

Generative Simulations of The Solar Corona Evolution With Denoising Diffusion : Proof of Concept

Grégoire Francisco^{1,2,3*}, Francesco Pio Ramunno^{4,5}, Manolis K. Georgoulis^{6,7}, João Fernandes⁸,
Teresa Barata² and Dario Del Moro¹

¹*Department of Physics, University of Rome Tor Vergata, Rome, Italy*

²*IA, Instituto De Astrofisica E Ciências Do Espaço, University of Coimbra, Coimbra, Portugal*

³*Department of Physics, University of Rome La Sapienza, Rome, Italy*

⁴*Institute for Data Science, University of Applied Sciences North Western Switzerland (FHNW), 5210 Windisch, Switzerland*

⁵*Department of Computer Science, University of Geneva, 1211 Geneva, Switzerland*

⁶*Johns Hopkins University Applied Physics Laboratory Laurel, MD 20735, USA*

⁷*Research Center for Astronomy and Applied Mathematics of the Academy of Athens, 11527 Athens, Greece*

⁸*CITEUC, Geophysical and Astronomical Observatory, University of Coimbra and Department of Mathematics, Coimbra, Portugal*

Abstract

The solar magnetized corona is responsible for various manifestations with a space weather impact, such as flares, coronal mass ejections (CMEs) and, naturally, the solar wind. Modeling the corona's dynamics and evolution is therefore critical for improving our ability to predict space weather. In this work, we demonstrate that generative deep learning methods, such as Denoising Diffusion Probabilistic Models (DDPM), can be successfully applied to simulate future evolutions of the corona as observed in Extreme Ultraviolet (EUV) wavelengths. Our model takes a 12-hour video of an Active Region (AR) as input and simulates the potential evolution of the AR over the subsequent 12 hours, with a time-resolution of two hours. We propose a light UNet backbone architecture adapted to our problem by adding 1D temporal convolutions after each classical 2D spatial ones, and spatio-temporal attention in the bottleneck part. The model not only produces visually realistic outputs but also captures the inherent stochasticity of the system's evolution. Notably, the simulations enable the generation of reliable confidence intervals for key predictive metrics such as the EUV peak flux and fluence of the ARs, paving the way for probabilistic and interpretable space weather forecasting. Future studies will focus on shorter forecasting horizons with increased spatial and temporal resolution, aiming at reducing the uncertainty of the simulations and providing practical applications for space weather forecasting. The code used for this study is available at the following link: https://github.com/gfrancisco20/video_diffusion.

1 Introduction

Key physical processes with a space weather impact occur within the solar corona. One example are the magnetic reconnection episodes resulting in solar flares (Priest et Forbes (2002)). Another one is the formation of coronal holes, known to be a determinant driver in the production of fast solar winds (Cranmer (2002), Wang (2024)). Reliable modeling of the solar corona's evolution is thus crucial for enhancing space weather forecasting capabilities. To that end, a significant body of research focuses on the extrapolation of coronal magnetic fields from photospheric magnetograms. To do so, many approaches leverage the fact that the dominant forces are magnetic and thereby model the magnetohydrodynamic (MHD) system as a magnetostatic one by neglecting all other forces (Wiegmann et Sakurai (2012)). However, such assumptions are less realistic at higher layers of the solar atmosphere, and the highly non-linear nature of coronal MHD processes like flares (Shibata et Magara (2011)), combined with the limitations of available boundary conditions, makes these models and other non-force-free ones computationally expensive and impractical for long-term forecasts. Moreover, the stochastic emergence of new magnetic flux from the solar interior adds further complexity to these models (Cheung et Isobe (2014)).

In this work, we explore the potential of generative deep learning methods to efficiently model such complex systems. Generative latent variable models have demonstrated their strong abilities in learning the mapping of high-dimensional data distributions (e.g., images or videos) onto latent probabilistic spaces, allowing for realistic sampling and extrapolation (Asperti et Tonelli (2023)). Among such methods Variational Auto Encoder (VAE)s (Kingma et Welling (2013)) offer computational efficiency but tend to struggle with high-frequencies and small-scale details. Generative Adversarial Network (GAN)s (Goodfellow *et al.* (2014)) partially address these shortcomings by producing higher-quality samples, but are difficult to train and often suffer from mode collapse, leading to poor diversity in generated data. Recently, Denoising Diffusion Probabilistic Model (DDPM)s

*Contact Author : gregoire.francisco@gmail.com

have been shown to consistently outperform both VAEs and GANs in terms of sample quality and diversity across several applications, including image generation, super-resolution, and image-to-image translation (Ho *et al.* (2020)). The stochastic generation process of diffusion models makes them particularly suitable for capturing the intrinsic stochastic nature of non-linear physical systems (Sohl-Dickstein *et al.* (2015)). Recent studies showed promising applications in the field of space weather. Ramunno *et al.* (2024a) successfully used DDPMs to model a latent distribution of solar EUV observations, and generate realistic new synthetic samples. Ramunno *et al.* (2024b) showed that DDPMs can forecast photospheric magnetogram 24 hours before flares, outperforming persistence models both in computer science and physical performance indicators. In this study, we aim to extend the application of DDPMs to video-to-video forecasting of Solar Dynamic Observatory (SDO)/Atmospheric Imaging Assembly (AIA) observations at the 94Å wavelength. The 94Å channel is particularly important for studying solar flares, as it is most sensitive to plasma temperatures around 6.3 million K, which are closely associated with flare activity Boerner *et al.* (2012). Previous work has shown that emissions in the 94Å and 131Å channels are strongly correlated with the peak flux of Soft X-Rays (SXR) during flares, and can provide reliable proxy for flare magnitude in SXR (van der Sande *et al.* (2022)). Reliable probabilistic simulations of the corona evolution in these wavelength could thus aid in both flare magnitude estimation and flare timing prediction, the latter being one of the hardest challenges in space weather forecasting (Boucheron *et al.* (2015)). The 94Å wavelength and a 2-hour temporal resolution used in this study serve as a proof of concept to illustrate the potential of the approach for space weather forecasting and solar physic studies. Future works will focus on extending this approach to additional wavelengths and higher temporal resolution to reduce the resulting simulations uncertainty and provide practical space weather applications. To reduce computational complexity while preserving the physical fidelity of active region (AR) dynamics, we limit our analysis to ARs within ± 230 arcsec from the solar disk center, thus minimizing the effects of projection and solar rotation. This enables to highlight the model’s ability to learn the intrinsic physical dynamics of ARs. The paper is organized as follows: Section 2 describes the model, starting with an introduction to key DDPM concepts followed by the description of our light UNet backbone to learn temporal patterns from videos. Section 3 presents the dataset used for training and evaluation. Section 4 discusses the performances, while Section 5 concludes with a discussion of our findings and future works.

2 Model

2.1 Background

Denoising Diffusion Probabilistic Model (DDPM)

A diffusion model (Sohl-Dickstein *et al.* (2015)) p_θ is defined as a reverse process modelling a variable X_0 as the $T - th$ state of a Markov Chain $(X_{T-t})_{t=0}^T$ of transitions defined

such as :

$$\begin{aligned} p_\theta(X_T) &:= \mathcal{N}(X_T; 0, I) \\ p_\theta(X_{t-1}|X_t) &:= \mathcal{N}(X_{t-1}; \mu_\theta(X_t, t), \Sigma_\theta(X_t, t)) \end{aligned} \quad (1)$$

In the DDPM case, the unknown transitions $p_\theta(X_{t-1}|X_t)$ are modeled as the reverse of a forward diffusion of Gaussian transitions. This forward process q gradually adds noise to X_0 , following a scheduled variance $(\beta_t)_{t=1}^T$, that results in equilibrium at T , thus reaching the pure noise state X_T :

$$q(X_t|X_{t-1}) := \mathcal{N}(X_t; \sqrt{1 - \beta_t}, \beta_t I) \quad (2)$$

This diffusion process can be resolved at any arbitrary timestep t , defining $\alpha_t := 1 - \beta_t$ and $\hat{\alpha}_t := \prod_{i=0}^t \alpha_i$, with :

$$q(X_t|X_0) := \mathcal{N}(X_t; \sqrt{\hat{\alpha}_t}X_0, (1 - \hat{\alpha}_t)I) \quad (3)$$

An autoencoder ϵ_θ can then be trained to learn predicting a random noise $\epsilon \sim \mathcal{N}(0, I)$, added during the diffusion process at timesteps t following the Equation 3, i.e. $x_t := \sqrt{\hat{\alpha}_t}x_0 + \sqrt{1 - \hat{\alpha}_t}\epsilon$. Specifically, the training is performed by minimising a distance - typically L2 - between the actual noise and its autoencoder’s estimations $\epsilon_\theta(x_t, t)$:

$$\min \|\epsilon - \epsilon_\theta(x_t, t)\|_2, \forall t \in \llbracket 1, T \rrbracket \quad (4)$$

Once ϵ_θ trained, the predicted noise can be used to approximate $\mu_\theta(X_t, t)$, the mean of the reverse process, as derived in Ho *et al.* (2020):

$$\frac{1}{\sqrt{\hat{\alpha}_t}}(x_t - \frac{\beta_t}{\sqrt{1 - \hat{\alpha}_t}}\epsilon_\theta(x_t, t)) \quad (5)$$

The variance $\Sigma_\theta(X_t, t)$ can be fixed to β_t to match the forward process’s variance schedule. Using these values in Equation 1, new samples x_0 can be generated by iteratively reversing the diffusion process, starting from pure random noises $x_T \sim \mathcal{N}(0, I)$. The computational cost of this sampling procedure depends on the noise schedule $(\beta_t)_{t=1}^T$, which must be chosen so that the Signal-To-Noise-Ratio (SNR) becomes null in T, i.e. $SNR_T = \frac{\hat{\alpha}_T}{1 - \hat{\alpha}_T} \approx 0$, in order for X_T to be pure noise. In the case of the linear noise schedule proposed by Ho *et al.* (2020), such condition start being satisfied from $T \approx 1000$ as $SNR_{1000} \approx 4e - 05$.

Conditional Diffusion

Finally, all the equations of this section can be extended to handle conditional diffusion by marginalizing the probability distributions over any conditioning variable c . In our case the conditions c represents the previous 12 hours of input data. These conditions are projected by the autoencoder into a continuous latent variable space, allowing the model to generalize to unseen conditions during training and generating predictions for new video sequences. Specifically, for any c representing 12 hours of an active region at 2 hours resolution, our models learn to simulate possible realisations of the next 12 hours X_0 , with a latent space approximating the conditional transitions:

$$p_\theta(X_{t-1}|X_t, c) := \mathcal{N}(X_{t-1}; \mu_\theta(X_t, t, c), \Sigma_\theta(X_t, t)) \quad (6)$$

2.2 Light Video UNet

For the autoencoder, we use a modified version of the Palette’s UNet architecture (Saharia *et al.* (2021)) also used by Ramunno *et al.* (2024b) for magnetogram-to-magnetogram forecasting. In our case, to adapt the model to videos and improve its ability to capture temporal patterns, we add temporal 1D-convolutions after each spatial 2D-convolutions. This approach provides spatio-temporal feature learning abilities similar to 3D-convolutions but with lower computational complexity as the complexity of a 2D+1D convolutions block is $O(n^2 + n)$ against $O(n^3)$ for 3D-convolutions, with n , the size of the kernels. For more complex patterns, we incorporate spatio-temporal self-attention (Vaswani *et al.* (2017)) blocks between each convolution block in the bottleneck. As the feature-maps size is smallest in the bottleneck - 16x16 in our model -, this placement allows for attention scores computation at the smallest possible complexity.

3 Dataset

For the of AIA Extreme Ultraviolet (EUV) images we start from the SDO-2H-ML dataset that we defined in Francisco *et al.* (2024). Specifically, the images are aligned, exposure normalised, instrument degradation corrected, downsampled at 1024 by 1024 pixels and available at a 2 hour resolution. Their pixel values are logarithmically scaled and depth-downsampled to 8-bit, which allow to preserve most of the original pixel distribution with marginal information loss, while making the dataset more compact. In this work we use the resulting images before the original JPEG compression used in Francisco *et al.* (2024). As we seek to test the model’s ability to learn and simulate Active Region (AR)s’ dynamics, while limiting the impact of the solar rotation and projection effects, we focus on crops of 614 by 614 arcsec centered around known SHARPs (Bobra *et al.* (2014)) that are within ± 230 arcsec from the solar disk center. The important size of our crops, relatively to original SHARPs bounding boxes, is meant to ensure capturing all the coronal loops that can be associated to a given AR. On the other end, this approach imposes us a strict chronological split between training and test data, as some of our crops may partially overlap over each other. The resulting crops are further downsampled to 128 by 128 pixels for computational reasons. Samples are created at a 2 hour cadence when possible by pairing the crops of a given AR within [-10h,0h] for the input video, with the crops within [+2h, +12h] for the targets, resulting in pairs of 6 frames videos. While our target is only the 94Å wavelenght, we add the 193Å and 211Å wavelength in the input to benefit from more information on coronal dynamics at different temperatures. We train on the resulting samples between 2010-05 and 2022-03, which amount to a total of 34000 paired videos. We test on samples spanning from 2022-05 to 2023-04 that are buffered by at least one month from the training data. Due to our spatial and temporal constraints, our test samples do not exhibit any X-flares, and only 41 samples exhibit M-flares. To build a a balanced test, we randomly select 41 C-flares, and 41 samples without any flare above the C-threshold, such sample will be referred

a quiet samples. This results in a small test of 123 samples that allow to estimate our probabilistic model performances at a moderate cost.

4 Results

4.1 Computer vision metrics

To evaluate the quality of the generated videos, we propose using a set of classical computer vision metrics: Peak Signal-to-Noise Ratio (Peak Signal To Noise Ratio (PSNR)), Structural Similarity Index (Structural Similarity Index Measure (SSIM)), and Learned Perceptual Image Patch Similarity (Learned Perceptual Image Patch Similarity (LPIPS)). Although these metrics are typically designed for image comparison, we apply them frame-wise and calculate their averages to derive an overall score for the video. The PSNR compares the pixel-wise mean squared error to the maximum possible value of the signal, normalizing the distortion of the generated videos by the potential dynamic range. In the case of Extreme Ultraviolet (EUV) observations, the signal can vary across several orders of magnitude, corresponding to different physical processes. To more accurately assess the model’s ability to preserve the underlying physics, we apply this metric to the descaled pixel distribution, reverting it to its original dynamic range. Consequently, the resulting PSNR scores appear much higher than those typically computed over standard 8-bit images. For context, we refer to Chervyakov *et al.* (2020), which derived a threshold formula indicating that a PSNR value above $Q = 5 * \log(\max_{signal-value}) / \log(2)$ suggests low distortion and high-quality images. With our saturation value of 6099 DN/s for the 94 Å EUV wavelength, this roughly corresponds to $Q \approx 63$. However, pixel-wise metrics like PSNR may not be the most relevant in our case. Given the non-deterministic nature of our problem, it is more suitable to compare the presence of similar structures in the images with slightly relaxed conditions regarding their pixel-wise shape and locations. To this end, the Structural Similarity Index (SSIM) is particularly relevant, as it evaluates a combination of indicators derived from the first and second-order moments of the pixel distribution over a small window - in our case, 11 by 11 pixels. Images are considered to have good similarity for SSIM values above 0.7. Similar to PSNR, we compute SSIM using the original (descaled) pixel dynamic range. Finally, for a more intuitive interpretation of perceptual similarity between the images, we use LPIPS (Zhang *et al.* (2018)) directly on the 8-bit generated videos, with values below 0.1 indicating good similarity.

4.2 Space weather related metrics

To further evaluate the model’s ability to capture key physical quantities, we complement the previous metrics with physical metrics derived from the EUV flux measured within the frames. While an estimate of the actual EUV flux can be obtained using the corresponding instrument’s response function (Boerner *et al.* (2012), Barnes *et al.* (2020)), we simplify our metrics by using the DN/s values as proxies for the flux. This simplification is particularly relevant, as van der Sande *et al.* (2022) demonstrated that DN/s values from AIA)observations can accurately approximate Soft X-Ray

(SXR) flux, which is of particular interest in space weather forecasting. Consequently, for each frame, our proxy for the corresponding region’s flux is the sum of the pixel DN/s values. From the resulting flux time series, we focus on two essential derived quantities: the Maximum Peak Flux (MPF) and the fluence. The MPF is defined as the maximum value of the flux within the entire time series. It is particularly relevant as it provides an estimate of the maximum activity expected during the forecasted time window. This metric is well suited to the stochastic nature of the problem, as we anticipate some uncertainty regarding the precise timing of flare occurrences, especially when forecasting over several hours. To complement the MPF, we introduce the Time To Peak Flux (T2PF), which indicates the time remaining before reaching the MPF and poses a particularly challenging problem in flare forecasting (Boucheron *et al.* (2015)). The fluence is defined as the integral of the flux over the considered period, approximated by:

$$fluence = \sum_t (\Delta_t * flux_t) \quad (7)$$

where Δ_t corresponds to our 2-hour resolution converted to seconds. Due to our relatively low 2-hour time resolution, the resulting MPF and fluence values are not meaningful approximations of the actual physical values for the corresponding time windows. Therefore, these metrics are not yet intended for practical applications but serve as indicators of the model’s ability to capture key physical quantities at a given resolution. Specifically, we focus on the model’s Mean Absolute Percentage Error (MAPE) in MPF and fluence, as well as the Mean Absolute Error (MAE) in T2PF. Finally, we evaluate the model’s ability to generate realistic trajectories of the flux time series by computing the Dynamic Time Warping (DTW) Euclidean distance (Piersol (1981)) between the generated time series and the ground truth. The DTW metric allows us to estimate the similarity between two time series while correcting for the impact of small delays or variations in speed. This feature enables better comparisons when the sequences may be out of phase, as is the case here. To enhance interpretability, we focus on the DTW distance normalized by the Euclidean norm of the observed time series.

4.3 Performances

The performance metrics of the models are presented in Table 1. For each sample 20 simulations are generated, and the final score per sample is calculated as the average score over all the simulations. The good scores in the first three columns, corresponding to the computer science metrics, suggest that the results are perceptually realistic. Both PSNR and LPIPS show slight degradation during periods when stronger flare events occurred within the forecasting window, indicating greater uncertainty at the pixel level and a lower perceptual similarity, respectively. On the other hand, the SSIM improves, suggesting that the model performs better at a structural level, as the metric compare the images similarity over small local windows of a few pixels. This apparent contradiction between the PSNR and SSIM may stem from the model’s ability to predict accurately brightenings, such as flares, while exhibiting some uncertainty about their precise pixel-wise shape and location. During higher solar activity, such brightenings constitute a larger portion of the overall signal, leading to the increased

SSIM. The last four columns indicate that the model successfully captures essential physical properties. For instance, the mean absolute percentage error for both MPF and fluence is approximately 40%. When breaking down the performance by activity levels, the errors remain within the same order of magnitude of the observed values. While promising for practical space weather applications, it is noted that the low temporal resolution of this study smooth the flux time series, simplifying the task and rendering the model not suitable for operational space weather forecasting, as most flares’ peak do not appear in the training and evaluation data.

From the 20 simulations generated for each sample, estimates of the fluence, MPF, and T2PF distributions were derived. Their reliability are assessed with the reliability diagrams shown in Figure 1. The blue curves depict how frequently the ground truth falls within the corresponding Confidence Interval (CI). For a perfectly reliable probabilistic model, these curves would align with the diagonal ($y = x$), meaning that observations would fall within the $p\%$ -CIs $p\%$ of the time for any p in the range $[0,100]$. The red bars indicate the size of the CIs, reflecting the level of uncertainty at a given confidence level. For practical purposes, smaller uncertainties are preferable, with the ideal model having CIs as small as possible.

For all three metrics, the generated probability distributions appear reliable, with the blue curves aligning with the diagonal at a few point of percentage. This suggests that the model effectively captures the stochastic nature of coronal evolution, which arises from the system’s inherent non-linearity and the stochastic emergence of new flux. The model’s resulting stochasticity is then further emphasized by the limited input information, such as the low spatial and temporal resolutions. As a result, the uncertainty of the generated simulation is quite large, with 96%-CIs reaching approximately $\pm 55\%$ of the actual observation for the fluence and the MPF, and $\pm 4H$ for the T2PF.

4.4 Prediction examples

To illustrate the diversity of the model’s predictions, we present three simulations for each of the following cases:

- AR 8195 on 2022-05-04 at 20:00, shown in Figure 2
- AR 8977 on 2023-01-18 at 06:00, shown in Figure 3
- AR 9188 on 2023-03-08 at 22:00, shown in Figure 4

For each figure, the first row displays the sequence of frames from the 12 hours preceding the forecasting date. The second row corresponds to the actual frames for the next 12 hours, which the model aims to simulate. The third to fifth rows present three distinct simulations generated by the model. The sixth row shows the simulations’ percentage deviation from the ground truth, averaged over 20 distinct simulations. The seventh and final row displays standard deviation maps, computed as the pixel-wise standard deviation across the same 20 simulations, highlighting the regions of the images where uncertainty is highest.

Comparing the mean percentage error across the three examples, the deviation from the ground truth appears balanced in the case of AR 8195 (Figure 2), while the simulations for

Models	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	Flux-DTW	MPF (MAPE)	Fluence (MAPE)	T2PF (MAE)
DDPM	73 ± 2.4	0.77 ± 0.13	0.09 ± 0.02	1.00 ± 0.57	42% ± 25	41% ± 24	4.84H ± 2.53
DDPM (quiet)	79 ± 3.1	0.67 ± 0.15	0.07 ± 0.03	0.99 ± 0.64	38% ± 0.27	40% ± 0.26	4.93H ± 2.65
DDPM (C)	72 ± 2.3	0.80 ± 0.13	0.08 ± 0.02	0.93 ± 0.55	38% ± 0.23	38% ± 0.23	5.31H ± 2.64
DDPM (M)	67 ± 1.9	0.84 ± 0.11	0.11 ± 0.03	1.09 ± 0.53	49% ± 0.24	44% ± 0.23	4.29H ± 2.31

Table 1: Models Performances. The first row corresponds to the model’s performances over the all test set presented in section 3. The three following row indicates the performances of the model on the restriction of samples which exhibited the following activity during the forecasted window : quiet for the second row, at least one C-class flare for the third row, at least one M-class flare for the fourth row. The first three columns are computer science metrics (section 4.1) where \uparrow indicates that higher scores are better, and \downarrow that lower scores are better. The last columns correspond display the error in the physical metrics presented in section 4.2. For every sample 20 simulations are generated. The first rows of the cells indicate the average score over all the simulations. The second rows of the cells (\pm symbol) indicate the average standard deviation of the sample in the corresponding scores.

Reliability Diagrams

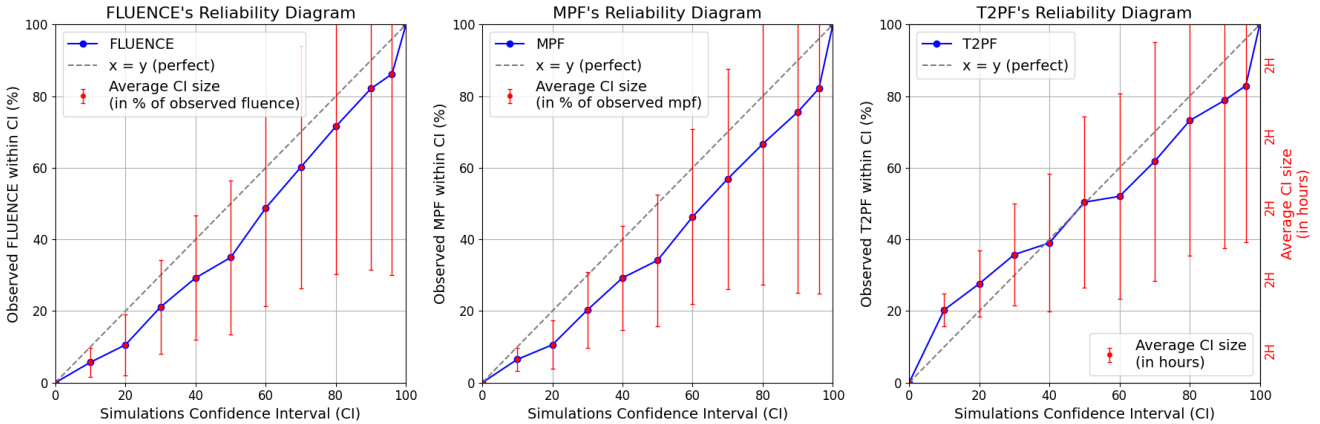


Figure 1: Reliability diagrams. The X-axis represents the Confidence Intervals (CIs) derived from the simulations. The left Y-axis represents the frequency at which actual observations fall into the derived CIs, corresponding to the blue curve. The red bars represent the size of the CIs, expressed as a percentage of the observed value for fluence and MPF, with one tile on the Y-axis grid representing 20% uncertainty. For the T2PF, the red bar scale (right Y-axis) is in hours, with each tile representing 2H of uncertainty. A perfectly reliable model has a blue curve aligned with the diagonal and provides minimal uncertainty for a given confidence interval, represented by smaller red bars.

AR 8977 (Figure 3) and AR 9188 (Figure 4) displays in average an undercasting tendency. The standard deviation maps highlight regions that are estimated to be at higher risk for extreme events. They mostly coincide with the areas and timesteps exhibiting the most intense activity in the ground truth. This suggests that, while the model does not assign a high probability to extreme events (indicated by the negative mean deviation), it captures their possibility with uncertainty around their precise timing and exact configuration.

The case of AR 8977 (Figure 3) is particularly illustrative. An M1.8 flare began at 10:21 and ended at 10:52, 1 hour and 8 minutes before the +6H frame, where the ground truth exhibits strong remnants of the flare. A very similar brightening is observed in the +6H frame of Simulation 0 but not in the other two simulations. Across the 20 generated simulations, this brightening appears twice at the same frame, resulting in an estimated probability of occurrence around 10% at that

specific time step. As a result, the standard deviation map is brightest for that frame and in the region of the flare.

Similar observations can be made for AR 9188 (Figure 4), which exhibits a bright flare remnant in the +12H frame. In this case, the model generates similar brightening in the +10H frame for Simulations 0 and 1, and in the +8H frame for Simulation 3. This suggests that the model is relatively confident about the occurrence of such events within the final hours of the forecasted time window but remains uncertain about the precise timing. This uncertainty is also highlighted by the bright standard deviation maps for the +8H to +12H frames.

Finally, AR 8195 is marked by a series of several C-class and M-class flares during both the forecasted time window and the 12 previous hours. This high level of activity results in complex coronal loop dynamics and configurations, which appear well understood by the model. Indeed, the model generates realistic and diverse possible evolutions for the active

Predictions example on AR 8195 the 2022-05-04 at 20:00

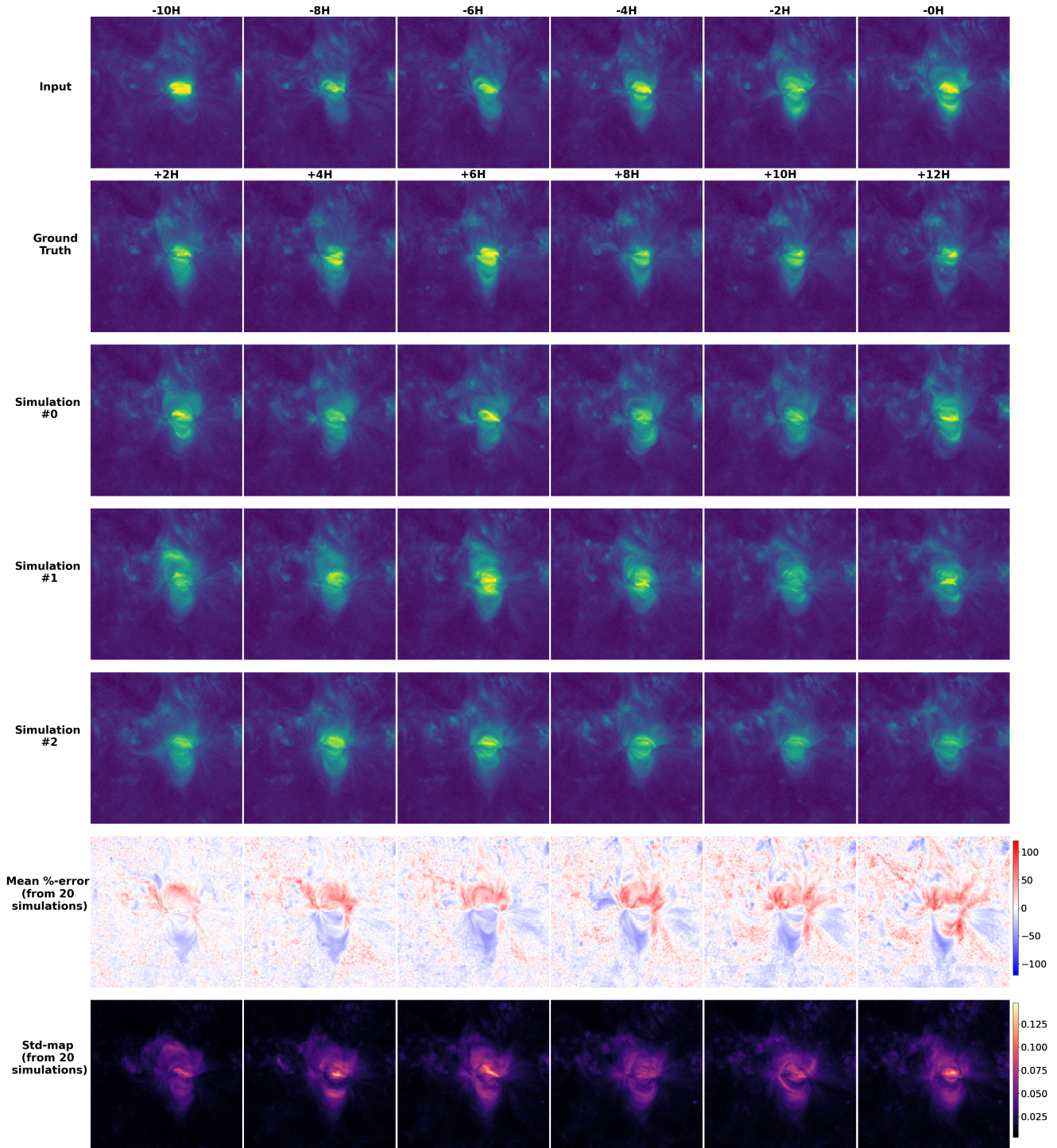


Figure 2: Predictions for AR 8195 on 2022-05-04 at 20:00. The AR exhibited a series of numerous C-class and M-class flares during the 12 hours prior to the prediction time, as well as during subsequent 12 hours. Remnants of these flares can be observed as brightenings in the frames, both in the first row (previous 12 hours) and in the second row (next 12 hours ground truth). Similar brightening can be observed in the three simulation results exhibited from the third to fifth row. The sixth row presents the pixel-wise percentage deviation from the ground truth, averaged over 20 simulations. The final row displays standard deviation maps computed over the same 20 simulations. Animations links : [Simulation #0](#), [Simulation #1](#), [Simulation #2](#).

region, illustrating again the strength of the DDPM in modeling the stochastic aspects of the problem.

5 Discussion & Conclusions

Despite the constraints of using a low temporal resolution (2 hours) and focusing only on ARs at the solar disk center, our current approach demonstrate promising potential. In particular, our findings suggest that DDPMs can be used to efficiently generate realistic evolutions of the solar corona as observed in EUV. Our probabilistic video forecasting model appears well-calibrated and reliable for deriving key space weather indicators. As seen in our prediction examples, the model is able to simulate complex MHD dynamics and forecast brightening events with time steps closely aligned to their actual occurrences. These encouraging outcomes suggest such models could represent a major milestone in the fields of Solar Physics and Space Weather. While current space weather forecasting abilities are sometimes compared to earth weather forecasting abilities of 50 years ago, these new methods could significantly reduce that gap, by providing computationally efficient simulation tools to visualize the possible trajectories of the system's evolution. Major limitations of this proof of concept must however be acknowledged. The low temporal resolution of 2 hours, statistically excludes most actual flare peaks from the training and evaluation data, making the resulting model not suited to model the occurrence of such event. The temporal and spatial restrictions of this work were intended solely to validate the concept with limited computational resources. A time resolution of 15 to 10 minutes could already allow to capture much better extreme transient fluctuations in variables of interest like the flux, resulting in improved flare predictions with reliable likelihood estimates and interpretable simulations of their occurrence. At a time resolution of 2 minutes or less, a DDPM might further model the physical dynamics of the corona at a much finer scale, potentially resulting in reduced uncertainty and promising application for solar physics studies. For instance, such models could be used to estimate the intrinsic stochasticity of flare occurrences given current instrument limitations in spatial and temporal resolution. Moving forward, future work could thus focus on generating forecasts for 6-hour periods at a 10-minute resolution and for 1-hour periods at a 2-minute resolution. Additionally, increasing the spatial resolution to 2.4 arcsec/pixel might also be worth exploring, while training the model at any longitudes for full-disk coverage will be indispensable for operational applications.

Acknowledgements

This research is part of the SWATNet project which is funded by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Grant Agreement No [955620](#).

This study was also produced within the IA and the CITEUC. IA is supported by Fundação para a Ciência e a Tecnologia (FCT, Portugal) through the research grants [UIDB/04434/2020](#) and [UIDP/04434/2020](#). CITEUC is funded by National Funds through FCT - project UIDP+UIDB/00611/2019.

Project partially funded under the National Recovery and Resilience Plan (PNRR), Missione 4 "Istruzione e Ricerca" – Componente C2 – Investimento 1.1, "Fondo per il Programma Nazionale di Ricerca e Progetti di Rilevante Interesse Nazionale (PRIN)" – Call for tender No. 1409 of 14/09/2022 of Italian Ministry of University and Research funded by the European Union – NextGenerationEU Award Number: P2022RKXH9, Concession Decree No. 1397 of 06/09/2023 adopted by the Italian Ministry of University and Research, project CORonal mass ejection, solar eNERgetic particle and flare forecaSTing from phOtospheric sigNaturEs (CORNERSTONE).

Predictions example on AR 8977 the 2023-01-18 at 06:00

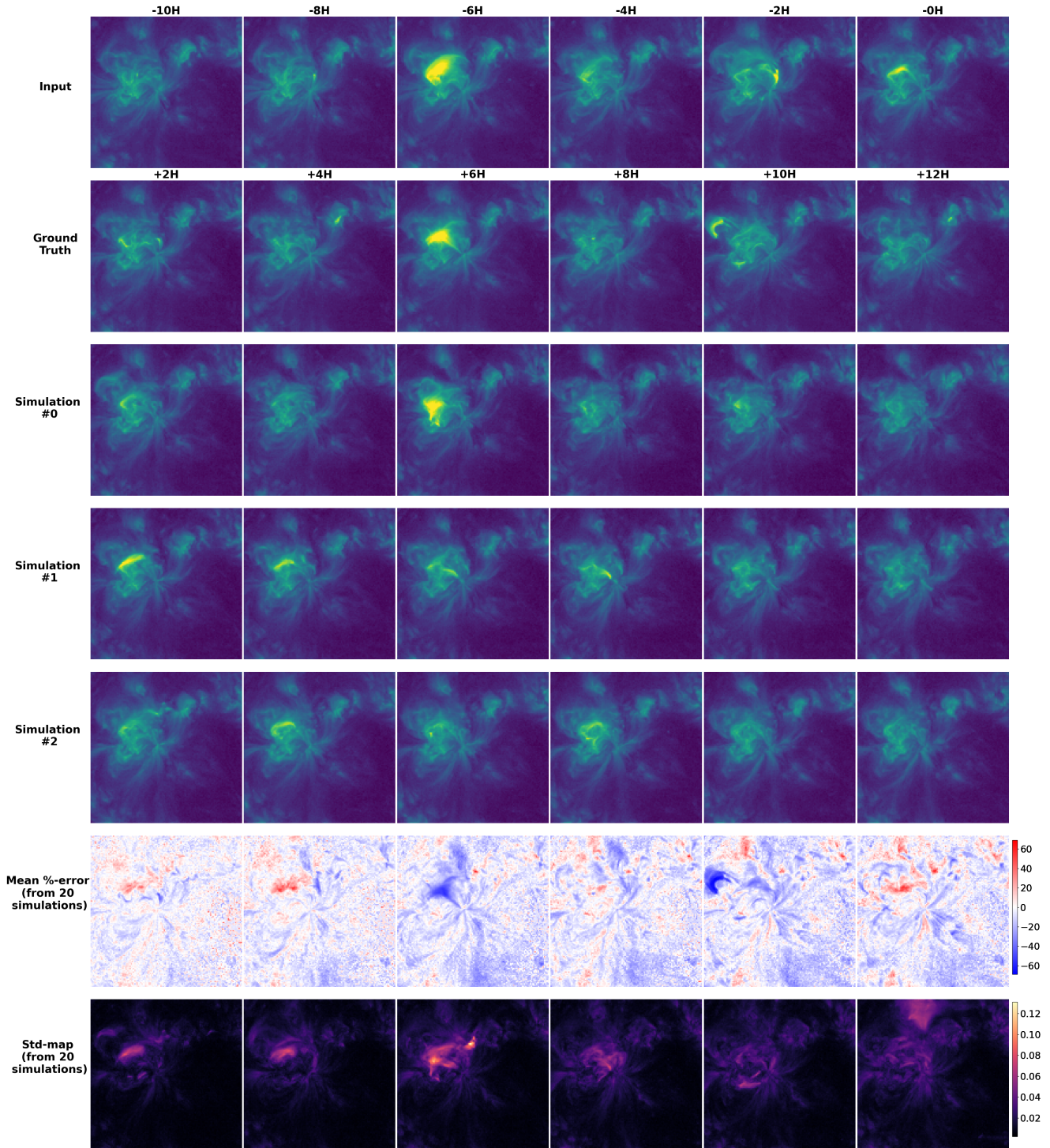


Figure 3: Predictions for AR 8977 on 2023-01-18 at 06:00. The -6H input frame shows the remnants of an M1.8 flare that occurred approximately 6.5 hours prior to the time of prediction. In the +6H ground truth frame, which represents the forecasted target, we observe the remnants of another M1.8 flare that concluded on 2023-01-18 at 10:52. A similar brightening is visible in the +6H frame of Simulation 0. The mean percentage pixel-wise error, averaged over 20 simulations, indicates that errors are larger during this event. The standard deviation map shows the pixel-wise standard deviation across 20 simulations, providing insight into the areas where the model is most uncertain. This highlights regions at higher risk for extreme events, such as the flare remnants seen in the +6H frame. Animations links : [Simulation #0](#), [Simulation #1](#), [Simulation #2](#).

Predictions example on AR 9188 the 2023-03-08 at 22:00

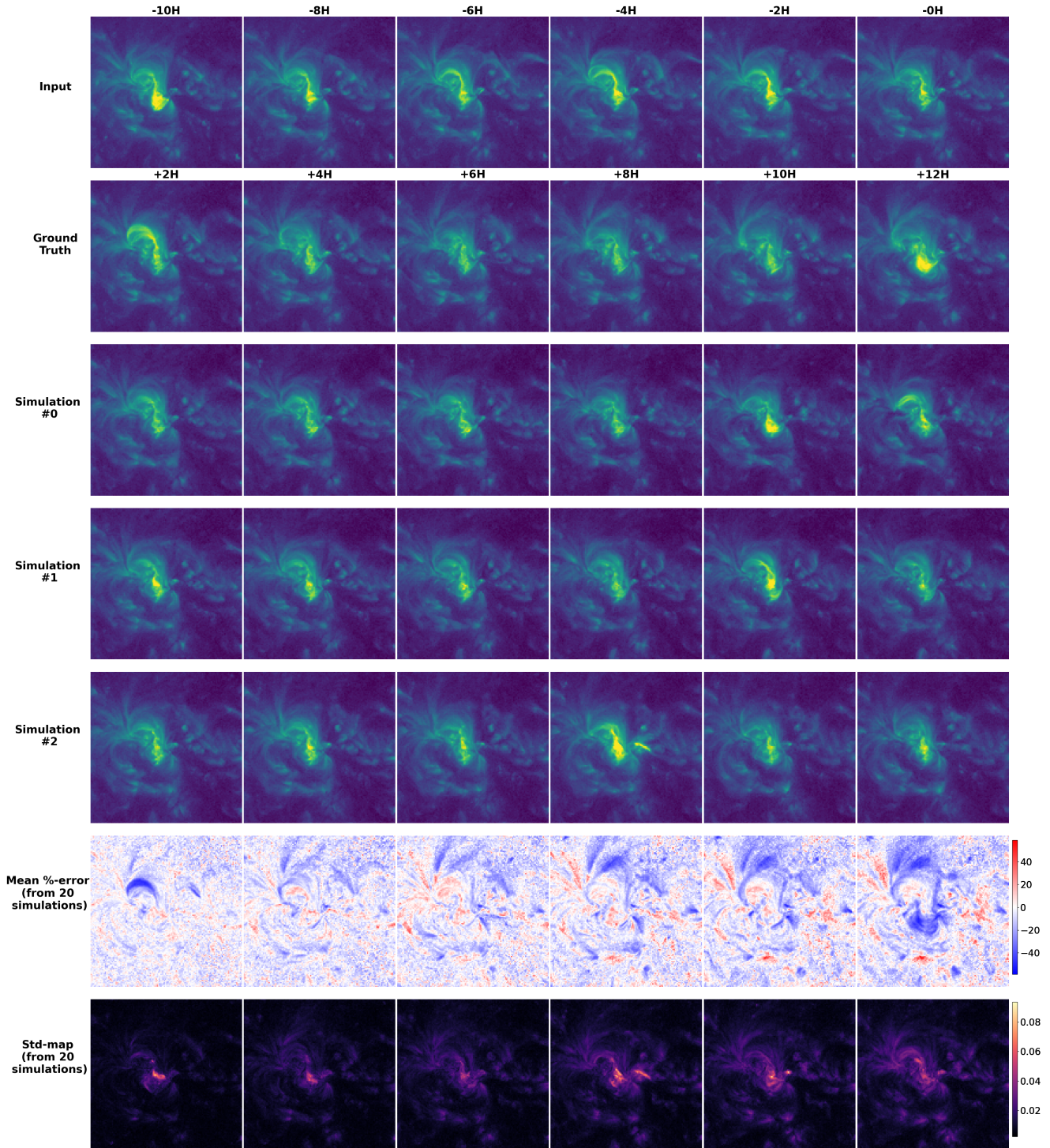


Figure 4: Predictions for AR 9188 on 2023-03-08 at 22:00. The AR exhibited a series of C-class and M-class flares during the [-10H, +12H] time window around the forecasting time. The -10H input frame shows strong remnants of a C flare that ended 30 minutes before the frame’s timestamp. The +2H ground truth frame displays the fading remnants of an M1.3 flare, which ended about 1 hour and 10 minutes earlier. The +12H frame exhibits strong remnants of a C3.1 flare ending 40 minutes before. Simulations 0 and 1 display a similar brightening in the +10H frame, while Simulation 2 shows one in the +6H frame. The mean percentage error indicates that overall activity is mostly underestimated, particularly around the upper loop where the M-class flare occurred, showing difficulty in predicting this flare and its aftermath. The standard deviation map shows increased uncertainty in the lower part of the sigmoid, where the C3.1 flare occurs between the +10H and +12H frames, suggesting that the model forecasted the possibility of similar events between the +6H and +12H frames, where the standard deviations are higher. Animations links : [Simulation #0](#), [Simulation #1](#), [Simulation #2](#).

6 List of Acronyms

AIA	Atmospheric Imaging Assembly
AR	Active Region
CI	Confidence Interval
DDPM	Denoising Diffusion Probabilistic Model
DTW	Dynamic Time Warping
EUV	Extreme Ultraviolet
GAN	Generative Adversarial Network
LPIPS	Learned Perceptual Image Patch Similarity
MAPE	Mean Absolute Percentage Error
MAE	Mean Absolute Error
MPF	Maximum Peak Flux
PSNR	Peak Signal To Noise Ratio
SDO	Solar Dynamic Observatory
SSIM	Structural Similarity Index Measure
SXR	Soft X-Rays
T2PF	Time To Peak Flux
VAE	Variational Auto Encoder

References

- ASPERTI, A. et TONELLI, V. (2023). Comparing the latent space of generative models. *Neural Computing and Applications*, 35(4):3155–3172.
- BARNES, W. T., CHEUNG, M. C. M., BOBRA, M. G., BOERNER, P. F., CHINTZOGLOU, G., LEONARD, D., MUMFORD, S. J., PADMANABHAN, N., SHIH, A. Y., SHIRMAN, N., STANSBY, D. et WRIGHT, P. J. (2020). aiapy: A python package for analyzing solar euv image data from aia. *Journal of Open Source Software*, 5(55):2801.
- BOBRA, M. G., SUN, X., HOEKSEMA, J. T., TURMON, M., LIU, Y., HAYASHI, K., BARNES, G. et LEKA, K. D. (2014). The helioseismic and magnetic imager (hmi) vector magnetic field pipeline: Sharps – space-weather hmi active region patches. *Solar Physics*, 289(9):3549–3578.
- BOERNER, P., EDWARDS, C., LEMEN, J., RAUSCH, A., SCHRIJVER, C., SHINE, R., SHING, L., STERN, R., TARBELL, T., TITLE, A., WOLFSON, C. J., SOUFLI, R., SPILLER, E., GULLIKSON, E., MCKENZIE, D., WINDT, D., GOLUB, L., PODGORSKI, W., TESTA, P. et WEBER, M. (2012). Initial Calibration of the Atmospheric Imaging Assembly (AIA) on the Solar Dynamics Observatory (SDO). *Solar Phys.*, 275(1-2):41–66.
- BOUCHERON, L. E., AL-GHRAIBAH, A. et MCATEER, R. T. J. (2015). Prediction of Solar Flare Size and Time-to-Flare Using Support Vector Machine Regression. *Astrophys. J.*, 812(1):51.
- CHERVYAKOV, N., LYAKHOV, P. et NAGORNOV, N. (2020). Analysis of the quantization noise in discrete wavelet transform filters for 3d medical imaging. *Applied Sciences*, 10(4).
- CHEUNG, M. C. M. et ISOBE, H. (2014). Flux Emergence (Theory). *Living Reviews in Solar Physics*, 11(1):3.
- CRANMER, S. R. (2002). Coronal Holes and the High-Speed Solar Wind. *Space Sci. Rev.*, 101(3):229–294.
- FRANCISCO, G., BERRETTI, M., CHIERICHINI, S., MUGATWALA, R., FERNANDES, J., BARATA, T. et MORO, D. (2024). Limits of solar flare forecasting models and new deep learning approach. *ESS Open Archive*.
- GOODFELLOW, I. J., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAI, S., COURVILLE, A. et BENGIO, Y. (2014). Generative Adversarial Networks. *arXiv e-prints*, page arXiv:1406.2661.
- HO, J., JAIN, A. et ABBEEL, P. (2020). Denoising Diffusion Probabilistic Models. *arXiv e-prints*, page arXiv:2006.11239.
- KINGMA, D. P. et WELLING, M. (2013). Auto-Encoding Variational Bayes. *arXiv e-prints*, page arXiv:1312.6114.
- PIERSOL, A. (1981). Time delay estimation using phase data. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(3):471–477.
- PRIEST, E. et FORBES, T. (2002). *Magnetic Reconnection: MHD Theory and Applications*. Cambridge University Press.
- RAMUNNO, F. P., HACKSTEIN, S., KINAKH, V., DROZDOVA, M., QUÉTANT, G., CSILLAGHY, A. et VOLOSHYNOVSKIY, S. (2024a). Solar synthetic imaging: Introducing denoising diffusion probabilistic models on SDO/AIA data. *Astron. Astrophys.*, 686:A285.
- RAMUNNO, F. P., JEONG, H.-J., HACKSTEIN, S., CSILLAGHY, A., VOLOSHYNOVSKIY, S. et GEORGIOULIS, M. K. (2024b). Magnetogram-to-Magnetogram: Generative Forecasting of Solar Evolution. *arXiv e-prints*, page arXiv:2407.11659.
- SAHARIA, C., CHAN, W., CHANG, H., LEE, C. A., HO, J., SALIMANS, T., FLEET, D. J. et NOROUZI, M. (2021). Palette: Image-to-Image Diffusion Models. *arXiv e-prints*, page arXiv:2111.05826.
- SHIBATA, K. et MAGARA, T. (2011). Solar Flares: Magneto-hydrodynamic Processes. *Living Reviews in Solar Physics*, 8(1):6.
- SOHL-DICKSTEIN, J., WEISS, E. A., MAHESWARANATHAN, N. et GANGULI, S. (2015). Deep Unsupervised Learning using Nonequilibrium Thermodynamics. *arXiv e-prints*, page arXiv:1503.03585.
- VAN DER SANDE, K., FLYER, N., BERGER, T. E. et GAGNON, R. (2022). Solar flare catalog based on SDO/AIA EUV images: Composition and correlation with GOES/XRS X-ray flare magnitudes. *Frontiers in Astronomy and Space Sciences*, 9:354.
- VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, L. et POLOSUKHIN, I. (2017). Attention Is All You Need. *arXiv e-prints*, page arXiv:1706.03762.
- WANG, Y. M. (2024). Coronal Holes, Footpoint Reconnection, and the Origin of the Slow (and Fast) Solar Wind. *Solar Phys.*, 299(4):54.
- WIEGELMANN, T. et SAKURAI, T. (2012). Solar force-free magnetic fields. *Living Reviews in Solar Physics*, 9:5.
- ZHANG, R., ISOLA, P., EFROS, A. A., SHECHTMAN, E. et WANG, O. (2018). The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. *arXiv e-prints*, page arXiv:1801.03924.