universe

MDPI

Article

JW-FD: A 15-Year Multimodal Dataset for Solar Flare Forecasting

Mingfu Shao [1], Jiaben Lin [1,*], Hui Wang [1,2], Liyue Tong [1], Chen Yang [1,2], Yin Zhang [1,2] and Yuyang Li [1,2]

[1] State Key Laboratory of Solar Activity and Space Weather, NAOC, Beijing 100101, China; shaomf@bao.ac.cn (M.S.); wanghui@nao.cas.cn (H.W.); lytong@bao.ac.cn (L.T.); chenyang@bao.ac.cn (C.Y.); yzhang@bao.ac.cn (Y.Z.); liyuyang22@mails.ucas.ac.cn (Y.L.)
[2] University of Chinese Academy of Sciences, Beijing 101408, China
* Correspondence: jiabenlin@bao.ac.cn

Abstract

Solar flares drive severe space weather hazards, and forecasting their occurrence remains a central challenge for both heliophysics and operational space weather services. Data-driven methods require long-horizon datasets in which images, magnetic features, and flare labels are coregistered in space and time. We present JW-FD (JinWu-Flare Dataset), a 15-year multimodal release spanning 1 January 2011 through 31 December 2025, constructed from SDO/HMI line-of-sight magnetograms, NOAA Solar Region Summary reports, and NOAA X-ray flare event lists. The dataset comprises 3,064 independent active regions and 1,991,247 coregistered magnetogram crops, together with FITS, PNG, CSV, and MP4 modalities. Each sample provides 29 magnetic features linked to configurable flare labels under a strict pre-eruption window spanning seven forecast horizons and four GOES intensity thresholds. An 8:1:1 split at the active region level is adopted to prevent temporal leakage between partitions. PNG branches are released at ten magnetic saturation thresholds, and internal Transformer and four-model 24 h experiments on ≥C1.0 and ≥M1.0 forecasting suggest $B_{\mathrm{th}} = 1000$ G as a preliminary default, although skill remains modest and the optimal saturation is model- and task-dependent. The open-source construction pipeline and dataset are publicly available .

Keywords: solar flare forecasting; magnetograms; multimodal dataset; machine learning

1. Introduction

Solar flares are among the most energetic manifestations of solar magnetic activity. Major (M- and X-class) flares can disrupt satellites, communications, and power grids, making reliable flare forecasting a priority for both fundamental solar physics and operational space weather services [1–3].

Photospheric line-of-sight (LoS) magnetograms encode signatures of magnetic complexity (gradient concentrations, neutral line topology, and flux imbalance) that precede many eruptive events [4,5]. Quantitative features derived from magnetograms have long supported statistical and machine learning (ML) flare forecasting, and recent work extends this paradigm to convolutional neural networks (CNNs), vision transformers, and multimodal models that consume both images and structured parameters [6,7]. These methods require datasets in which images, physical features, and flare labels are coregistered in space and time, under a partitioning scheme that prevents temporal leakage across splits.

Several public datasets now support data-driven flare forecasting, but they differ in temporal span, modality coverage, and evaluation design. Boucheron et al. released an AR LoS magnetogram image corpus covering 2010–2018 [8]. Angryk et al. released SWAN-SF,

Check for updates

Academic Editor: Costas Alissandrakis

Received: 27 July 2026
Revised: 9 September 2026
Accepted: 14 September 2026
Published: 16 September 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

a multivariate time series corpus of Space weather HMI Active Region Patch (SHARP) parameters over a similar interval [9,10]. Nishizuka et al. assembled feature databases associated with Deep Flare Net (DeFN) [11,12], and related work continues to organize SHARP parameter sets for operational forecasting [13,14]. Complementary studies have assembled magnetogram or multiwavelength datasets for CNNs, long short-term memory (LSTM) networks, and related models [15–18], yet many remain study-specific and are not publicly redistributed as reusable, long-horizon, multimodal corpora [1,19,20].

More recently, SuryaBench provides a large multitask heliophysics benchmark for foundation model research, with solar flare forecasting as one of several applications [21]. Its flare subset uses full disk SDO context with a fixed 24 h labeling window and a calendar-based train/test partition, rather than NOAA AR crops with configurable multi-horizon labels and an AR level split. Relative to Boucheron et al. [8], SWAN-SF [9], and the SuryaBench flare task [21], the JinWu-Flare Dataset (JW-FD) supplies a long-horizon, AR-centered multimodal release in which images, magnetic features, flare labels, and evolution videos share a common index under a leakage-aware AR-level partition [3].

JW-FD is designed to meet this need. This paper presents the dataset, its design rationale, construction pipeline, and recommended usage. Our principal contributions are as follows:

1. A 15-year multimodal release (2011–2025) covering solar Cycles 24–25 with 1.99 million active region magnetogram crops, supporting studies of cycle-dependent flare statistics and rare X-class events.
2. Coregistered images, magnetic features, flare labels, and evolution videos, with 66 dimensional records that link 29 magnetic features to configurable labels defined before eruption.
3. Configurable labels over seven forecast horizons (1–72 h) and four GOES intensity thresholds (C1.0–X1.0), defined on a strict pre-eruption window.
4. An 8:1:1 training, validation, and test split. The data are divided into training, validation, and test sets in an 8:1:1 ratio, with all observations from the same active region assigned to a single partition. This prevents information from the same region at different times from leaking between the datasets.

2. Methods

JW-FD is constructed with a six-step automated pipeline that ingests Solar Region Summary (SRS) reports and NOAA Events X-ray (XRA) flare lists from the National Oceanic and Atmospheric Administration/Space Weather Prediction Center (NOAA/SWPC), together with Solar Dynamics Observatory (SDO)/HMI [22,23] full disk LoS magnetograms. The pipeline exports Flexible Image Transport System (FITS) crops, Portable Network Graphics (PNG) images, comma separated values (CSV) tables, and MPEG-4 (MP4) videos. Figure 1 summarizes the workflow; Table 1 lists the inputs and outputs of each step.

Table 1. Pipeline step inputs and outputs. Here B_{th} denotes the magnetic saturation threshold (in Gauss) applied when converting FITS magnetograms to 8-bit PNG images; the release includes ten branches from 200 to 2000 G in steps of 200 G.

Step	Script	Input	Output
1	`step1_extract_flare_labels`	NOAA Events `.txt` (XRA)	`flare_labels.csv`
2	`step2_crop_active_regions`	SRS + HMI FITS (4096^2)	`fits_600/{AR}/`
3	`step3_convert_to_png`	Step 2 FITS	`png_600_Th{`B_{th}`}/`
4	`step4_generate_dataset_fits`	FITS + labels	FITS branch CSV
5	`step5_generate_dataset_png`	PNG + labels	PNG branch CSV
6	`step6_make_movie`	PNG sequences	`movies/AR{num}_evolution.mp4`

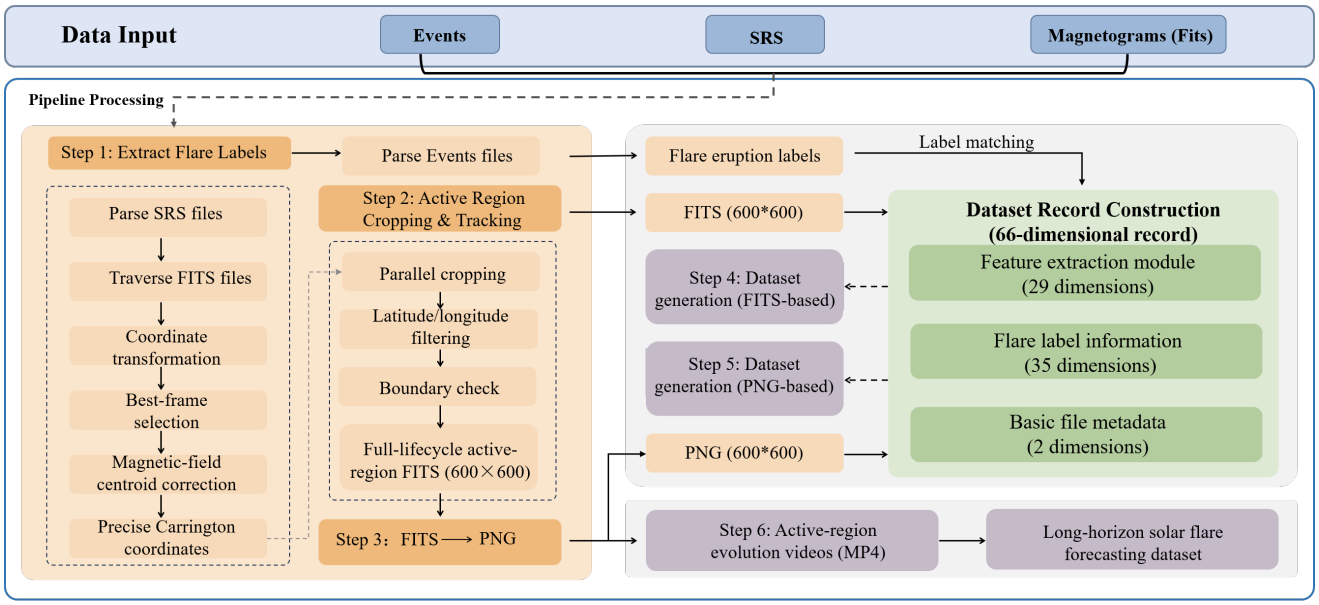


Figure 1. JW-FD dataset construction workflow. Step 1 extracts NOAA X-ray flare events. Step 2 crops and tracks NOAA active regions on SDO/HMI full-disk line-of-sight magnetograms. Step 3 converts FITS crops to 8-bit PNG images at ten magnetic saturation thresholds. Steps 4 and 5 extract the same 29-dimensional feature set and share label-matching logic on the FITS and PNG branches, respectively. Step 6 encodes per-active-region PNG sequences as evolution videos.

2.1. Data Sources

SDO/HMI provides full-disk LoS magnetograms at 4096×4096 pixels (∼1.5″ per pixel) with a cadence of approximately 12 min. The source frames include limb-darkening correction but are not projection-corrected: they remain in the observer plane, and B_{LoS} is not converted to an approximate radial field. NOAA/SWPC SRS reports list daily NOAA AR numbers and heliographic locations. NOAA Events files record GOES soft X-ray flare begin, maximum, and end times, together with GOES class strings (C, M, or X followed by a numeric subclass).

2.2. Design Choices

Four design choices distinguish JW-FD from SHARP-based patch datasets and earlier AR image datasets [5,8].

Long temporal coverage. Fifteen calendar years (2011–2025) span the declining phase of Solar Cycle 24, the 2018–2020 minimum, and the rising phase of Cycle 25. This span supports cycle-dependent statistics and supplies enough rare M- and X-class events for threshold-specific studies.

Multimodal coregistration. Every sample is indexed by NOAA AR number and Coordinated Universal Time (UTC) timestamp across FITS magnetograms (Gauss scale physics), PNG frames (vision model input), 29-dimensional magnetic feature vectors, configurable flare labels, and per-AR MP4 evolution videos. Users may train end-to-end vision models on PNG or MP4 inputs, parameter-based classifiers on extracted features, or multimodal models that fuse imagery with structured metadata [6].

Configurable pre-eruption labels. Binary and class labels are defined for seven forecast horizons ($h \in \{1, 3, 6, 12, 24, 48, 72\}$ h) and four GOES thresholds ($\tau \in \{\mathrm{C1.0}, \mathrm{M1.0}, \mathrm{M5.0}, \mathrm{X1.0}\}$). Positive labels are assigned only within the half-open interval $[\mathrm{Begin} - h, \mathrm{Begin})$, so frames at or after eruption onset are not treated as positive (Figure 2).

AR-level evaluation split. Consecutive HMI frames of the same AR exhibit strong temporal autocorrelation, so assigning different timestamps of one AR to both training and

test sets constitutes information leakage [3,20]. JW-FD therefore adopts an 8:1:1 random split at the AR level (Section 4.1).

Additional practical choices include: (i) 600×600 pixel crops (∼300″ field of view); (ii) NOAA AR-centered directory organization rather than HMI Active Region Patch (HARP)/SHARP patch identifiers, matching operational numbering; (iii) a disk center filter ($|\text{longitude}| \leq 60°$) that reduces the most severe limb effects but does not apply a geometric projection correction; and (iv) dual FITS and PNG CSV branches so that physics-faithful (Gauss scale) and vision-ready (normalized intensity) feature extraction share identical label logic.

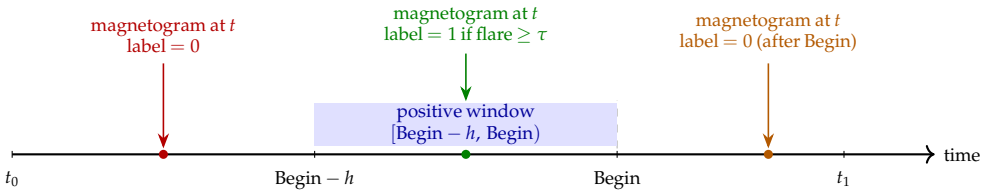


Figure 2. Pre-eruption labeling window for horizon h. A magnetogram observed at time t receives a positive binary label for threshold τ only if $t \in [\text{Begin} - h, \text{Begin})$ and the matched GOES class meets or exceeds τ. Observations at or after Begin (dashed line) are excluded from the positive class.

2.3. Active Region Cropping and Tracking

Tracking begins from the SRS Stonyhurst location of each NOAA AR at its first available magnetogram time, which is converted to a Carrington anchor. That anchor is then carried across subsequent epochs so that the same heliographic region remains centered as the Sun rotates. Local refinement masks pixels stronger than 100 G inside an 80-pixel neighborhood of the projected center and updates the center to their mass centroid. From each full-disk FITS frame we extract a 600×600 crop and discard ARs with $|\text{longitude}| > 60°$.

2.4. FITS-to-PNG Conversion

Cropped FITS magnetograms are converted to 8-bit PNG images for subsequent vision-based analysis. Each crop is mirrored left to right to follow the conventional north–up, east–left display orientation. Magnetic field values are then clipped at a chosen saturation threshold and linearly mapped to the integer range $[0, 255]$:

$$B' = \text{clip}(B, -B_{\text{th}}, B_{\text{th}}), \quad I = \left\lfloor \frac{B' + B_{\text{th}}}{2B_{\text{th}}} \times 255 \right\rfloor, \tag{1}$$

Invalid (Not a number, NaN) pixels are assigned a value of 0 prior to clipping, so that 0 G occupies grey level 127. We publish PNG trees for ten saturation levels, $B_{\text{th}} = 200$–2000 G in steps of 200 G, each with matched 29-dimensional features and flare labels; Section 3.2 discusses how this choice affects downstream models.

2.5. Feature Extraction and Label Matching

Steps 4 and 5 generate CSV datasets from FITS and PNG inputs. Both branches extract the same 29 magnetic features (Table 2) and apply the same label-matching logic, but they operate on different physical domains. FITS features are computed on Gauss-scale magnetograms with no grey-level offset; invalid (NaN) pixels are replaced by 0 G before extraction. PNG features are computed on 8-bit intensity images after subtracting a 127-level zero point, which is the uint8 bin that contains 0 G under a fixed $\pm B_{\text{th}}$ stretch (Equation (1)). The 29 features follow the snapshot (non-temporal) subset of the photospheric proxies introduced by Al-Ghraibah et al. [24] and subsequently used in a multivariate time-series

setting by Boucheron et al. [4]. Al-Ghraibah et al. extracted flux, gradient, neutral-line, flux-evolution, and wavelet descriptors from line-of-sight magnetograms and classified whether an active region would produce a flare of GOES class ≥C1.0. JW-FD retains the 7 gradient, 13 neutral-line, 5 wavelet, and 4 flux descriptors (29 in total) as per-frame attributes; the 9 flux-evolution descriptors of Al-Ghraibah et al. are not precomputed, because the released image and video sequences already allow users to form temporal differences if desired. Operationally, (i) Sobel filters yield a gradient-magnitude map from which we store the mean, standard deviation, median, extrema, skewness, and kurtosis; (ii) a 10×10 boxcar mean (implemented as a uniform filter, with near-zero window means reset to 0 G to avoid spurious contours from floating-point residue) and zero-Gauss contours define polarity inversion lines, which are then weighted by gradient magnitude to obtain length, fragment count, gradient-weighted length, and statistics of curvature and bending energy; (iii) a five-level Haar decomposition yields orientation-summed sub-band energies L1–L5; and (iv) total positive, negative, signed, and unsigned flux are obtained by summing pixel values over the corresponding polarity masks. These quantities are physically interpretable proxies of magnetic complexity and free-energy storage, and the open-source pipeline regenerates the same numbers from the released FITS crops.

Table 2. Twenty-nine magnetic features and their CSV column names.

Category	Feature Names
Gradient (7)	Gradient mean; Gradient std; Gradient median; Gradient min; Gradient max; Gradient skewness; Gradient kurtosis
Neutral line (13)	NL length; NL no. fragments; NL gradient-weighted length; NL curvature mean/std/median/min/max; NL bending energy mean/std/median/min/max
Wavelet (5)	Wavelet Energy L1; Wavelet Energy L2; Wavelet Energy L3; Wavelet Energy L4; Wavelet Energy L5
Flux (4)	Total positive flux; Total negative flux; Total signed flux; Total unsigned flux

Figure 3 illustrates the four feature families on NOAA AR 11158 at 17:00 UT on 13 February 2011. Panel (a) shows the line-of-sight magnetogram from which the four flux integrals are computed. Panel (b) is the Sobel gradient-magnitude map that supplies the seven gradient statistics. Panel (c) shows the neutral-line descriptors: thin lime dashes are all zero-Gauss contours after 10×10 smoothing (the noisy inversion line), and thin magenta markers mark only the high-gradient subset of the gradient-weighted neutral line (top 10% of $|\nabla B|$ on the contour), with marker size scaled by $|\nabla B|$. Panel (d) displays the five-level Haar coefficient array whose orientation-summed sub-band energies are stored as L1–L5.

Relative to image-only AR corpora [8], SHARP parameter time series [9], and full-disk flare benchmarks [21], JW-FD is distinctive here in that the 29 magnetic features, configurable flare labels, and FITS/PNG/MP4 products share a common AR-and-time index and can be regenerated from the released crops by the open-source pipeline.

Each CSV row corresponds to one magnetogram time stamp. We define seven forecast horizons $h \in \{1, 3, 6, 12, 24, 48, 72\}$ h and four GOES thresholds $\tau \in \{\text{C1.0}, \text{M1.0}, \text{M5.0}, \text{X1.0}\}$. The binary column `flare_label_{`$\tau$`}_{`$h$`hr}` equals 1 when at least one flare in the same AR has GOES class $\geq \tau$ and a begin time Begin in the half-open window

$$\text{Begin} - h \;\leq\; t \;<\; \text{Begin} \tag{2}$$

and equals 0 otherwise. Separately, `flare_class_{`h`hr}` stores the strongest GOES class among flares that fall in the same window for horizon h, and equals `0` when no such flare occurs.

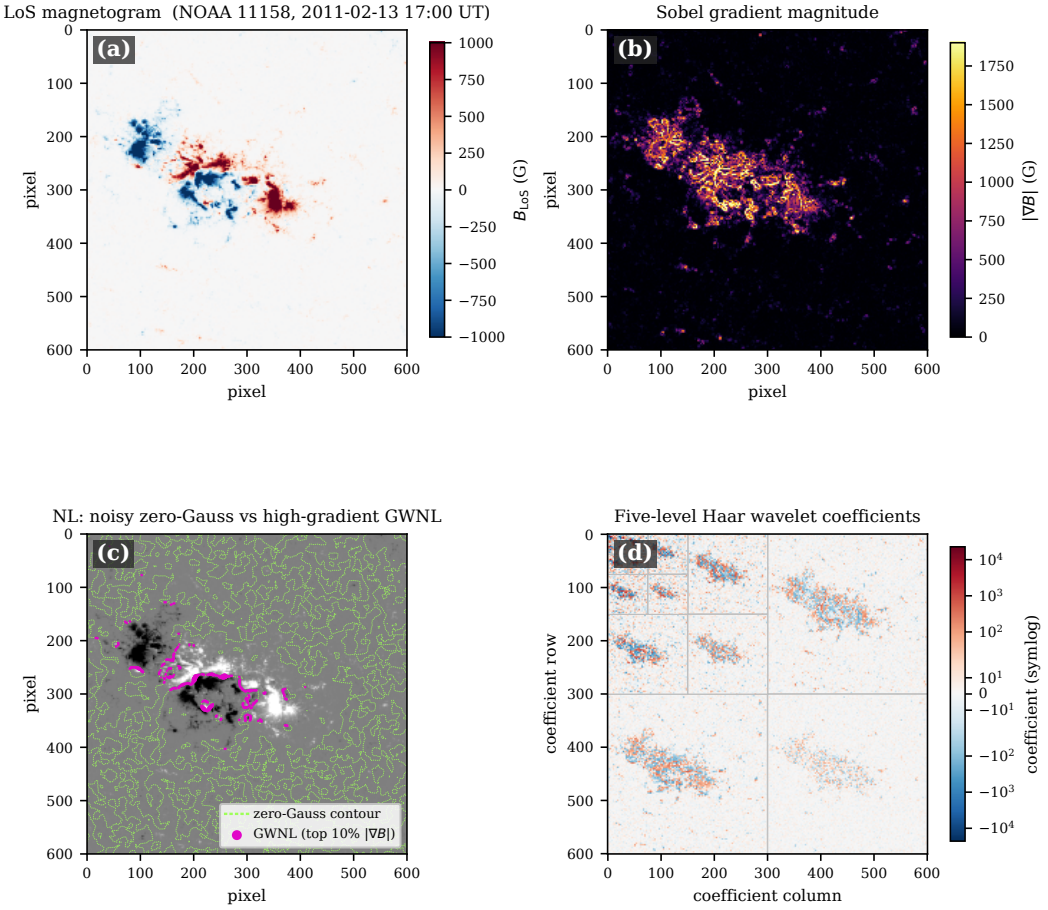


Figure 3. Feature extraction on NOAA AR 11158 at 17:00 UT on 13 February 2011 (FITS crop, 600 × 600 pixels), following Al-Ghraibah et al. [24]. (**a**) Line-of-sight magnetogram. (**b**) Sobel gradient magnitude. (**c**) Neutral-line contours and the gradient-weighted subset. (**d**) Five-level Haar wavelet coefficients.

2.6. Evolution Videos

For each AR directory, PNG frames are sorted chronologically and encoded as H.264 MP4 at 60 frames per second. Each AR produces one `AR{num}_evolution.mp4` file in the `movies/` directory. The 60 fps setting is a playback rate for visualization, not an observational cadence: consecutive frames remain the native HMI LoS crops sampled at approximately 12 min. At this encoding, one second of video spans about 12 h of solar time, so the MP4 files provide a time-compressed overview of AR evolution for inspection and qualitative analysis rather than a high-temporal-resolution data stream.

2.7. Quality Control and Expert Verification

Quality control is enforced throughout the pipeline. Flare lists are first screened to remove records with invalid AR identifiers, missing peak times, or unparseable GOES class strings. During active region extraction, crops are retained only when their centers remain on the usable disk and the AR longitude lies within the prescribed limit. Subsequent conversion and feature extraction stages record truncated or unreadable inputs in step-specific logs, skip empty files, and replace NaN values encountered during PNG normalization and FITS feature extraction (filled with 0 G). These automated checks operate together with the feature pipeline in Section 2.5 so that invalid flare records, unusable disk positions, and truncated or NaN inputs are excluded before the 29 magnetic parameters and flare labels are written.

In addition to these automated checks, solar physicists inspected a randomly selected 5% subset of the processed samples. Inspection criteria included crop centering, completeness of the field of view, and consistency of M/X class labels with the intended windows before eruption. Samples affected by telemetry gaps or uncommon processing failures were excluded or regenerated.

3. Results

3.1. Dataset Records and Statistics

Table 3 summarizes the scale of the JW-FD release. Here each magnetogram sample is one 600 × 600 active region crop; the remaining rows inventory the accompanying modalities, flare event records, and evaluation split.

Table 3. Summary statistics of the JW-FD release (2011–2025).

Quantity	Specification
Temporal coverage	1 January 2011–31 December 2025 (15 yr)
Active regions	3064 independent NOAA ARs
Magnetogram samples	1,991,247
Spatial resolution	600 × 600 pixels per AR crop
Released modalities	FITS, PNG, CSV, and MP4 (3064 videos)
Flare event records	12,298 NOAA XRA events (`flare_labels.csv`)
Evaluation split	AR level 8:1:1 (training/validation/test)
Data sources	SDO/HMI LoS; NOAA/SWPC SRS; NOAA Events (XRA)

Annual sample totals are strongly modulated by the solar cycle (Figure 4). Counts peak near the Cycle 24 maximum (2013–2014) and again during the Cycle 25 rise (2023–2025), whereas the 2018–2020 activity minimum yields far fewer NOAA numbered ARs and therefore fewer retained samples.

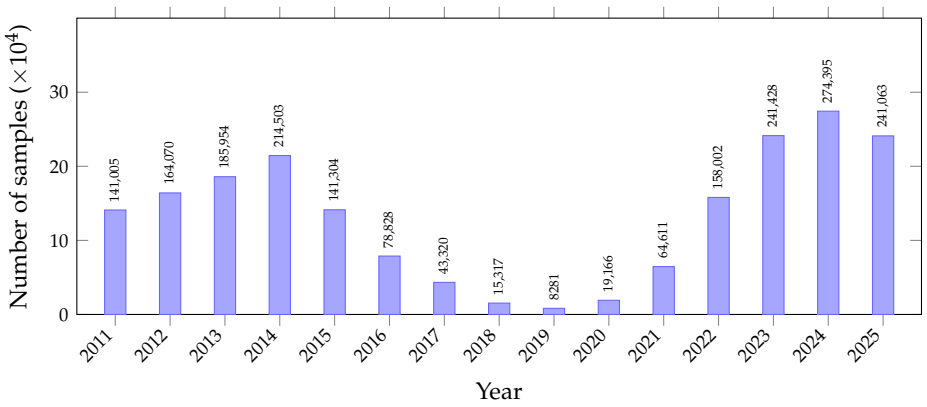


Figure 4. Annual distribution of magnetogram samples in JW-FD. Sample counts track solar activity, with peaks near the Cycle 24 maximum (2013–2014) and the Cycle 25 rise (2023–2025), and a clear minimum in 2018–2020.

Processed products share a common release root and are partitioned by modality (Table 4). Magnetogram and label directories are indexed by NOAA AR number; PNG branches further separate data by the saturation threshold B_{th}.

Each CSV row indexes one sample under a fixed 66-dimensional schema (Table 5). The same record can therefore support tabular learning, vision models, or multimodal fusion without reconstructing labels externally.

For evaluation, JW-FD provides an AR-level 8:1:1 random split with seed 62 (Table 6). All samples from one AR are assigned to a single subset; sample fractions may depart slightly from 8:1:1 because ARs contribute unequal numbers of magnetograms.

Table 4. Released data modalities and directory layout.

Modality	Content	Directory Layout
FITS	LoS magnetogram in Gauss; 600 × 600 crop	`fits_600/{ARnum}/...`
PNG	8 bit grayscale; saturation B_{th} encoded in path	`png_600_Th{B_th}/{ARnum}/...`
CSV	66 dimensional record per sample; FITS and PNG branches	`label/{fits|png}/Th{B_th}/...`
MP4	Per AR time compressed evolution (60 fps playback; ∼12 min source cadence)	`movies/AR{num}_evolution.mp4`
Labels	NOAA XRA flare event list	`flare_labels.csv`

Table 5. Composition of each 66-dimensional CSV record.

Component	*N*	Description
Magnetic features	29	Gradient, neutral line, wavelet, and flux descriptors (Table 2)
Binary flare labels	28	Seven forecast horizons × four GOES thresholds
Class labels	7	Strongest GOES class within each forecast horizon
File metadata	2	`image_filename` and `image_path`
Total	66	

Table 6. Active region-level partitioning into training, validation, and test subsets (nominal ratio 8:1:1; random seed 62).

Subset	Samples (*N*)	Fraction	ARs (*N*)	Fraction
Training	1,592,952	80.0%	2450	80.0%
Validation	193,375	9.7%	307	10.0%
Test	204,920	10.3%	307	10.0%
Total	1,991,247	100%	3064	100%

3.2. Magnetic Saturation Threshold Ablation

Converting Gauss-scale FITS magnetograms to 8-bit PNG requires a saturation level B_{th} in Equation (1). This choice affects visual contrast, feature statistics on PNG-derived branches, and model input distributions. JW-FD therefore releases PNG branches at ten saturation levels, $B_{\mathrm{th}} = 200$–2000 G in steps of 200 G, each accompanied by 29-dimensional magnetic features and the same configurable flare labels.

To provide an initial reference point, we ran a controlled ablation over six of these values (200–1000 G and 2000 G) with a Transformer-based vision classifier on the PNG branch, using binary flare grade ≥C1.0 and the fixed AR-level 8:1:1 split (random seed 62). Table 7 summarizes accuracy (ACC), true-positive rate (TPR), true-negative rate (TNR), true skill statistic (TSS), F1 score, Matthews correlation coefficient (MCC), precision, and false-alarm rate (FAR). In this internal study, $B_{\mathrm{th}} = 1000$ G attains the highest TPR (0.6320), TSS (0.5225), and MCC (0.2107), whereas 200 G is stronger on ACC, TNR, F1, Precision, and FAR.

We emphasize that this threshold selection is a preliminary exploration limited to the Transformer architecture and the ≥C1.0 binary task. The optimal saturation may differ for other models (CNN, LSTM, ViT), flare thresholds (M1.0, M5.0, X1.0), or forecast horizons. Because JW-FD releases PNG branches at all ten B_{th} values, users should treat 1000 G as a recommended starting point rather than a universal optimum and re-ablate when performance is critical.

3.3. Baseline Forecasting Experiments

To illustrate that JW-FD can support both parameter-based and vision-based models, we trained four baselines on the official AR-level 8:1:1 split (random seed 62) at $B_{\mathrm{th}} = 1000$ G: an RBF support vector machine on the 29 magnetic features, and ResNet-152,

ViT-B/16, and Swin-B classifiers on the PNG crops. The task is binary 24 h forecasting at GOES thresholds ≥C1.0 and ≥M1.0. Training used class-balanced losses; all active regions were retained, while magnetogram frames were temporally subsampled. Testing used the full 204,920-frame test partition at natural class prevalence. The decision threshold was selected on the validation set to maximize TSS.

Table 7. Forecast performance under different magnetic saturation thresholds B_{th} (Transformer, ≥C1.0, AR level split). For FAR, lower is better; for all other metrics, higher is better. Best values in each column are shown in bold.

B_{th}	ACC	TPR	TNR	TSS	F1	MCC	Precision	FAR
200 G	**0.9346**	0.3736	**0.9445**	0.3180	**0.1648**	0.1731	**0.1058**	**0.8942**
400 G	0.8955	0.5844	0.9010	0.4854	0.1620	0.2043	0.0940	0.9060
600 G	0.8997	0.5438	0.9059	0.4497	0.1577	0.1938	0.0923	0.9077
800 G	0.8877	0.6102	0.8926	0.5027	0.1581	0.2045	0.0908	0.9092
1000 G	0.8861	**0.6320**	0.8906	**0.5225**	0.1609	**0.2107**	0.0922	0.9078
2000 G	0.9254	0.3046	0.9363	0.2409	0.1236	0.1248	0.0776	0.9224

Table 8 reports test-set skill. For ≥C1.0, the three vision models attain TSS of order 0.53–0.54 and F1 of order 0.52–0.54, while SVM is slightly lower (TSS 0.50). For ≥M1.0, TSS is about 0.59–0.61, while precision and F1 are lower because positives are rarer. The area under the receiver operating characteristic curve (ROC-AUC) and the area under the precision–recall curve (PR-AUC) do not depend on the TSS-tuned decision threshold. ROC-AUC asks only whether flare frames receive higher scores than non-flare frames; it remains 0.82–0.85 for ≥C1.0 and 0.85–0.90 for ≥M1.0, so ranking is preserved as the class becomes rarer. PR-AUC asks how often a high score is a true flare, and therefore tracks prevalence: it falls from 0.49–0.57 to 0.15–0.19, in line with the drop in precision and the FAR above 0.88 for ≥M1.0. These C- and M-class scores are comparable in magnitude to earlier LoS-parameter studies [24]. Dedicated SHARP-based systems such as FlareCast [25] report higher test-set skill, and two points help to place that comparison in context. First, FlareCast takes as input sequences of SHARP radial-field (B_r) magnetograms (six frames spanning 6 h), rather than line-of-sight PNG crops, and it first regresses the maximum X-ray flux before assigning a flare class. Second, the 24 h test-set TSS of 0.891 for ≥C is accompanied by lower values for ≥M (0.792) and ≥X (0.831); in an operational window from 1 August to 24 November 2024, the four-class confusion-matrix diagonals are 0.58 (N), 0.57 (C), 0.63 (M), and 0.81 (X). These operational figures provide additional perspective on deployed skill.

Table 8. Baseline 24 h forecast skill on the official AR-level test set (204,920 frames; $B_{th} = 1000$ G). Vision models use PNG crops; SVM uses the 29 magnetic features. Decision thresholds are selected on the validation set to maximize TSS. ROC and PR denote ROC-AUC and PR-AUC (threshold-free). For FAR, lower is better; for all other metrics, higher is better. Best values in each task block are shown in bold.

Task	Model	ACC	TPR	TNR	TSS	F1	MCC	Prec.	FAR	ROC	PR
≥C1.0	SVM	0.7010	**0.8255**	0.6753	0.5008	0.4865	0.3838	0.3449	0.6551	0.8234	0.4900
	ResNet-152	**0.7826**	0.7532	**0.7886**	0.5419	**0.5431**	**0.4440**	**0.4246**	**0.5754**	0.8461	**0.5733**
	ViT-B/16	0.7599	0.7715	0.7575	0.5289	0.5243	0.4230	0.3971	0.6029	0.8432	0.5275
	Swin-B	0.7763	0.7632	0.7790	**0.5423**	0.5393	0.4405	0.4170	0.5830	**0.8507**	0.5657
≥M1.0	SVM	0.7721	0.7775	0.7719	0.5495	0.1580	0.2095	0.0880	0.9120	0.8506	0.1547
	ResNet-152	0.8173	0.7855	0.8182	0.6038	0.1913	0.2476	0.1089	0.8911	**0.8959**	0.1837
	ViT-B/16	0.7823	**0.8258**	0.7811	**0.6069**	0.1727	0.2339	0.0964	0.9036	0.8791	0.1450
	Swin-B	**0.8297**	0.7571	**0.8317**	0.5888	**0.1965**	**0.2483**	**0.1129**	**0.8871**	0.8891	**0.1946**

4. Discussion

4.1. Evaluation Protocol and Class Imbalance

JW-FD adopts an 8:1:1 random split at the AR level using random seed 62, ensuring that all magnetogram crops from a given AR appear exclusively in the training, validation, or test partition (Table 6). This protocol prevents information leakage arising from temporal autocorrelation within ARs [3,20]. Split CSV files (`*_train.csv`, `*_val.csv`, `*_test.csv`) are distributed under `label/png/Th{threshold}/` and `label/fits/`.

Positive labels are rare for higher flare thresholds (M5.0, X1.0), producing pronounced class imbalance. At longer forecast horizons the positive fraction increases, because a qualifying flare anywhere in a longer future window yields a positive label; shorter horizons are therefore more imbalanced. Over the full release, sample-level GOES class labels at the 72 h horizon comprise 15,094 X-class, 129,776 M-class, and 422,548 C-class events, together with 1,423,829 non-flare samples. Skill oriented metrics such as TSS, F1, and MCC therefore provide a more balanced assessment than raw accuracy alone [19].

4.2. Recommended Usage

JW-FD is intended for the heliophysics and space weather ML communities. Suggested applications include:

- Vision and video models: Train CNNs, ViTs, or temporal models on PNG crops or MP4 evolution sequences, selecting B_{th} by ablation (Section 3.2) and comparing against the 24 h baselines in Section 3.3.
- Parameter-based ML: Apply classical classifiers or gradient boosted trees to the 29 dimensional magnetic parameter vectors, comparing FITS-derived (Gauss scale) and PNG-derived (intensity scale) branches.
- Multimodal fusion: Combine PNG/MP4 inputs with magnetic parameters and labels from the same indexed record [6,7].
- Multi-horizon benchmarking: Evaluate models across seven forecast horizons and four GOES thresholds using the precomputed label columns.

The approximately 1.99 million magnetogram frames originate from 3064 active regions and are strongly autocorrelated in time within each region; they should not be treated as two million independent physical samples. The 24 h baselines in Section 3.3 make this concrete for snapshot models. Training retained every active region but temporally subsampled frames, and on the official AR-level test set an SVM on the 29 magnetic features is only modestly below ResNet-152, ViT-B/16, and Swin-B. JW-FD therefore supports parameter-based ML and single-frame vision on a similar footing; it is not a vision-only corpus, and the image count should not be read as an ImageNet-scale advantage. From $\geq$C1.0 to $\geq$M1.0 the TSS remains of order 0.50–0.61, whereas precision and F1 fall and the false-alarm rate exceeds 0.88, so a single-frame score is a poor operational alert for rarer classes.

Sequence and video models were not trained here, but they are a natural match to the same statistics: the within-AR autocorrelation that invalidates i.i.d. frame counts is the temporal structure those models are meant to use, and the release already provides tracked 12 min crops, seven forecast horizons, and timestamp-aligned PNG, CSV, and MP4 records. The MP4 files are a 60 fps playback of those native frames, not a high-cadence video stream, so video architectures should be treated as temporal models on the HMI cadence rather than as conventional video recognizers. In all of these settings the independent unit remains the active region.

Analyses that assume independent and identically distributed observations, that require cycle-to-cycle hold-out, that treat MP4 clips as independent videos, or that target the

rarest X-class events with fine-grained uncertainty from frame counts alone are less well supported by the nominal sample size. JW-FD is particularly suitable for AR-level generalization studies, multi-horizon and multi-threshold benchmarking, snapshot parameter or vision baselines, and sequence or multimodal experiments that exploit the aligned records. For rare-class tasks, we recommend reporting skill scores that are robust to imbalance (TSS, F1, MCC, together with precision or FAR), quantifying uncertainty at the active-region level, and, where appropriate, constructing chronological splits as noted in Section 4.3.

Beyond offline forecasting benchmarks, JW-FD can also support AI-driven solar observing systems that need flare-aware context or multimodal solar representations [26–28].

To reproduce or extend the dataset, users can clone the open-source pipeline at https://github.com/Xiaoxuan-1/JW-FD (accessed on 13 September 2026) and set `START_YEAR=2011`, `END_YEAR=2025` in `config.py`. We recommend using the distributed split files so that evaluations remain comparable across publications.

4.3. Limitations

JW-FD is restricted to HMI line-of-sight magnetograms and does not ingest vector-field products such as SHARP [5]. The source full-disk frames include limb-darkening correction, but no geometric projection correction is applied: the released 600×600 crops remain in the observer plane, and B_{LoS} is not converted to an approximate radial field. Foreshortening and the viewing-angle dependence of the LoS component therefore remain in both the images and the 29 derived features. The $|\mathrm{longitude}| \leq 60^{\circ}$ filter introduces a disk-center bias and excludes high-latitude and near-limb active regions; it mitigates the most severe limb effects but is not a substitute for heliographic or cylindrical equal-area remapping. AR identification depends on NOAA numbering and SRS daily positions, which can lag rapid flux emergence. The native HMI cadence of approximately 12 min may also undersample rapidly evolving pre-flare configurations. In addition, the recommended 8:1:1 split is a random allocation of active regions across the full 2011–2025 interval. It therefore tests generalization to unseen active regions under similar solar-cycle conditions, and it should not be interpreted as demonstrating generalization across solar cycles or to independent future epochs. The present coverage spans the declining phase of Cycle 24 through the rising and maximum phase of Cycle 25 and cannot by itself establish cycle-to-cycle generalization. Users interested in temporal generalization may construct chronological evaluation splits from the released timestamps.

5. Conclusions

JW-FD is a 15-year multimodal dataset for solar-flare forecasting. It comprises 3064 independent NOAA active regions and 1.99 million coregistered line-of-sight magnetogram crops spanning 1 January 2011 through 31 December 2025, released as aligned FITS, PNG, CSV, and MP4 products. PNG images, 29-dimensional magnetic features, and flare labels are provided at ten saturation thresholds from 200 to 2000 G, together with an AR-level 8:1:1 split that limits temporal leakage. JW-FD therefore supplies a large, leakage-controlled benchmark that links HMI magnetic observations with NOAA active-region and flare information across multiple prediction timescales and flare thresholds.

The dataset is intended as common, reusable infrastructure for traditional machine learning, deep learning, and multimodal methods. The baseline experiments show that the released records can support both parameter-based and vision-based models. Subsequent work can build on these records by exploiting the evolution videos, fusing image and parameter branches, incorporating vector magnetic-field products, and constructing chronological evaluation splits when cycle-to-cycle generalization is the scientific goal. We regard the present release as a foundation for that future investigation and optimization.

Author Contributions: Conceptualization, M.S. and J.L.; methodology, M.S.; software, M.S. and L.T.; validation, M.S., H.W. and Y.L.; data curation, M.S.; writing—original draft preparation, M.S.; writing—review and editing, M.S., J.L., H.W., L.T., C.Y., Y.Z. and Y.L.; visualization, M.S.; project administration, M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Astronomical Observatories, Chinese Academy of Sciences, under the "Seizing the High Ground of Science and Technology" Research Program (project title: AI Modeling and Analysis of Intense Solar Activity and Its Space Weather Effects), grant number E4TQ2101, and by the Specialized Research Fund for the State Key Laboratory of Solar Activity and Space Weather.

Data Availability Statement: The JW-FD construction pipeline is openly available at https://github.com/Xiaoxuan-1/JW-FD (accessed on 13 September 2026). A representative approximately 10% subset of the processed release (FITS crops, PNG frames at $B_{\mathrm{th}} = 1000$ G, movies, and labels, aligned at the active-region level) is deposited on Zenodo at https://doi.org/10.5281/zenodo.21672850. Processed data are organized under a release root containing `fits_600/`, `png_600_Th{threshold}/`, `label/`, `movies/`, and `flare_labels.csv`. The full dataset is planned for deposition at the National Astronomical Data Center (NADC); until then, interested researchers may contact the corresponding author.

Acknowledgments: Open SDO/HMI magnetograms and NOAA/SWPC flare products made this release possible; we gratefully acknowledge both teams. Support from the National Astronomical Data Center (NADC) is also acknowledged.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Shao, M.; Liu, S.; Xu, H.; Jia, P.; Wang, H.; Tong, L.; Bai, Y.; Yang, C.; Li, Y.; Li, N.; et al. Advances and Challenges in Solar Flare Prediction: A Review. *Astron. Tech. Instrum.* **2026**. [CrossRef]
2. Toriumi, S.; Wang, H. Flare-productive Active Regions. *Living Rev. Sol. Phys.* **2019**, *16*, 3. [CrossRef] [PubMed]
3. Park, S.H.; Leka, K.D.; Kusano, K.; Andries, J.; Barnes, G.; Bingham, S.; Bloomfield, D.S.; McCloskey, A.E.; Delouille, V.; Falconer, D.; et al. A Comparison of Flare Forecasting Methods. IV. Evaluating Consecutive-Day Forecasting Patterns. *Astrophys. J.* **2020**, *890*, 124. [CrossRef]
4. Boucheron, L.E.; Al-Ghraibah, A.; McAteer, R.T.J. Prediction of Solar Flare Size and Time-to-Flare Using Support Vector Machine Regression. *Astrophys. J.* **2015**, *812*, 51. [CrossRef]
5. Bobra, M.G.; Couvidat, S. Solar Flare Prediction Using SDO/HMI Vector Magnetic Field Data with a Machine-learning Algorithm. *Astrophys. J.* **2015**, *798*, 135. [CrossRef]
6. Shao, M.; Wang, H.; Li, Y.; Lin, J.; Liu, J.; Tan, B.; Guo, J.; Zhang, Y.; Huang, J.; Su, J.; et al. JW-Flare: Accurate Solar Flare Forecasting Method Based on Multimodal Large Language Models. *arXiv* **2025**. [CrossRef]
7. Shao, M.; Wang, H.; Tong, L.; Li, Y.; Wang, C.; Lin, J.; Liu, S.; Xu, H.; Zhang, Y.; Huang, J. JW-VL: A Vision-Language Model for Solar Physics. *Res. Astron. Astrophys.* **2026**, *26*, 095007. [CrossRef]
8. Boucheron, L.E.; Vincent, T.; Grajeda, J.A.; Wuest, E. Solar Active Region Magnetogram Image Dataset for Studies of Space Weather. *Sci. Data* **2023**, *10*, 825. [CrossRef] [PubMed]
9. Angryk, R.A.; Martens, P.C.; Aydin, B.; Kempton, D.; Mahajan, S.S.; Basodi, S.; Ahmadzadeh, A.; Cai, X.; Filali Boubrahimi, S.; Hamdi, S.M.; et al. Multivariate Time Series Dataset for Space Weather Data Analytics. *Sci. Data* **2020**, *7*, 227. [CrossRef] [PubMed]
10. Bobra, M.G.; Sun, X.; Hoeksema, J.T.; Turmon, M.; Liu, Y.; Hayashi, K.; Barnes, G.; Leka, K.D. The Helioseismic and Magnetic Imager (HMI) Vector Magnetic Field Pipeline: SHARPs—Space-Weather HMI Active Region Patches. *Sol. Phys.* **2014**, *289*, 3549–3578. [CrossRef]
11. Nishizuka, N.; Sugiura, K.; Kubo, Y.; Den, M.; Watari, S.; Ishii, M. Solar Flare Prediction Model with Three Machine-learning Algorithms Using Ultraviolet Brightening and Vector Magnetograms. *Astrophys. J.* **2017**, *835*, 156. [CrossRef]
12. Nishizuka, N.; Sugiura, K.; Kubo, Y.; Den, M.; Ishii, M. Deep Flare Net (DeFN) Model for Solar Flare Prediction. *Astrophys. J.* **2018**, *858*, 113. [CrossRef]
13. Abduallah, Y.; Wang, J.T.L.; Wang, H.; Xu, Y. Operational Prediction of Solar Flares Using a Transformer-based Framework. *Sci. Rep.* **2023**, *13*, 13665. [CrossRef] [PubMed]
14. Florios, K.; Kontogiannis, I.; Park, S.H.; Guerra, J.A.; Benvenuto, F.; Bloomfield, D.S.; Georgoulis, M.K. Forecasting Solar Flares Using Magnetogram-based Predictors and Machine Learning. *Sol. Phys.* **2018**, *293*, 28. [CrossRef]

15. Huang, X.; Wang, H.; Xu, L.; Liu, J.; Li, R.; Dai, X. Deep Learning Based Solar Flare Forecasting Model. I. Results for Line-of-sight Magnetograms. *Astrophys. J.* **2018**, *856*, 7. [CrossRef]
16. Li, X.; Zheng, Y.; Wang, X.; Wang, L. Predicting Solar Flares Using a Novel Deep Convolutional Neural Network. *Astrophys. J.* **2020**, *891*, 10. [CrossRef]
17. Sun, Z.; Bobra, M.G.; Wang, X.; Wang, Y.; Sun, H.; Gunturkun, B.; Camporeale, E.; Gombosi, T.; Hero, A.; Chen, Y. Predicting Solar Flares Using CNN and LSTM on Two Solar Cycles of Active Region Data. *Astrophys. J.* **2022**, *931*, 163. [CrossRef]
18. Jonas, E.; Bobra, M.; Shankar, V.; Hoeksema, J.T.; Recht, B. Flare Prediction Using Photospheric and Coronal Image Data. *Sol. Phys.* **2018**, *293*, 48. [CrossRef]
19. Bloomfield, D.S.; Higgins, P.A.; McAteer, R.T.J.; Gallagher, P.T. Toward Reliable Benchmarking of Solar Flare Forecasting Methods. *Astrophys. J. Lett.* **2012**, *747*, L41. [CrossRef]
20. Barnes, G.; Leka, K.D.; Schrijver, C.J.; Colak, T.; Qahwaji, R.; Ashamari, O.W.; Yuan, Y.; Zhang, J.; McAteer, R.T.J.; Bloomfield, D.S.; et al. A Comparison of Flare Forecasting Methods. I. Results from the "All-Clear" Workshop. *Astrophys. J.* **2016**, *829*, 89. [CrossRef]
21. Roy, S.; Hegde, D.V.; Schmude, J.; Lal, R.; Gaur, V.; Lin, A.; Mandal, K.; Singh, T.; Muñoz-Jaramillo, A.; Yang, K.; et al. SuryaBench: Benchmark Dataset for Advancing Machine Learning in Heliophysics and Space Weather Prediction. *Sci. Data* **2026**, *13*, 712. . [CrossRef] [PubMed]
22. Scherrer, P.H.; Schou, J.; Bush, R.I.; Kosovichev, A.G.; Bogart, R.S.; Hoeksema, J.T.; Liu, Y.; Duvall, T.L.; Zhao, J.; Title, A.M.; et al. The Helioseismic and Magnetic Imager (HMI) Investigation for the Solar Dynamics Observatory (SDO). *Sol. Phys.* **2012**, *275*, 207–227. [CrossRef]
23. Pesnell, W.D.; Thompson, B.J.; Chamberlin, P.C. The Solar Dynamics Observatory (SDO). *Sol. Phys.* **2012**, *275*, 3–15. [CrossRef]
24. Al-Ghraibah, A.; Boucheron, L.E.; McAteer, R.T.J. An Automated Classification Approach to Ranking Photospheric Proxies of Magnetic Energy Build-up. *Astron. Astrophys.* **2015**, *579*, A64. [CrossRef]
25. Lv, J.; Jia, P.; Shi, Y.; Song, Y.; Chen, F.; Guo, Y.; Liu, T.; Ji, K.; Jin, Z.; Lin, J.; et al. FlareCast: A Solar Flare Forecasting System Utilizing Deep Bayesian Neural Networks and the Concept of Machine Learning Operations. *Astrophys. J. Suppl. Ser.* **2026**, *283*, 46. [CrossRef]
26. Lin, J.; Tong, L.; Wang, H.; Shao, M.; Yang, C. Scientific-Intention Driven Embodied Intelligent Solar Telescope: Conceptual Design. *Laser Optoelectron. Prog.* **2026**, *63*, 0839014. [CrossRef]
27. Tong, L.; Lin, J.; Deng, Y.; Sun, Y.; Shao, M.; Wang, H.; Yang, C. JW-ASTClaw: A Generalizable Multi-Agent Framework for Autonomous Solar Telescope and Its Implementation within Chinese Meridian Project. *arXiv* **2026**. [CrossRef]
28. Tong, L.; Lin, J.; Hou, J.; Deng, Y.; Wang, D.; Liu, G.; Xu, S.; Ren, S.; Zhao, L.; Feng, Z.; et al. Design and Implementation of a Microservice-Architecture Master Control System for AIMS. *arXiv* **2026**. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.