± 0.12) compared with mixed ones (TSS = 0.69 ± 0.05, HSS = 0.58 ± 0.08), underscoring the benefits of isolating individual ARs during training.

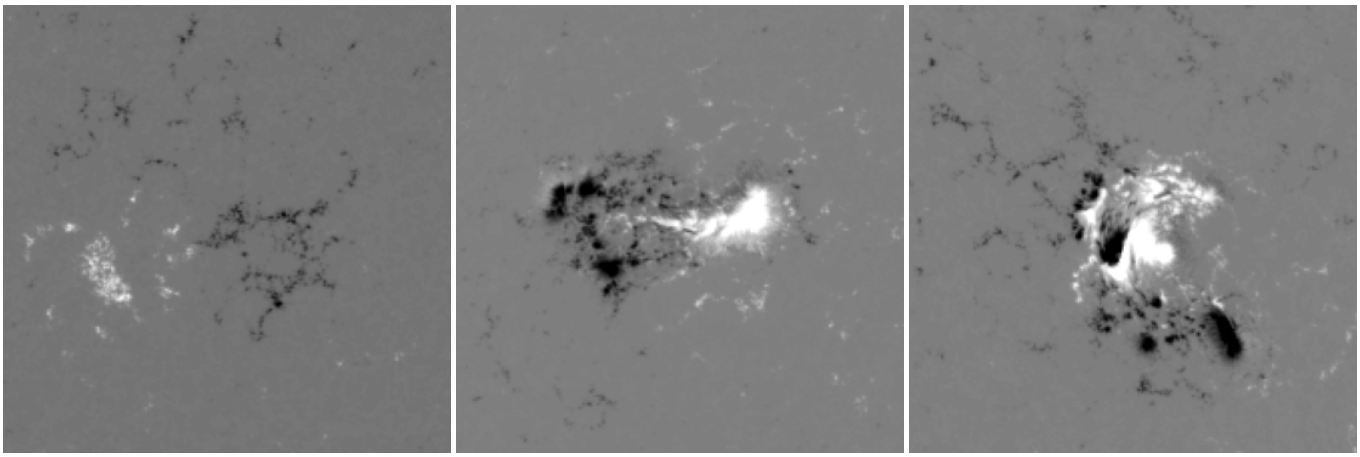
Grim & Gradvohl (2024) further extended transformer-based approaches by applying Improved Multiscale Vision Transformers (MViTv2) to 24-hour SDO/HMI magnetogram sequences (16 frames at a 96-minute cadence) for 24- and 48-hour flare forecasting. Using a weighted loss and oversampling to mitigate class imbalance, their best M+ model achieved TSS = 0.70 ± 0.04 and HSS = 0.25 ± 0.01 for the 24-hour horizon.

Collectively, these studies mark a clear progression from static image classifiers to architectures capable of capturing the dynamic evolution of solar magnetic fields. While performance gains have been moderate compared to single-image models, spatio-temporal approaches provide a more physically grounded framework for flare forecasting and set the stage for multimodal models that jointly exploit temporal, spectral, and contextual information from multiple instruments.

2.3. *Time series-based approaches*

Parallel to the development of image and video models, another major line of research has focused on forecasting solar flares directly from time series of magnetic and radiative parameters that characterize ARs. Rather than modeling spatial morphology, these approaches capture the temporal evolution of key physical quantities, such as those derived from the SHARP or HMI datasets, to identify dynamical precursors of flare activity. Deep learning methods including recurrent networks (RNNs, LSTMs, GRUs), one-dimensional convolutional networks (1D-CNNs), and more recently, transformer-based architectures have been employed to uncover such temporal dependencies and correlations.

Early contributions demonstrated the promise of recurrent architectures for this task. Chen et al. (2019) trained a two-layer LSTM model on multivariate time series of 20 SHARP



(a) No-Flare *(AR=2563, t=20160714 00:00 UT)* (b) C-class *(AR=2396, t=20120806 19:00 UT)* (c) X-class *(AR=2673, t=20170905 06:00 UT)*

Figure 1: Sample of HMI magnetograms from the image dataset, illustrating examples of the NONE (left), C-class (middle), and X-class (right) categories. The NONE example corresponds to an AR with no flares stronger than the C class occurring within the following 24 hours. For each image, the associated AR identifier and observation times are reported below.

parameters sampled at 1-hour cadence, predicting flares across multiple thresholds and forecast horizons ranging from 1 to 72 hours. Their model maintained stable performance for short-term forecasts but exhibited a gradual decline for longer lead times, achieving TSS = 0.68 and HSS = 0.68 for 24-hour M+ forecasting. This work highlighted both the effectiveness of LSTMs in modeling flare-related temporal dynamics and the inherent difficulty of long-horizon prediction.

Building on this foundation, Abduallah (2023) introduced *SolarFlareNet*, a more sophisticated hybrid model that integrates 1D-CNN, LSTM, and transformer encoder blocks to capture both short- and long-term dependencies in the SHARP parameters. Trained on 108-minute sequences of nine magnetic parameters sampled at 12-minute cadence, the model forecasted flares at various intensity thresholds (≥C, ≥M, ≥M5.0) and horizons (24, 48, and 72 hours). Using the start of each sequence as the reference time for labeling,[4] their best model achieved a mean TSS of 0.84 ± 0.03 for 24-hour M+ forecasting − one of the highest values reported for time-series methods to date.

More recently, Donahue & Inceoglu (2024) applied a pure transformer architecture to 24-hour and 48-hour sequences of 18 SHARP parameters, directly leveraging the transformer's ability to model long-range temporal dependencies without recurrent connections. For a 24-hour horizon, their M+ forecasting model reached TSS = 0.66 and HSS = 0.16, confirming the growing potential of transformer-based encoders for time-domain flare prediction.

For completeness, we also note several recent studies that have focused on high-cadence time series of soft X-ray fluxes (XRS) recorded by the GOES. Unlike the previously discussed works, which frame flare prediction as a classification problem (i.e., predicting the flare class), these analyses approach the task as a multi-step regression problem, aiming to forecast the future evolution of the XRS flux itself or related observables over specific time horizons.

For instance, Dei & Fuentes (2020) compared three neural architectures − 1D CNN, LSTM, and N-BEATS − for forecasting X-ray fluxes at lead times ranging from 1 to 5 hours, using historical windows of 1, 2, and 3 times the prediction interval. Among these, the N-BEATS model consistently achieved the lowest mean squared error (MSE∼0.001 for the largest prediction window), outperforming both baseline estimators and CNN/LSTM networks.

More recently, van der Sande & Muñoz-Jaramillo (2025) evaluated a set of traditional machine learning and deep learning models (Linear, MLP, Random Forest, and CNN) to predict XRS fluxes at horizons up to ∼48 h. Their models used a diverse set of input sources, including magnetograms, EUV images at multiple wavelengths, and scalar magnetic features. Forecasting accuracy was found to improve with longer prediction windows, with their best CNN model reaching MSE ∼ 0.3 after 24 h. From the predicted flux values, they also derived binary classification metrics for ≥M1-class events, obtaining TSS and HSS values of approximately 0.6 for forecasting windows beyond 24 h.

3. Dataset

In this work, we employ three distinct data modalities as inputs to the flare forecasting models: images, videos, and time series. Table 1 summarizes the number of available samples and subsets for each modality. The following sections provide detailed descriptions of the original observations and the data preparation workflow, including event annotation procedures and the generation of training, validation, and test splits.

3.1. Image dataset

The image dataset used in this work was compiled by Boucheron et al. (2023) from full-disk LOS solar photospheric

[4]This differs from our approach, in which the last time step defines the forecast reference.

Table 1: Number of samples per solar flare class present in each dataset used in this work.

Data	Sample	All	No-Flare		C-class		M-class		X-class		NONE		M+	
			N	f(%)	N	f(%)	N	f(%)	N	f(%)	N	f(%)	N	f(%)
HMI	all	950,047	759,465	79.9	161,059	17.0	26,680	2.8	2,843	0.3	920,524	96.9	29,523	3.1
images	train	759,357	610,108	80.3	125,854	16.6	21,141	2.8	2,254	0.3	735,962	96.9	23,395	3.1
	train_ds	299,249	150,000	50.1	125,854	42.1	21,141	7.1	2,254	0.8	275,854	92.2	23,395	7.8
	train_ds_bal	48,395	12,500	25.8	12,500	25.8	21,141	43.7	2,254	4.7	25,000	51.7	23,395	48.3
	cv	95,933	76,142	79.4	16,975	17.7	2,578	2.7	238	0.2	93,117	97.1	2,816	2.9
	cv_ds_bal	5,816	1,500	25.8	1,500	25.8	2,578	44.3	238	4.1	3,000	51.6	2,816	XXX
	test	94,757	73,215	77.2	18,230	19.2	2,961	3.1	351	0.4	91,445	96.5	3,312	3.5
HMI	all	631,366	503,293	79.7	109,062	17.3	17,212	2.7	1,799	0.3	612,355	97.0	19,011	3.0
videos	train	504,526	404,790	80.2	84,365	16.7	13,979	2.8	1,392	0.3	489,155	97.0	15,371	3.0
(Δt=72 min)	train_ds	199,736	100,000	50.1	84,365	42.2	13,979	7.0	1,392	0.7	184,365	92.3	15,371	7.7
	train_ds_bal	30,771	7,700	25.0	7,700	25.0	13,979	45.4	1,392	4.5	15,400	50.0	15,371	50.0
	cv	62,730	48,714	77.7	12,094	19.3	1,773	2.8	149	0.2	60,808	96.9	1,922	3.1
	cv_ds_bal	3,922	1,000	25.5	1,000	25.5	1,773	45.2	149	3.8	2,000	51.0	1,922	49.0
	test	64,110	49,789	77.7	12,603	19.7	1,460	2.3	258	0.4	62,392	97.3	1,718	2.7
HMI	all	774,687	617,791	79.7	133,028	17.2	21,597	2.8	2,271	0.3	750,819	96.9	23,868	3.1
videos	train	618,864	496,211	80.2	103,501	16.7	17,380	2.8	1,772	0.3	599,712	96.9	19,152	3.1
(Δt=36 min)	train_ds	222,653	100,000	44.9	103,501	46.5	17,380	7.8	1,772	0.8	203,501	91.4	19,152	8.6
	train_ds_bal	38,352	9,600	25.0	9,600	25.5	17,380	45.3	1,772	4.6	19,200	50.1	19,152	49.9
	cv	77,924	61,122	78.4	14,450	18.5	2,159	2.8	193	0.2	75,572	97.0	2,352	3.0
	cv_ds_bal	4,752	1,200	25.3	1,200	25.3	2,159	45.4	193	4.1	2,400	50.5	2,352	49.5
	test	77,899	60,458	77.6	15,077	19.4	2,058	2.6	306	0.4	75,535	97.0	2,364	3.0
XRS	all	20,755	10,307	49.7	6,964	33.6	3,081	14,8	403	1.9	17,271	83.2	3,484	16.8
series	train	14,529	7,215	49.7	4,875	33.6	2,157	14.8	282	1.9	12,090	83.2	2,439	16.8
(T=24 h)	cv	2,076	1,031	49.7	697	33.6	308	14.8	40	1.9	1,728	83.2	348	16.8
	test	4,150	2,061	49.7	1,392	33.6	616	14.8	81	1.9	3,453	83.2	697	16.8
XRS	all	41,517	23,583	56.8	13,283	32.0	4,207	10.1	444	1.1	36,866	88.8	4,651	11.2
series	train	29,062	16,508	56.8	9,298	32.0	2,945	10.1	311	1.1	25,806	88.8	3,256	11.2
(T=12 h)	cv	4,151	2,358	56.8	1,328	32.0	421	10.1	44	1.1	3,686	88.8	465	11.2
	test	8,303	4,716	56.8	2,657	32.0	841	10.1	89	1.1	7,373	88.8	930	11.2

magnetograms acquired by the Helioseismic and Magnetic Imager (HMI) aboard NASA's SDO between 1 May 2010 and 31 December 2018, with a cadence of 720 s. The original full-disk HMI observations were cropped around the heliographic position (within ±60° in latitude and longitude) of 1,570 ARs reported by the National Oceanic and Atmospheric Administration (NOAA), yielding 600×600-pixel cutouts (approximately 300"×300"), large enough to encompass the most extended ARs. The final dataset comprises 950,047 magnetogram images, each centered on a NOAA-identified AR.

For this study, we adopted the lower-resolution version of the dataset, where the full-resolution images were resized to 224×224 pixels and preprocessed following the pipeline described in Boucheron et al. (2023):

- offset pixel intensities I by 2550, i.e., $I \rightarrow I + 2550$;
- clip intensities to the range [0, 5100];
- normalize values to [0, 255];
- downsample intensities to 8-bit depth (uint8).

Image labeling was performed using the Space Weather Prediction Center (SWPC) Event Reports, which list solar flares detected by the GOES and their associated NOAA ARs. Each HMI magnetogram was assigned a categorical label indicating whether a flare of class ≥C1.0 occurred within the subsequent 24 hours relative to the image timestamp. Specifically, images were labeled as No-Flare if no flares above the C class occurred, or otherwise by the major flare class (C, M, or X), disregarding subtypes. Representative examples of HMI images labeled as No-Flare, C, and X are shown in Fig. 1.

The total number of images per flare class is reported in Table 1 (first row). Following Boucheron et al. (2023), the dataset includes predefined training, validation (cv), and test subsets, whose class distributions are listed in rows 2, 5, and 7 of the same table. The dataset is markedly imbalanced, dominated by No-Flare samples (80%), followed by C-class (17–19%), M-class (3%), and the rare X-class (0.3%). We additionally report aggregated class-grouped versions labeled as NONE (combining No-Flare and C-class samples) and M+ (combining M- and X-class samples).

To mitigate this imbalance, we created additional downsampled datasets by randomly reducing the number of No-Flare and C-class images while retaining at least one image per AR (so ARs with more images were more aggressively downsampled). The target number of images for each downsampled

class was set approximately equal to the combined number of M+ samples. The resulting datasets, denoted with a "_ds" suffix (e.g., *train_ds*), and their class abundances are reported in Table 1. More strongly balanced versions, labeled "_ds_bal" (e.g., *train_ds_bal*), were produced by further downsampling No-Flare and C-class data to achieve a nearly 1:1 ratio between positive and negative examples. Unless otherwise specified, only the downsampled datasets were used for training the image-based forecasting models.

Finally, to assess statistical variability, we generated five-fold (K=5) cross-validation splits from the original dataset by randomly reassigning all images from each AR to a single split (ensuring that no AR appears in multiple folds). The overall class distributions were preserved to within ∼1% of the original dataset on average, with a standard deviation across folds below 2%. These K-fold splits were used to estimate the uncertainty of forecasting metrics reported for the image-based models.

3.2. Video dataset

The video dataset was constructed from the image data described above by grouping consecutive frames into sequences of N_{frames} images, each separated by a temporal interval of Δt minutes. We adopted $N_{\text{frames}} = 16$, matching the standard input length used by most pretrained video models such as VideoMAE (see Section 4.1.2). Two temporal sampling cadences were considered, $\Delta t = 36$ and 72 minutes, corresponding to total sequence durations T of approximately 9.6 hours and 19.2 hours, respectively. Each video sample was assigned the label of its last frame, representing the flare class (or absence thereof) associated with the final timestamp in the sequence.

The number of video samples available for each flare class is reported in Table 1. Following the same strategy adopted for the image data, we also generated a downsampled training subset to mitigate class imbalance, which was subsequently used for model training.

3.3. Time series dataset

The time series dataset was constructed from historical solar irradiance measurements collected by the GOES[5,6], covering nearly three decades of solar activity from 1995 onward. Specifically, we used the science-quality Level 2 (L2) series of 1-minute averaged soft X-ray fluxes in the long-wavelength band (0.1–0.8 nm; XRS-B channel). For each GOES satellite (08-18), the science fluxes were normalized by their corresponding daily background flux level (reported by GOES), producing a flux-ratio time series (science/background). These normalized series were then segmented into non overlapping 12h and 24h intervals (respectively containing 720 and 1440 entries with 1-minute cadence), each representing a distinct observation sample for forecasting. Samples with more than 70% missing data were discarded, while shorter data gaps were filled by linear interpolation. Finally, a logarithmic transformation was applied to the flux ratios.

We enriched the time series dataset using historical flare event information from the NOAA SWPC Event Reports[7], which provide start and end times, flare class, and associated active region identifiers (when available). For each XRS flux-ratio series, we generated a corresponding binary sequence with the same temporal cadence (1-minute), where each time step was assigned a value of 1 if a flare of any class occurred within that minute, and 0 otherwise.

For each selected interval (T=12 h, 24 h), this process yielded two-channel time series inputs for the forecasting models: (i) the normalized XRS flux ratio (*xrs-flux-ratio*), and (ii) the corresponding binary flare occurrence history (*flux-hist*). The same historical flare catalog was also used to define the forecasting labels for each time series, following the same scheme adopted for the image and video datasets—indicating whether a C-, M-, or X-class flare occurred within the 24-hour period following the last recorded time step. An example of an *xrs-flux-ratio* observation labeled as M-class is shown in Fig. 2, where the corresponding historical flare events are overplotted in blue.

4. Method

4.1. Forecasting models

Our forecasting framework is implemented using the `transformers`[8] and `PyTorch`[9] libraries and supports three distinct data modalities: images, videos, and time series. As illustrated in Fig. 3, each modality is processed by a dedicated backbone model designed to extract domain-relevant representations from its respective input type:

- *Image-based model*: Individual solar magnetogram images are processed by a vision encoder (e.g., transformer-based architecture) that extracts spatial features describing the magnetic field morphology of active regions.
- *Video-based model*: Sequences of magnetogram frames are processed through a video backbone that captures both spatial and temporal dependencies, enabling the encoder to represent the dynamic evolution of active regions over time.
- *Time-series model*: Solar irradiance and related measurements (e.g., X-ray flux or flux ratios) are modeled using transformer-based sequence encoders that learn temporal dependencies and compress them into latent representations.

For all modalities, the resulting feature vector—encoding spatial, spatio-temporal, or purely temporal patterns—is passed to a fully connected classification head that predicts the probability

[5] https://www.ncei.noaa.gov/products/goes-r-extreme-ultraviolet-xray-irradiance
[6] https://www.ncei.noaa.gov/products/goes-1-15/space-weather-instruments
[7] ftp://ftp.swpc.noaa.gov/pub/warehouse/
[8] https://github.com/huggingface/transformers
[9] https://github.com/pytorch/pytorch

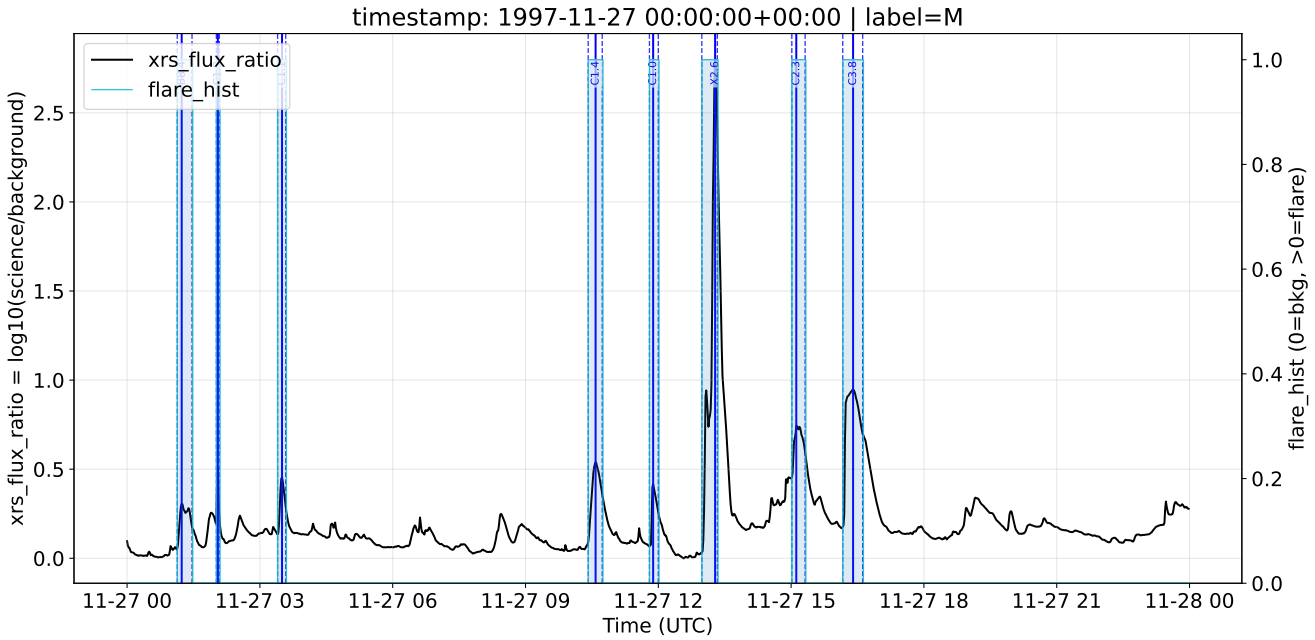


Figure 2: Sample XRS flux-ratio time series (black) spanning a 24-hour interval and labeled as M-class. Historical flare events occurring within the same interval are overplotted in blue.

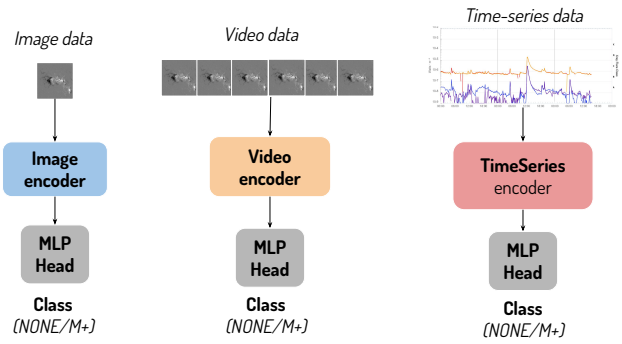


Figure 3: Overview of the three modality-specific solar flare forecasters. Each model processes a different data modality: (left) magnetogram images through a SigLIP2 vision transformer, (center) short sequences of magnetogram frames through a VideoMAE spatio-temporal transformer, and (right) soft X-ray flux time series from GOES through a Moirai2 temporal transformer. All models have the same classification head architecture for consistent comparison of forecasting skill across modalities.

of flare occurrence within a 24-hour forecasting window. The classification head consists of a linear layer producing a single output logit for binary forecasting tasks (e.g., `NONE` vs. `M+` or `NONE` vs. `C+`) or K output logits for single- or multi-label K-class forecasting (e.g., `NONE`, `C`, `M`, `X`).

4.1.1. Image model

For the image-based forecasting task we employ *SigLIP2* (Tschannen et al., 2025), a recent family of vision–language encoders that builds upon the original *SigLIP* architecture (Zhai et al., 2023). SigLIP builds on the *Contrastive Language–Image Pretraining* (CLIP) model (Radford et al., 2021), replacing the original contrastive objective with a sigmoid loss, which improves alignment between visual and textual modalities during pretraining. This formulation improves training stability and yields stronger alignment between modalities. SigLIP2 further extends this approach by integrating captioning-based pretraining, self-distillation, masked prediction, and active data curation into a unified training recipe. These additions lead to more semantically rich and spatially precise visual representations, which are especially valuable for tasks such as solar flare forecasting.

The vision encoder of SigLIP2 follows the ViT architecture, partitioning an image into non-overlapping patches, linearly projecting them into embeddings, and modeling global dependencies through self-attention layers. In SigLIP2, training is carried out on the large-scale WebLI dataset (Chen et al., 2023) comprising approximately 10 billion images and 12 billion associated texts across 109 languages. This massive multimodal pretraining endows the encoder with robust and transferable visual features. Although the language encoder is not used in our experiments, the visual backbone benefits from this large-scale multimodal supervision.

To adapt SigLIP2 for flare classification, we discard the multimodal projection layers and attach a task-specific classification head on top of the pooled visual representation. The head consists of a linear mapping that outputs logits over the flare classes of interest. During fine-tuning, we experiment with both frozen and trainable vision backbones, enabling us to evaluate how much domain-specific specialization benefits solar forecasting.

Practically, our implementation relies on the Hugging Face `AutoModelForImageClassification` interface, which allows us not only to load SigLIP2 checkpoints but also to

seamlessly substitute alternative pretrained vision encoders that support the same API. This design ensures flexibility and extensibility for future comparisons across image backbones.

4.1.2. Video model

For video-based forecasting we considered *VideoMAE*[10] model (Tong et al., 2022), a transformer-based architecture specifically designed for spatio-temporal representation learning in videos. VideoMAE extends the masked autoencoding paradigm from images to videos by introducing *tube masking*, where spatio-temporal cubes of a video are randomly masked at an extremely high ratio (90–95%). The model is trained in a self-supervised manner to reconstruct the missing patches from the remaining visible tokens. This strategy forces the encoder to capture high-level structures and motion dynamics rather than relying on pixel-level redundancy, making it particularly suitable for solar magnetogram sequences in which the temporal evolution of active regions is essential.

The encoder of VideoMAE is based on a ViT architecture that jointly models space–time dependencies. Input video clips are divided into spatio-temporal tokens (cubes), which are embedded and processed through multi-head self-attention layers. This allows the model to simultaneously capture fine-grained spatial patterns (magnetic field morphology) and their temporal evolution across frames. During pretraining, only the encoder processes the visible tokens, while a lightweight decoder reconstructs the masked tokens; at fine-tuning, the decoder is discarded and the encoder serves as a general-purpose video backbone.

A crucial aspect of VideoMAE's success lies in its ability to exploit large-scale pretraining datasets. The original version demonstrated that competitive video representations could be obtained from relatively small datasets such as Kinetics-400 (~240k videos) (Kai et al., 2017) or Something-Something V2 (~170k videos) (Goyal et al., 2017). VideoMAE v2 pushed this further by introducing large-scale pretraining on more than 1.3 million unlabeled video clips collected from diverse sources (YouTube, movies, web videos, and Instagram), complemented by a hybrid labeled dataset (~660k clips) for post-pretraining. This scaling was key to supporting billion-parameter video transformers and improving transferability across downstream tasks.

Recent improvements introduced in VideoMAE v2 (Wang et al., 2023) also make the approach more efficient and scalable. First, a *dual masking strategy* was proposed, extending masking to the decoder in addition to the encoder. This reduces the number of tokens processed during reconstruction and nearly halves memory and computation costs while preserving accuracy. Second, VideoMAE v2 demonstrated that scaling to billion-parameter video transformers becomes feasible when coupled with large and diverse multi-source datasets (over one million clips), and a *progressive training pipeline* that includes both unsupervised pretraining and supervised post-pretraining on hybrid datasets. Together, these innovations improve the generalization of the learned representations and allow the backbone to adapt more effectively to downstream tasks.

In our solar flare forecasting setup, we leverage the pretrained VideoMAE encoder and replace the reconstruction decoder with a classification head. Specifically, the pooled encoder representation of a magnetogram video is connected to a linear layer producing logits over the target flare classes. The model is then fine-tuned on labeled solar video sequences. This enables the backbone to adapt its spatio-temporal representations to the forecasting task, exploiting both morphological features of active regions and their dynamic evolution over time.

4.1.3. Time series model

For time-series forecasting we employ *Moirai*[11], a recently proposed foundation model for universal time-series forecasting (Woo et al., 2024). Moirai is built on a Transformer encoder architecture but introduces several innovations that make it particularly effective for heterogeneous, long, and irregularly sampled sequences. Its design combines (i) *multi patch-size projections*, which enable efficient handling of both high- and low-frequency data, (ii) *Any-variate Attention*, which flattens multivariate inputs and encodes both time and variate axes so that the model can flexibly process any number of input channels, and (iii) *mixture distribution heads*, which allow probabilistic forecasting with flexible distributions beyond simple Gaussian assumptions.

Moirai was pretrained on the *Large-scale Open Time Series Archive* (LOTSA) (Woo et al., 2024), a corpus of over 27 billion observations across nine domains (energy, transport, climate, sales, finance, healthcare, etc.), making it one of the largest open pretraining efforts for time series. This large-scale pretraining enables the model to capture universal temporal patterns and transfer them to downstream tasks.

For our solar flare forecasting task, we adapt the pretrained Moirai encoder to process solar irradiance measurements (GOES XRS flux ratios and flare history). In our code, both the *xrs_flux_ratio* and *flare_hist* are treated as *variates* of the multivariate time series. We modified the `DataLoader` to handle our domain-specific JSON metadata format, and adjusted the `forward` method to support classification rather than probabilistic regression: instead of forecasting future values, the encoder outputs latent representations that are passed through a lightweight classification head (a linear projection into flare classes). The patching procedure follows Moirai's standard approach: given an input of shape $[2, 1440]$ (two variates, each with 1440 time steps), a patch size of 16 divides each variate into $1440/16 = 90$ patches, resulting in $2 \times 90 = 180$ patch tokens. Each token corresponds to a contiguous window of 16 samples from a *single* variate. These tokens are embedded and processed jointly by the Moirai encoder, then pooled and passed to a lightweight classification head. Finally, we integrated these components into our training pipeline, adding various training losses and solar-specific evaluation metrics (see Section 4.2).

By leveraging both its pretrained temporal representations from LOTSA and our tailored adaptations for solar physics

[10] Video Masked Autoencoder, https://github.com/MCG-NJU/VideoMAE

[11] Masked EncOder-based UnIveRsAl TIme Series Forecasting Transformer, https://github.com/SalesforceAIResearch/uni2ts

Table 2: Summary of hyperparameter values used for training our forecaster models.

Parameter	Models		
	SigLIP2	VideoMAE	Moirai2
Input data size	[3,224,224]	[16,3,224,224]	[1440,2]
Preprocessing	RandomCenterCrop(0.65,1.0)	RandomCenterCrop(0.65,1.0)	Log stretch xrs_flux_ratio
	Resize(224×224)	Resize(224×224)	
	RandomFlip	RandomFlip	
	RandomRotate90	RandomRotate90	
Model base	`google/siglip2-base-patch16-224`	`MCG-NJU/videomae-base`	`Salesforce/moirai-2.0-R-small`
Frozen encoder layers	10	10	None
#GPUs	4	4	1
Epoch	10	10	100
Batch size	64	16	64
Gradient accumulation	16	16	16
Optimizer	AdamW	AdamW	AdamW
Learning rate	1e-4	1e-4	1e-4
LR scheduler	cosine	cosine	cosine
Warmup ratio	0.3	0.3	0.1

data, Moirai provides a powerful backbone for flare forecasting, combining universal sequence modeling capabilities with domain-specific fine-tuning.

4.2. Model training and evaluation

In this work, we restricted our analysis to the prediction of solar flares with the greatest potential impact on Earth, focusing on the binary classification of `M+` flares (including both `M`- and `X`-class events) against the background class `NONE` (including observations with no flares or flares weaker than `M`-class). We trained multiple models, systematically varying the input data processing schemes and model hyperparameters to identify the optimal configuration for the forecasting task. The main hyperparameters adopted for all models are summarized in Table 2.

All training experiments were conducted on the CINECA LEONARDO Booster supercomputing infrastructure[12], using a single compute node equipped with 4 GPUs (NVIDIA A100, 64 GB) and 1 CPU (Intel Xeon Platinum 8358 CPUs, 2.60 GHz), with 512 GB of RAM. The following subsection outlines the details of the adopted training strategy.

Input data and preprocessing. The analysis was performed using the downsampled datasets described in Section 3, since preliminary experiments showed no significant difference in the prediction performance of the negative class when compared with the full original datasets. Image and video frame data were normalized to the range [0, 1]. For the time series data, a logarithmic transformation was applied only to the *xrs_flux_ratio* channel, while the *flare_hist* channel was left unchanged.

Data augmentation. To enhance model generalization, we applied random data augmentation to the image and video datasets during training. Since the original dataset was constructed from full-disk images using a fixed cutout size (600×600 pixels), rather than dynamically adapting the crop to the size of each AR, some solar structures unrelated to ARs or flare activity may be present. To mitigate this, we introduced a `RandomCenterCrop` transform that crops input images using a random crop fraction uniformly sampled in the range [0.65, 1.0]. The cropped images were then resized to match the input resolution required by the pretrained backbone (e.g., 224×224 pixels). Additional augmentations included random horizontal and vertical flips, as well as random 90° rotations, each applied independently with equal probability. For validation and test data, only normalization and resizing transforms were applied to ensure consistency. The same augmentation pipeline was used for video data, with all transforms applied identically across all frames within each video sequence. No augmentation was applied to the time series data.

Training losses. Because the solar flare datasets are highly imbalanced toward negative (non-flaring) examples, we implemented a custom Hugging Face (HF) trainer capable of rebalancing the training data within each batch.[13] The trainer supports three alternative class-weighted loss functions to mitigate class imbalance: the class-weighted binary cross-entropy loss ($\mathcal{L}_{\text{bce}}$), the class-weighted focal loss ($\mathcal{L}_{\text{focal}}$), and the score-oriented loss (SOL; $\mathcal{L}_{\text{sol}}$) introduced by Marchetti et al. (2022) and Guastavino et al. (2023). A detailed mathematical description of these loss functions is provided in Appendix A.

Model selection and evaluation. For each sample $i = 1, \ldots, N$ with true target $y_i \in \{0, 1\}$ (with $y_i = 1$ denotes the occurrence of a flare), the model outputs a probability $\hat{p}_i \in [0, 1]$[14] representing

[12] https://www.hpc.cineca.it/systems/hardware/leonardo/

[13] When the number of negative samples greatly exceeds that of positive samples and the batch size is small, there is a high probability that a batch will be dominated – or entirely composed – of negative examples.

[14] Model outputs for trained binary forecasters are single logits, converted to probabilities through a sigmoid activation function.

Table 3: Forecasting metrics for M+ flare prediction obtained by fine-tuned SigLIP2 models on the HMI image validation set from Boucheron et al. (2023) (labeled as 'cv' in Table 1), using different training configurations and loss functions (binary cross-entropy, focal, and score-oriented TSS loss). Each metric is reported for two decision thresholds: $\tau = 0.5$ and $\tau = \tau^*$, where τ^* is estimated from the full TSS–τ curves across multiple training and evaluation runs ($N = 5$) as the threshold that maximizes the mean TSS, subject to minimum recall (0.5) and precision (0.15) constraints. For both thresholds, the table reports the mean and standard deviation of each metric computed over the $N = 5$ runs.

Model ID	Train Set	Val Set	Sampler	Loss	TSS		HSS		MCC	
					$\tau = 0.5$	$\tau = \tau^*$	$\tau = 0.5$	$\tau = \tau^*$	$\tau = 0.5$	$\tau = \tau^*$
siglip2_v1	train_ds	cv	yes	sol	0.680 ± 0.037	0.674 ± 0.037	0.211 ± 0.032	0.213 ± 0.031	0.310 ± 0.024	0.310 ± 0.022
siglip2_v2	train_ds	cv	no	focal_w	0.679 ± 0.022	0.656 ± 0.029	0.185 ± 0.021	0.213 ± 0.028	0.292 ± 0.018	0.306 ± 0.022
siglip2_v3	train_ds	cv	no	bce_w	0.672 ± 0.019	0.620 ± 0.042	0.175 ± 0.031	0.210 ± 0.026	0.283 ± 0.020	0.295 ± 0.009
siglip2_v4	train_ds_bal	cv_ds_bal	no	sol	**0.695 ± 0.025**	0.685 ± 0.025	0.207 ± 0.029	0.214 ± 0.031	0.311 ± 0.022	0.314 ± 0.024
siglip2_v5	train_ds_bal	cv_ds_bal	no	focal	0.659 ± 0.055	0.640 ± 0.074	0.200 ± 0.037	0.211 ± 0.036	0.298 ± 0.023	0.301 ± 0.026
siglip2_v6	train_ds_bal	cv_ds_bal	no	bce	0.682 ± 0.037	**0.701 ± 0.028**	**0.239 ± 0.034**	**0.216 ± 0.036**	**0.330 ± 0.019**	**0.318 ± 0.021**

Table 4: Forecasting metrics for M+ flare prediction obtained by fine-tuned VideoMAE models on the HMI video validation set from Boucheron et al. (2023) (labeled as 'cv' in Table 1), using different training configurations. Each metric is reported for two decision thresholds: $\tau = 0.5$ and $\tau = \tau^*$, where τ^* is estimated from the full TSS–τ curves across multiple training and evaluation runs ($N = 5$) as the threshold that maximizes the mean TSS, subject to minimum recall (0.5) and precision (0.15) constraints. For both thresholds, the table reports the mean and standard deviation of each metric computed over the $N = 5$ runs.

Model ID	Δt (min)	Train Set	Val Set	Loss	TSS		HSS		MCC	
					$\tau = 0.5$	$\tau = \tau^*$	$\tau = 0.5$	$\tau = \tau^*$	$\tau = 0.5$	$\tau = \tau^*$
videomae_v1	36	train_ds	cv_ds	bce_w	0.678 ± 0.037	**0.722 ± 0.038**	0.281 ± 0.049	0.228 ± 0.050	0.360 ± 0.026	0.333 ± 0.041
videomae_v2	36	train_ds_bal	cv_ds_bal	bce	0.696 ± 0.054	0.700 ± 0.044	0.266 ± 0.046	0.244 ± 0.047	0.354 ± 0.024	0.338 ± 0.026
videomae_v3	72	train_ds	cv	bce_w	0.696 ± 0.044	0.710 ± 0.035	0.225 ± 0.072	0.211 ± 0.070	0.325 ± 0.050	0.319 ± 0.051
videomae_v4	72	train_ds_bal	cv_ds_bal	bce	**0.703 ± 0.036**	0.711 ± 0.045	**0.298 ± 0.034**	**0.273 ± 0.035**	**0.378 ± 0.017**	**0.362 ± 0.015**

Table 5: Forecasting metrics for M+ flare prediction obtained by fine-tuned Moirai2 models on the GOES XRS time series validation set (labeled as 'cv' in Table 1), using different training configurations. Each metric is reported for two decision thresholds: $\tau = 0.5$ and $\tau = \tau^*$, where τ^* is estimated from the full TSS–τ curves across multiple training and evaluation runs ($N = 5$) as the threshold that maximizes the mean TSS, subject to minimum recall (0.5) and precision (0.15) constraints. For both thresholds, the table reports the mean and standard deviation of each metric computed over the $N = 5$ runs.

Model ID	T (h)	Train Set	Val Set	Loss	TSS		HSS		MCC	
					$\tau = 0.5$	$\tau = \tau^*$	$\tau = 0.5$	$\tau = \tau^*$	$\tau = 0.5$	$\tau = \tau^*$
moirai2_v1	12	train	cv	bce_w	0.731 ± 0.020	0.736 ± 0.020	0.657 ± 0.079	0.635 ± 0.087	0.665 ± 0.071	0.648 ± 0.075
moirai2_v2	24	train	cv	bce_w	**0.744 ± 0.027**	**0.747 ± 0.018**	**0.690 ± 0.026**	**0.680 ± 0.022**	**0.694 ± 0.026**	**0.686 ± 0.021**

the likelihood of a flare occurring within the 24-hour forecasting window. The predicted class $\hat{y}_i$ is obtained by thresholding this probability at τ (=0.5 by default), i.e., $\hat{y}_i = \mathbb{1}[\hat{p}_i \geq \tau]$. From the pairs $\{\hat{y}_i, y_i\}_{i=1}^{N}$ we computed the entries of the confusion matrix cin the table below:

		Predicted	
		NONE	M+
True	NONE	TN	FP
	M+	FN	TP

where TP, TN, FP, and FN denote the number of true positives, true negatives, false positives and false negatives, respectively.

Binary forecasting performance was evaluated using the following metrics: *True Skill Statistic* (TSS), *Heidke Skill Score* (HSS), *Matthews correlation coefficient* (MCC). Additional details and discussion of these metrics are provided in Appendix B.

All metrics were computed on the validation dataset at the end of each training epoch, and the model achieving the highest TSS score was selected as the best-performing checkpoint. The TSS was chosen as the primary selection criterion because it is largely insensitive to class imbalance, in contrast to metrics such as the F1-score. The selected model was then evaluated on the held-out test set to derive the final forecasting metrics.

A comprehensive discussion of the advantages and limitations of different forecasting metrics is provided by Francisco et al. (2025), who emphasize that HSS and MCC are more robust to variations in dataset composition. Accordingly, both metrics are reported alongside TSS when presenting the forecasting results in Section 5.

5. Results

We began our analysis with the image dataset introduced by Boucheron et al. (2023), aiming to identify optimal training parameters that could later be extended to the video models. As a first step, we conducted multiple training runs on the *train_ds* dataset (see Table 1) using different loss functions to determine the most effective configuration. Specifically, we evaluated a weighted binary cross-entropy loss, a weighted focal loss ($\gamma = 2$), and a score-oriented loss based on the TSS metric. In the latter case, we employed a weighted random sampler to rebalance positive and negative samples within each batch, whereas for the other loss configurations, class imbalance was already compensated through loss weighting. Model selection across epochs was based on performance evaluated on the *cv* dataset. We further investigated three additional configurations in which both training and model selection were performed on balanced downsampled datasets (*train_ds_bal* and *cv_ds_bal* in Table 1). In these cases, no class weighting or batch resampling was applied.

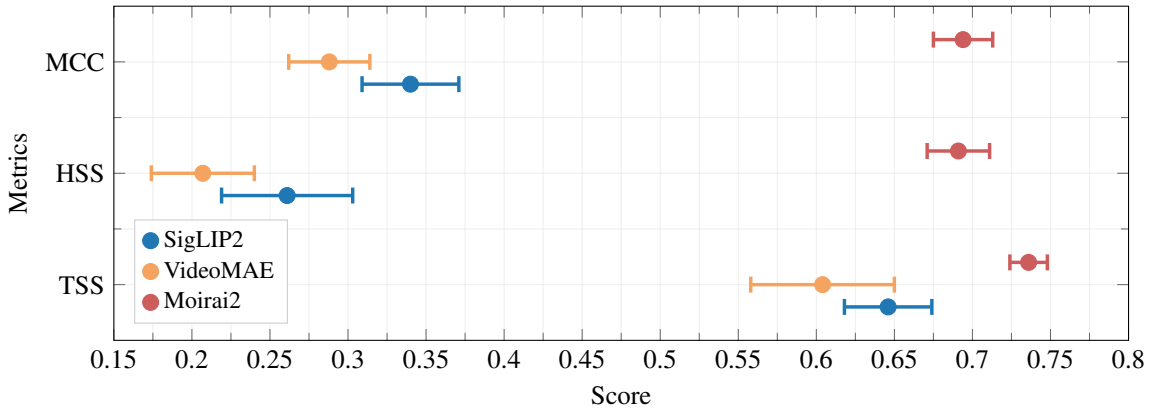


Figure 4: Forecasting metrics for M+ flare prediction obtained with models trained on each data modality (blue: SigLIP2, orange: VideoMAE, red: Moirai2) and evaluated on the test set. For the image and video models, the box plot reports the average score and its standard deviation computed over $N = 5$ independent runs, all trained and evaluated on the same original data split from Boucheron et al. (2023) (downsampled), labelled as "default" in Table 6. For the time series model, the same quantities are reported over $k = 5$ independent data splits.

Table 6: Forecasting metrics for M+ flare prediction obtained with models trained on each data modality and evaluated on the test set. For the image and video models, the table reports the mean and standard deviation of the scores computed over $N = 5$ independent runs, all trained and evaluated on the same original data split from Boucheron et al. (2023) (downsampled), labelled as "default". Additionally, scores averaged over $k = 5$ alternative splits – each randomly drawn from the original dataset as described in Section 3 – are reported and labelled as "alt.". For the time series model, metrics are likewise averaged over $k = 5$ independent data splits.

Model ID	Data split	TSS		HSS		MCC	
		$\tau = 0.5$	$\tau = \tau^*$	$\tau = 0.5$	$\tau = \tau^*$	$\tau = 0.5$	$\tau = \tau^*$
siglip2_v6	default	**0.646 ± 0.028**	**0.671 ± 0.027**	**0.261 ± 0.042**	**0.235 ± 0.044**	**0.340 ± 0.031**	**0.329 ± 0.034**
	alt.	0.645 ± 0.165	0.653 ± 0.161	0.197 ± 0.061	0.165 ± 0.074	0.278 ± 0.076	0.264 ± 0.079
videomae_v4	default	**0.604 ± 0.046**	**0.613 ± 0.050**	**0.207 ± 0.033**	**0.191 ± 0.034**	**0.288 ± 0.026**	**0.278 ± 0.027**
	alt.	0.535 ± 0.073	0.569 ± 0.083	0.179 ± 0.084	0.167 ± 0.082	0.249 ± 0.058	0.249 ± 0.060
moirai_v2		0.736 ± 0.012	0.740 ± 0.009	0.691 ± 0.020	0.683 ± 0.017	0.694 ± 0.019	0.688 ± 0.015

The forecasting metrics (TSS, HSS, and MCC) obtained on the *cv* dataset are summarized in Table 3. For each model configuration (labeled as siglip2_v1-v6), training was repeated $N = 5$ times. Metrics were evaluated at two flare-decision thresholds: τ=0.5 and $\tau = \tau^*$, where τ^* corresponds to the threshold that maximizes the mean TSS, estimated from the full TSS–τ curves across the N runs under the constraint of a minimum recall of 0.5 and a minimum precision of 0.15. For both thresholds, we report the mean and standard deviation of each metric over the repeated runs. Within the run-to-run uncertainties, no statistically significant differences were observed among the different configurations. However, configuration siglip2_v6 – trained on balanced datasets using the BCE loss – achieved slightly higher values for most metrics and was therefore selected for the final evaluation of the image model on the test set.

We investigated two training configurations for the video models at each frame step size (36 and 72 minutes). The first used a weighted BCE loss on unbalanced training and validation datasets, while the second employed a standard BCE loss on balanced datasets. The forecasting metrics obtained on the validation sets for each configuration (labeled as videomae_v1-v4) are summarized in Table 4 for different decision thresholds. Overall, model performances did not vary significantly across configurations. Slightly higher metric values were achieved with the 72-minute frame step and balanced datasets. Compared to the image-based models, the video models exhibited a modest improvement only in the HSS and MCC scores. Configuration videomae_v4 was therefore selected for the final evaluation on the video test set.

We trained the time series models on GOES XRS data using two alternative input sequence lengths, $T = 12$ and 24 hours. The adopted model configurations, labeled as moirai2_v1-v2, along with their corresponding validation metrics averaged over N=5 runs, are reported in Table 5. The model trained with the longer input window (moirai2_v2), reaching slightly higher forecasting scores within the run-to-run uncertainties, was selected for the final evaluation.

Table 6 summarizes the forecasting metrics obtained on the test set for the best-performing models selected through validation across all data modalities. For the image and video models, the table reports the mean and standard deviation of the scores computed over $N = 5$ independent runs, each trained and evaluated on the same original downsampled data split from Boucheron et al. (2023), labelled as "default". In addition, metrics averaged over $k = 5$ alternative splits – each randomly drawn from the original dataset as described in Section 3 – are also reported and labelled as "alt.". For the time series model, metrics are likewise averaged over $k = 5$ independent data splits.

The best performance metrics for each model are shown in Fig. 4 as colored markers – blue for SigLIP2, orange for VideoMAE, and red for Moirai2 – indicating the mean scores with their corresponding standard deviations. The SigLIP2 image model achieved TSS=0.646, HSS=0.261, and MCC=0.340,

confirming the ability of vision–language transformers to extract meaningful magnetic features from single magnetograms. The VideoMAE model, trained on 16-frame magnetogram sequences, yielded slightly lower scores (TSS=0.604, HSS=0.207, MCC=0.288), suggesting a limited ability to fully exploit temporal information and capture pre-flare evolution, even when varying cadence intervals (32, 72).

Overall, both the image and video models achieve slightly higher metric values when evaluated on the original data split compared to the alternative samples. The Moirai2 time-series model achieved the best overall performance, with TSS=0.736, HSS=0.691, and MCC=0.694, outperforming both the image and video models, particularly for the HSS and MCC metrics.

The TSS scores obtained with single-image data (∼0.65) are consistent with values reported in previous studies using SDO/HMI AR magnetograms for the same forecasting task (M+ class prediction with a 24-hour horizon), despite differences in data splits and model architectures. In contrast, models trained on multi-wavelength full-disk data (e.g., Sun et al. 2023; Francisco et al. 2025) have achieved higher HSS and MCC values, typically around 0.4. For the video models, all tested configurations tend to slightly underperform on the test set compared with both our image-based models and recently published video-based approaches (e.g., Li et al. 2025; Xu et al. 2025). While the inclusion of temporal information is physically expected to provide a measurable advantage over single-image models, such improvements have generally been modest. Other studies (e.g., Francisco et al. 2024) report performance gains of ∼0.02–0.03 in skill scores – often comparable to the typical run-to-run uncertainty. We therefore performed additional training runs to further investigate these results.

We first examined whether alternative video cadences and sequence lengths could yield performance improvements. In addition to the previously tested shorter sequences (frame cadence Δt = 32 min, total duration $T \approx 9.6$ h), we considered much longer sequences (Δt = 288 min, $T \approx 3.2$ days). The latter configuration reduced the size of the training dataset by a factor of 2.6 compared to the 72-minute cadence case. Models were trained and evaluated under the same conditions as the selected `videomae_v4` configuration. In both cases, the resulting HSS and MCC metrics on the test set were comparable, while the TSS scores were lower: 0.565 ± 0.059 for Δt = 36 min and 0.507 ± 0.062 for Δt = 288 min.

Since no significant differences were observed across alternative video data settings, we next investigated whether performance could be improved on the Δt=72 min dataset through alternative hyperparameter configurations. One possible explanation for the limited gains is that the pretrained VideoMAE representations – learned from approximately 300,000 natural videos – may transfer less effectively to the solar domain compared to SigLIP2, which was pretrained on a much larger dataset of about 10 billion images. Because a large portion of the encoder weights was frozen during fine-tuning, this may have constrained the model's capacity to adapt to the solar data. To test this hypothesis, we performed a new experiment in which the entire VideoMAE encoder was fine-tuned. To further mitigate overfitting, we introduced a dropout layer (p = 0.3) in the classification head and applied weight decay (λ = 0.07) relative to the previous configurations. The resulting model reached comparable HSS and MCC scores and a slightly better TSS (0.636 ± 0.046), still compatible with previous results within the run-to-run uncertainties.

We also tested alternative learning rate and warm-up settings, reducing the learning rate from 1e-4 to 5e-5 and varying the warm-up ratio between 0.1 and 0.3. These configurations, however, yielded forecasting scores comparable to the default setup.

6. Summary

In this work, we conducted a systematic evaluation of recent transformer-based architectures applied to solar flare forecasting across three complementary data modalities – images, videos, and time series. Using publicly available datasets of SDO/HMI magnetograms and GOES soft X-ray fluxes, we benchmarked the *SigLIP2*, *VideoMAE*, and *Moirai2* models, each representing a state-of-the-art transformer design in its respective domain. All models were trained and validated under consistent data splits, balanced sampling strategies, and identical evaluation protocols based on the TSS, HSS, and MCC forecasting metrics. All trained model weights, software code, and datasets have been publicly released to promote reproducibility and collaboration, supporting the development of next-generation, physics-informed, and operationally deployable flare forecasting systems.

From a physical standpoint, the comparative behavior of the three transformer models reflects the distinct nature of the information encoded in each data modality. The SigLIP2 model, trained on single-frame magnetograms, captures the instantaneous spatial distribution of magnetic fields in active regions – primarily the polarity inversion lines, magnetic shear, and δ-spot structures – which are well-established indicators of stored free magnetic energy and potential flare productivity (e.g., Romano et al., 2014, 2018, 2024). However, without temporal context, the model remains sensitive only to static morphological proxies of magnetic non-potentiality.

The VideoMAE architecture extends this representation to the temporal domain, tracking the short-term evolution of active regions and thus probing the processes of flux emergence, shear accumulation, and magnetic cancellation that precede energy release. Nevertheless, its forecasting gain over static images remains moderate. This suggests that, within the ∼10–20 h windows adopted, the temporal variability of the photospheric field still traces quasi-stationary phases of magnetic energy build-up, while the rapid reconfiguration associated with reconnection occurs on shorter coronal timescales not fully sampled by the input sequences. For longer temporal windows (∼3 days), the limited size and cadence of the available video dataset likely prevent the model's ability to learn these subtle pre-eruptive dynamics.

The superior performance of the Moirai2 time-series model highlights instead the predictive value of irradiance-based temporal information. The GOES X-ray flux encodes the integrated coronal response to energy release and heating, thus serving as a direct tracer of magnetic energy dissipation. The model's ability

to identify flare precursors from smooth irradiance variations and flare history suggests that the temporal evolution of coronal emission already contains the statistical imprint of the magnetic processes governing flare occurrence. In this sense, Moirai2 acts as a data-driven proxy for the coronal energy storage and release cycle, whereas image and video models probe only its boundary conditions at the photosphere.

Overall, the comparative results indicate that magnetic morphology alone provides a necessary but not sufficient condition for flare occurrence, while irradiance time series effectively encode the integrated temporal signature of the coronal response. A future unified multimodal architecture combining both the spatial topology of magnetic fields and the temporal evolution of coronal emission could thus enable physically consistent, end-to-end forecasting of solar eruptive activity.

Building on these findings, we plan to extend this work to a revised and enriched multimodal dataset integrating full-disk video sequences of HMI and EUV observations across multiple wavelengths, together with multivariate time series from GOES irradiance sensors and other observables (e.g., AR features). The new dataset will include additional metadata and diagnostic flags – currently unavailable in the Boucheron et al. (2023) dataset – enabling more comprehensive investigations, such as distinguishing the influence of single versus multiple active regions and assessing event directionality and potential geoeffectiveness at Earth.

Future developments will also explore generative modeling (e.g., GANs or diffusion-based models) to expand and rebalance the training dataset, and self-supervised pretraining of video models to enhance temporal representation learning. These efforts aim to overcome current data limitations and move toward a robust, physics-aware multimodal architecture for operational solar flare forecasting.

Acknowledgements

Supported by Italian Research Center on High Performance Computing Big Data and Quantum Computing (ICSC), project funded by European Union - NextGenerationEU - and National Recovery and Resilience Plan (NRRP) - Mission 4 Component 2 within the activities of Spoke 3 (Astrophysics and Cosmos Observations).

We acknowledge ISCRA for awarding this project access to the LEONARDO supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CINECA (Italy). Additional computing resources for this work were provided by the INAF "CIRASA" (*Collaborative and Integrated platform for Radio Astronomical Source Analysis*) project, and the Italian PNRR Project IR0000034 "STILES" (*Strengthening the Italian leadership in ELT and SKA*) project. S.R. acknoledges the INAF SCIARADA grant for financial support.

Data Availability

The software code used in this work is publicly available under the GNU General Public License v3.0[15] on the GitHub repository https://github.com/inaf-oact-ai/solar-flare-forecaster/.

The datasets and trained model weights have been made available in this *Hugging Face* repository: https://huggingface.co/inaf-oact-ai.

[15] https://www.gnu.org/licenses/gpl-3.0.html

References

Abduallah, Y., et al., 2023, Sci Rep, 13, 13665, https://doi.org/10.1038/s41598-023-40884-1

Benz, A. O., 2017, *Living Reviews in Solar Physics*, 14, 1, 2, https://doi.org/10.1007/s41116-016-0004-3

Bloomfield, D. S., et al, 2012, ApJL, 747, L41, https://doi.org/10.1088/2041-8205/747/2/L41

Boucheron, L. E., et al., 2023, Sci Data 10, 825, https://doi.org/10.1038/s41597-023-02628-8

Chen, Y., et al., 2019, Space Weather, 17, 1404, https://doi.org/10.1029/2019SW002214

Chen, X., et al., 11th International Conference on Learning Representations (ICLR), 2023, https://doi.org/10.48550/arXiv.2209.06794

Dei, D., and Fuentes, O., 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 2020, pp. 1-7, https://doi.org/10.1109/IJCNN48605.2020.9207284

Donahue, P. K., and Inceoglu, F., 2024, Front. Astron. Space Sci., 10, 1298609, https://doi.org/10.3389/fspas.2023.1298609

Dosovitskiy, A., et al., 2020, https://doi.org/10.48550/arXiv.2010.11929

Deng, Z., et al., 2021, ApJ, 922, 232, https://doi.org/10.3847/1538-4357/ac2b2b

Francisco, G., et al., 2024, https://doi.org/10.48550/arXiv.2410.16116

Francisco, G., et al., 2025, ApJ, 985, 108, https://doi.org/10.3847/1538-4357/adc56d

Gan, W., et al., 2023, Sol Phys, 298, 68, https://doi.org/10.1007/s11207-023-02166-x

Goyal, R., et al., Proceedings of the IEEE international conference on computer vision, 2017, 5842–5850, https://doi.org/10.48550/arXiv.1706.04261

Grim, L. F.L., Gradvohl, A. L.S., 2024, Sol Phys, 299, 33, https://doi.org/10.1007/s11207-024-02276-0

Guastavino, S., et al., 2022, A&A, 662, A105, https://doi.org/10.1051/0004-6361/202243617

Guastavino, S., et al., 2023, Front. Astron. Space Sci., 9, 1039805, https://doi.org/10.3389/fspas.2022.1039805

Hanssen, A., & Kuipers, W., 1965, *On the Relationship between the Frequency of Rain and Various Meteorological Parameters*, 58 (Koninklijk Nederlands Meteorologisch Instituut)

Huang, X., et al., 2018, ApJ, 856, 7, https://doi.org/10.3847/1538-4357/aaae00

Kay, W., et al., 2017, https://doi.org/10.48550/arXiv.1705.06950

Kilpua, E. K. J., et al., 2017, *Living Reviews in Solar Physics*, 14, 5, https://doi.org/10.1007/s41116-017-0009-6

Legnaro, E., et al., 2025, ApJ, 981, 157, https://doi.org/10.3847/1538-4357/adb41a

Li, X., et al., 2025, ApJS, 278, 63, https://doi.org/10.3847/1538-4365/add149

Marchetti, F., et al., 2022, Pattern Recognition, 132, 108913, https://doi.org/10.1016/j.patcog.2022.108913

Müller, D., et al., 2020, A&A, 642, A1, https://doi.org/10.1051/0004-6361/202038467

Pandey, C., et al., 2023 IEEE Sixth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), Laguna Hills, CA, USA, 2023, pp. 83-90, https://doi.org/10.1109/AIKE59827.2023.00021

Pesnell, W. D., et al., 2012, Sol Phys, 275, 3, https://doi.org/10.1007/s11207-011-9841-3

Priest, E. R. & Forbes, T. G. 2002, A&A Review, 10, 4, 313, https://doi.org/10.1007/s001590100013

Pulkkinen, A., et al., 2017, Space Weather, 15, 828, https://doi.org/10.1002/2016SW001501
Radford, A., et al., *Learning transferable visual models from natural language supervision*, In International Conference on Machine Learning, 2021, PMLR, 8748, https://doi.org/10.48550/arXiv.2103.00020
Raouafi, N. E., et al., 2023, Space Sci Rev, 219, 8, https://doi.org/10.1007/s11214-023-00952-4
Romano, P., et al., 2014, ApJ, 794, 2, 118, https://doi.org/10.1088/0004-637X/794/2/118
Romano, P., et al., 2018, ApJL, 852, 1, L10, https://doi.org/10.3847/2041-8213/aaa1df
Romano, P., et al., 2024, ApJL, 973, 1, L31, https://doi.org/10.3847/2041-8213/ad77cb
Shibata, K. & Magara, T., 2011, *Living Reviews in Solar Physics*, 8, 1, 6, https://doi.org/10.12942/lrsp-2011-6
Sun, P., et al., 2022, ApJ, 941, 1, https://doi.org/10.3847/1538-4357/ac9e53
Sun. D., et al., 2023, ApJS, 266, 8, https://doi.org/10.3847/1538-4365/acc248
Temmer, M., 2021, *Living Reviews in Solar Physics*, 18, 1, 4, https://doi.org/10.1007/s41116-021-00030-3
Tong, Z., et al., *VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training*, In NeurIPS, 2022
Toriumi, S. & Wang, H. 2019, *Living Reviews in Solar Physics*, 16, 1, 3, https://doi.org/10.1007/s41116-019-0019-7
Tschannen, M., et al., *SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features*, 2025, https://doi.org/10.48550/arXiv.2502.14786
Tripathi, D., et al., 2022, Proceedings of the International Astronomical Union, 18(S372), 17-27, https://doi.org/10.1017/S1743921323001230
van der Sande, K., and Muñoz-Jaramillo, A., 2025, ApJ, 984, 87, https://doi.org/10.3847/1538-4357/ad8de5
Vaswani, A., et al., Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 5998-6008
Xu, D., et al., 2025, ApJS, 276, 68, https://doi.org/10.3847/1538-4365/ada281
Yan, P., et al., 2024, Astrophys Space Sci, 369, 110, https://doi.org/10.1007/s10509-024-04374-8
Wang, L., et al., *Videomae v2: Scaling video masked autoencoders with dual masking*, In CVPR, 2023
Woo, G., et al., 2024, Proceedings of the 41st International Conference on Machine Learning (ICML'24), Vienna, Austria, 2178, pp 53140-53164, https://doi.org/10.48550/arXiv.2402.02592
Zhai, X., et al., *Sigmoid loss for language image pretraining*, In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, 11975, https://doi.org/10.48550/arXiv.2303.15343
Zhang, J., et al., 2001, ApJ, 559, 1, 452, https://doi.org/10.1086/322405
Zheng, Y., et al., 2019, ApJ, 885, 73, https://doi.org/10.3847/1538-4357/ab46bd
Zhukov, A. N., et al., 2025, https://doi.org/10.48550/arXiv.2509.00253

Appendix A. Training losses

For single-label multi-class forecasting, the framework currently supports these possible losses:

- *Cross entropy (CE) loss*: $\mathcal{L}_{\rm ce}$

$$\mathcal{L}_{\rm ce} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1} w_k\, \mathbb{1}\{y_i = k\}\left(-\log\frac{\exp(z_{ik})}{\sum_{j=0}^{K-1}\exp(z_{ij})}\right).$$

- *Focal loss*: $\mathcal{L}_{\rm focal}$

$$\mathcal{L}_{\rm focal} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1} w_k\, \mathbb{1}\{y_i = k\}\,(1-p_{ik})^{\gamma}\,\log(p_{ik}).$$

- *Score-oriented loss*: $\mathcal{L}_{\rm sol}$

$$\begin{aligned}
\mathcal{L}_{\rm sol} &= -\frac{1}{K}\sum_{k=0}^{K-1}\mathrm{TSS}_k \\
\mathrm{TSS}_k &= \frac{\mathrm{TP}_k}{\mathrm{TP}_k+\mathrm{FN}_k} + \frac{\mathrm{TN}_k}{\mathrm{TN}_k+\mathrm{FP}_k} - 1 \\
\mathrm{TP}_k &= \sum_{i=1}^{N} y_{ik}\, p_{ik} \\
\mathrm{TN}_k &= \sum_{i=1}^{N}(1-y_{ik})(1-p_{ik}) \\
\mathrm{FP}_k &= \sum_{i=1}^{N}(1-y_{ik})\, p_{ik} \\
\mathrm{FN}_k &= \sum_{i=1}^{N} y_{ik}(1-p_{ik})
\end{aligned}$$

where N is the batch size (i.e. number of training samples in the batch), K is the number of classes, z_{ik} are the model logits for sample i and class k, $y_i \in \{0,\ldots,K-1\}$ are the true class indices with y_{ik} their one-hot encoding, p is the predicted probability of the true class, w_k are the class weights, γ is a focusing parameter, and $\mathbb{1}\{\cdot\}$ is the indicator function. TSS denote the *True Skill Statistic* ($\in$[-1,1]), while TP, TN, FP, and FN are the number of true positives, true negatives, false positives and false negatives, respectively. Similarly, the framework provides the corresponding losses for the binary forecasting task.
The class weights w_k are pre-computed according to the forecasting task:

$$\text{(Multiclass)}\quad w_k = \frac{N}{K\,n_k},$$

$$\text{(Binary)}\quad w_k = \frac{n_{neg}}{n_{pos}}$$

where N is the total number of samples, n_k the number of samples belonging to class k, n_{neg} and n_{pos} are the number of negatives and positive samples, respectively.

Appendix B. Forecasting metrics

F1-score. The F1 score summarizes the balance between misses and false alarms at a fixed threshold by the harmonic mean of precision $\mathcal{P}$ and recall $\mathcal{R}$:

$$\mathrm{F1} \;=\; 2\,\frac{\mathcal{P}\times\mathcal{R}}{\mathcal{P}+\mathcal{R}} = \frac{2\,\mathrm{TP}}{2\,\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} \tag{B.1}$$

where the recall $\mathcal{R}$ (or true positive rate/sensitivity TPR) and the precision $\mathcal{P}$ are respectively computed as:

$$\mathcal{R} = \text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{B.2}$$

$$\mathcal{P} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{B.3}$$

F1 ranges in $[0, 1]$, with 1 indicating perfect classification.

True Skill Statistic (TSS). The TSS score, also known as *Hanssen & Kuipers' Discriminant* (H&KSS) (Hanssen & Kuipers, 1965) measures the forecast's ability to correctly predict both positive (event) and negative (non-event) outcomes:

$$\text{TSS} = \frac{\text{TP}}{\text{TP} + \text{FN}} - \frac{\text{FP}}{\text{FP} + \text{TN}} = \text{TPR} + \text{TNR} - 1 \tag{B.4}$$

where TNR is the true negative rate (specificity), computed as:

$$\text{TNR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{B.5}$$

TSS ranges from -1 (perfectly wrong) through 0 (no skill, random/-constant forecast) to 1 (perfect). TSS is widely used in space-weather because it is relatively insensitive to class imbalance, while other metrics like the F1-score are more affected when the proportion of the minority class (rare event) changes.

Heidke Skill Score (HSS). HSS score, like TSS, includes all confusion matrix entries to measure the forecasting accuracy relative to random chance:

$$\text{HSS} = \frac{2\,(\text{TP} \times \text{TN} - \text{FN} \times \text{FP})}{(\text{TP} + \text{FN}) \times (\text{FN} + \text{TN}) + (\text{TP} + \text{FP}) \times (\text{FP} + \text{TN})}. \tag{B.6}$$

Its range is $(-\infty, 1]$, with 1 perfect forecasting and 0 indicating no improvement over chance. Unlike TSS, HSS is sensitive to a varying event/no-event sample ratio (Bloomfield et al., 2012).

Matthews correlation coefficient (MCC). MCC is the Pearson correlation between predicted and observed binary outcomes:

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP}) \times (\text{TP} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TN} + \text{FN})}}. \tag{B.7}$$

It ranges from -1 (total disagreement) through 0 (average random prediction) to 1 (perfect prediction). MCC uses all four entries of the confusion matrix and is robust to imbalance, making it suitable when both classes are important.