Precipitation Downscaling
with Spatiotemporal Video Diffusion

Prakhar Srivastava[1] **Ruihan Yang**[1] **Gavin Kerrigan**[1] **Gideon Dresdner**[2]
Jeremy McGibbon[2] **Christopher Bretherton**[2] **Stephan Mandt**[1]
[1]University of California, Irvine   [2]Allen Institute for AI, Seattle
{prakhs2,ruihan.yang,gavin.k,mandt}@uci.edu
{gideond,jeremym,christopherb}@allenai.org

Abstract

In climate science and meteorology, high-resolution local precipitation (rain and snowfall) predictions are limited by the computational costs of simulation-based methods. Statistical downscaling, or super-resolution, is a common workaround where a low-resolution prediction is improved using statistical approaches. Unlike traditional computer vision tasks, weather and climate applications require capturing the accurate conditional distribution of high-resolution given low-resolution patterns to assure reliable ensemble averages and unbiased estimates of extreme events, such as heavy rain. This work extends recent video diffusion models to precipitation super-resolution, employing a deterministic downscaler followed by a temporally-conditioned diffusion model to capture noise characteristics and high-frequency patterns. We test our approach on FV3GFS output, an established large-scale global atmosphere model, and compare it against six state-of-the-art baselines. Our analysis, capturing CRPS, MSE, precipitation distributions, and qualitative aspects using California and the Himalayas as examples, establishes our method as a new standard for data-driven precipitation downscaling.
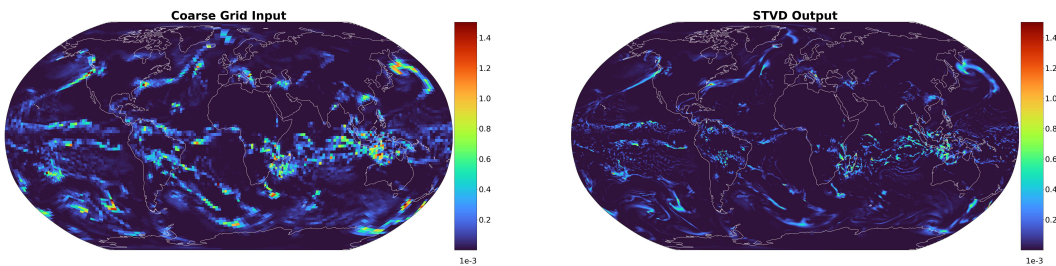
1   Introduction

Figure 1: Static snapshot from the Spatiotemporal Video Diffusion (STVD) model, illustrating input (left) and output (right) precipitation frames. The input panel displays simulated coarse-resolution precipitation (rain, snow) fields (Section 3), super-resolved into the high-resolution output shown in the right panel. Both frames use Robinson projection and cover six tiles of the cubed-sphere grid, providing a detailed global view (optimal viewing with zoom). For dynamics, see Fig. 3.

Precipitation patterns are central to human and natural life. In a rapidly warming climate, reliable simulations of changing precipitation patterns can help adapt to climate change. However, these simulations are challenging due to the multi-scale variability of weather systems and the influence of complex surface features (like mountains and coastlines) on precipitation trends and extremes
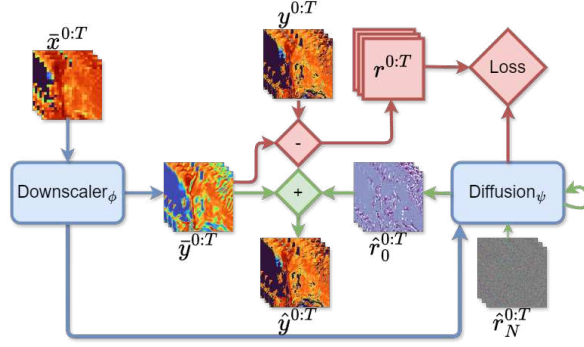
Figure 2: Our model's training and inference pipelines: Blue blocks apply to both phases, red blocks to training only, and green blocks to inference only. It deterministically downscales a low-resolution precipitation sequence using spatio-temporal factorized attention and models residuals with conditional diffusion (with factorized attention). Here, T denotes sequence length and N denotes diffusion steps. The parameters ($\theta = \phi, \psi$) are optimized jointly during training. See A.1 for details.

[58]. For many purposes, such as estimating flood hazards, precipitation must be estimated at spatial resolutions of only a few kilometers. Fluid-dynamical models of the global atmosphere are too expensive to run routinely at such fine scales [65], so the climate adaptation community relies on "downscaling"[1] of coarse-grid simulations to a finer grid. Traditional downscaling methods are either dynamical (running a fine-grid fluid-dynamical model limited to the region of interest, which requires specialized knowledge and computational resources) or statistical (typically restricted to simple univariate methods) [67]. Our work builds on vision based super-resolution methods to improve statistical downscaling and is a natural follow-up to recent deep-learning-based weather/climate prediction methods, which have revolutionized data-driven forecasting. These approaches boast improvements of orders of magnitude in runtime without sacrificing accuracy [46, 29].

We address the downscaling problem for a sequence. Our objective is to transform a sequence ("video") of low-resolution precipitation frames into a sequence of high-resolution frames. Despite differences from natural videos, precipitation's hourly temporal continuity allows us to use video super-resolution techniques to leverage multiple context frames for stochastic downscaling [53, 38].

Recent efforts to enhance the resolution of climate states like precipitation have relied on deterministic regression methods using convolutions or transformers. However, super-resolution is a one-to-many mapping with a continuum of "correct" answers. Supervised learning for these problems often leads to visual artifacts from *mode averaging*, where the network predicts an average of incompatible solutions, causing blurriness in visual data [30, 76]. Besides visual artifacts, mode averaging can have even more dramatic implications in climate and weather modeling, such as the underestimation of extreme precipitation [44], which is mainly induced by regional weather patterns on the unresolved scale. A natural alternative to supervised super-resolution methods [12, 27, 75, 11, 23] to prevent mode averaging is conditional *generative modeling*, which captures multimodal conditional distributions.

To that end, recent works propose using generative adversarial networks (GANs) for precipitation downscaling. These methods often face challenges, tending to converge on specific modes of the data distribution and occasionally fixating on isolated points in extreme cases. Despite their perceptual appeal, the scientific utility of super-resolution requires accurate modeling of the statistical *distribution* of high-resolution data given low-resolution input, which GANs typically fail to capture.

We propose SpatioTemporal Video Diffusion (STVD)[2] for precipitation downscaling. We use a deterministic regression model ("downscaler") for a coarse prediction, refined by a conditional video diffusion model that captures the residual error for adding fine-grained details. Both modules rely on spatio-temporal factorized attention to process the input sequence. Diffusion models are well-suited for precipitation downscaling as they successfully capture high dimensional and multimodal distributions, alleviating a key drawback of GAN-based methods for climate science applications.

---

[1]This is the climate science terminology for super-resolution.
[2]Code : https://github.com/mandt-lab/STVD

This study highlights the capability of conditional diffusion models to meet the specific needs of statistical precipitation downscaling, with our key contributions being:

1. We introduce a novel framework for temporal precipitation downscaling using diffusion models. Our model combines a deterministic downscaling module with a diffusion-based residual module. It leverages spatio-temporal factorized attention to process information from multiple low-resolution frames.

2. Our model outperforms six strong super-resolution baselines across multiple criteria, including MSE and several distributional metrics. We compare against two image super-resolution models and four video super-resolution models using the FV3GFS global atmosphere simulation dataset [77, 10].

3. Our approach captures key characteristics of precipitation, including extreme precipitation probabilities and spatial patterns of annual precipitation in mountainous regions, which are crucial for domain science applications.

Our paper is structured as follows: we first describe our method (Sec. 2), followed by our experimental findings (Sec. 3). Finally, we discuss relevant literature (Sec. 4) and its connection to our work. Fig. 1 shows a global view of the input and predicted precipitation of our model. The code for our model is available at `https://github.com/mandt-lab/STVD`.

2 Downscaling via Spatiotemporal Video Diffusion

Problem Statement At training time, we assume access to a collection of high-resolution precipitation frame sequences $\mathbf{y}^{0:T}$ and their corresponding low-resolution precipitation frame sequences $\mathbf{x}^{0:T}$. Such a low-resolution sequence can be obtained through area-weighted coarsening [39] of the corresponding high-resolution sequence. The dataset is discussed extensively in Sec. 3. Frame indices are represented by superscripts, where we assume that each sequence consists of $T+1$ frames for simplicity. While it may be possible to roll out predictions for multiple sequences autoregressively using techniques such as reconstruction guidance [21], we leave this exploration for future work. Our objective is to train a model to effectively *downscale*, or *super-resolve* a given sequence $\mathbf{x}^{0:T}$ with $\mathbf{y}^{0:T}$ serving as the target. We use "downscaling" and "super-resolution" interchangeably.

More formally, let $\mathbf{x}^t \in \mathbb{R}^{C \times H \times W}$ and $\mathbf{y}^t \in \mathbb{R}^{1 \times sH \times sW}$ represent individual low-resolution and high-resolution frames. Here, $s \in \mathbb{N}$ denotes the downscaling factor, C is the number of channels (quantities used as input to the model to characterize the atmospheric state in each low-resolution grid cell so as to add skill to the precipitation prediction), and H, W indicate the height and width of the low-resolution frame. For our study, we adopt a downscaling factor of $s = 8$ and have $C = 12$ total low-resolution channels. In addition to the low-resolution precipitation state, we provide eleven channels of information to the model, such as topography, wind velocity, and surface temperature; see A.2 for details.

Solution Sketch Our approach treats the downscaling problem as a conditional generative modeling task. We devise a model to learn the conditional distribution of high-resolution precipitation frames, incorporating contextual information from the low-resolution precipitation frame sequence.

Our proposed solution, **SpatioTemporal Video Diffusion (STVD)** (Fig. 2), relies on two modules: a deterministic downscaler and a stochastic component based on conditional diffusion models [20, 63], both using spatio-temporal factorized attention. The first module uses a UNet with factorized attention to integrate information from a low-resolution frame sequence, resulting in an initial prediction frame sequence $\bar{\mathbf{y}}^{0:T}$. The second module is a conditional diffusion model that stochastically generates a sequence of additive residual frames $\mathbf{r}^{0:T}$ which serves to add fine-grained details to the initial prediction. Together, these two modules produce a high-resolution frame sequence $\hat{\mathbf{y}}^{0:T} = \bar{\mathbf{y}}^{0:T} + \mathbf{r}^{0:T}$. Both modules are trained end-to-end.

Decomposing the prediction into a deterministic mean and a stochastic residual is inspired by predictive-coding-based video decompression. This approach aims to predict a sequence of video frames while compressing the sparse residuals [2, 74] which are easier to model than dense frames. Similarly, it is easier to generate residuals than dense images when using diffusion models [73, 41].
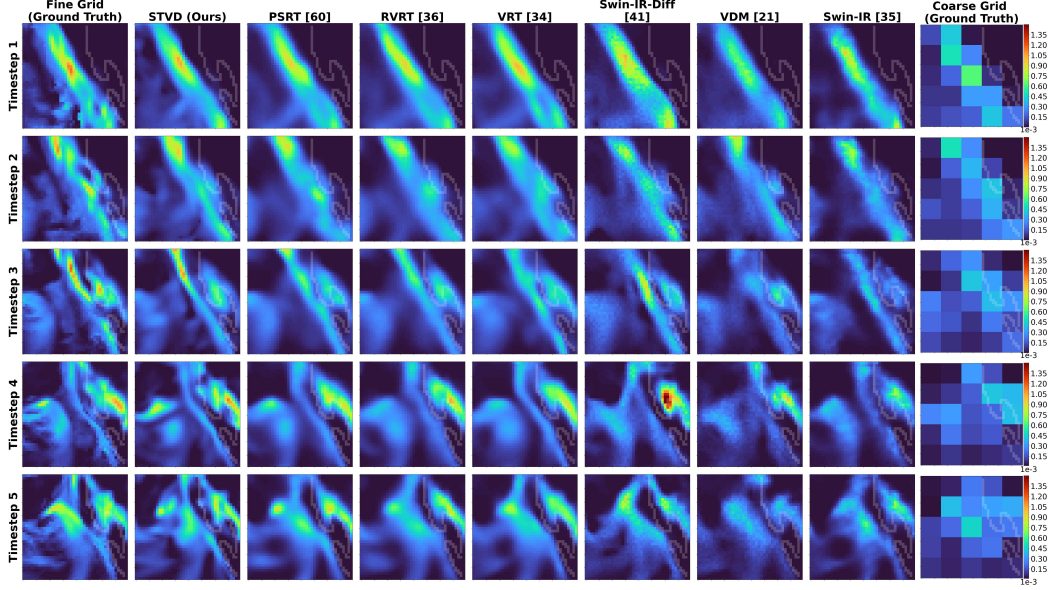
Figure 3: A qualitative comparison between our proposed model and top baseline for a precipitation event associated with a cold front impinging on the Northern California coast and then the Sierra mountain range (coastline marked in hazy white). Fig. 6 plots the regional topography. The time interval between adjacent frames is 3 hours; the plotted region is 1000×1000 km. Our model resolves the fine-grid precipitation structure better than the considered baselines. See A.3 for full-page high quality samples from Himalayas and Sierra.

In what follows, we first describe our overall probabilistic framework for downscaling. Then, we discuss the deterministic module along with spatio-temporal attention, followed by the remaining residual prediction module based on diffusion generative modeling. See A.1 for architecture details.

2.1 Probabilistic Modeling of Downscaling

Given a sequence of low-resolution frames $\mathbf{x}^{0:T}$ and the corresponding high-resolution frames $\mathbf{y}^{0:T}$, we aim to learn a parametric approximation p_θ of the conditional distribution $p\left(\mathbf{y}^{0:T} \mid \mathbf{x}^{0:T}\right) \approx p_\theta\left(\mathbf{y}^{0:T} \mid \mathbf{x}^{0:T}\right)$. Importantly, we do not assume independence across time; each generated frame $\mathbf{y}^t$ can depend on all other generated frames. The generated high-resolution frame sequence is conditioned on the entire low-resolution frame sequence, capturing long-range temporal correlations and enhancing the fidelity and cohesion of the high-resolution reconstruction.

As noted earlier, the likelihood $p_\theta\left(\mathbf{y}^{0:T} \mid \mathbf{x}^{0:T}\right)$ is modeled using a deterministic downscaler and a residual diffusion model. We will discuss how the model parameters $\theta = (\phi, \psi)$ decompose into those for a downscaler (ϕ) and a diffusion model (ψ).

2.1.1 Deterministic Downscaling

Our first module is a deterministic downscaler that predicts an initial high-resolution frame sequence $\bar{\mathbf{y}}^{0:T} = \mu_\phi\left(\mathbf{x}^{0:T}\right)$ where μ_ϕ is a network generating a deterministic high-resolution prediction with parameters ϕ. We perform bicubic interpolation on each frame of $\mathbf{x}^{0:T}$ before passing the sequence through the network μ_ϕ. Since the diffusion network operates on high-resolution inputs (i.e. denoising the high-resolution residuals), this choice allows us to use the same UNet [52] architecture (with different weights) for both the downscaling module μ_ϕ and the residual diffusion module. This enables us to easily share features across the modules via concatenation. See A.1 for further details.

Importantly, μ_ϕ incorporates a temporal attention mechanism that allows any frame at time t, or its corresponding feature map, to attend to all context frames from 0 to T. This architecture enables the concurrent inference of all frames within the sequence $\bar{\mathbf{y}}^{0:T}$. The attention weights differ for each frame, allowing for the flexible incorporation of information across time.

4

2.1.2 Stochastic Residual Modeling via Diffusion

After computing the initial prediction $\bar{\mathbf{y}}^{0:T}$, finer details are modeled by residuals learned from a conditional diffusion model. Our final stochastic high-resolution frame sequence $\hat{\mathbf{y}}^{0:T}$ is generated by sampling an additive residual sequence $\mathbf{r}^{0:T}$ from this model: $\hat{\mathbf{y}}^{0:T} = \bar{\mathbf{y}}^{0:T} + \mathbf{r}^{0:T}$. Thus, we seek to model the residuals $\mathbf{r}^{0:T} = \mathbf{y}^{0:T} - \bar{\mathbf{y}}^{0:T}$. Our diffusion model generates the entire residual sequence $\mathbf{r}^{0:T}$ concurrently, with the generation of each residual $\mathbf{r}^t$ dependent on the others. This is achieved via a UNet architecture with spatio-temporal attention, similar to the mechanism used for the deterministic downscaling module. See A.1 for further details.

To model the distribution of $\mathbf{r}^{0:T}$, we use DDPM [20]. To that end, we introduce a collection of latent variables $\mathbf{r}_{0:N}^{0:T}$, where the lower subscripts indicate the denoising diffusion step. In the *forward process*, the latent variable $\mathbf{r}_n^{0:T}$ is created from $\mathbf{r}_{n-1}^{0:T}$ via additive noise. In the *reverse process* for generation, a denoising model (with parameters ψ) is trained to predict $\mathbf{r}_{j-1}^{0:T}$ from $\mathbf{r}_j^{0:T}$. N denotes the total number of denoising steps. Note that $\mathbf{r}^{0:T} = \mathbf{r}_0^{0:T}$, i.e. the first diffusion step corresponds to the true residual. Additionally, $\mathbf{r}_0^{0:T}$ implicitly depends on the downscaler parameters ϕ, allowing us to simultaneously optimize all model parameters $\theta = (\phi, \psi)$ within the context of diffusion modeling.

As is standard in diffusion models [20], we parameterize the reverse process via a Gaussian distribution with a mean determined by a neural network M_ψ,

$$p_\psi\left(\mathbf{r}_{n-1}^{0:T}|\mathbf{r}_n^{0:T}, \mathbf{c}\right) = \mathcal{N}\left(\mathbf{r}_{n-1}^{0:T}|M_\psi\left(\mathbf{r}_n^{0:T}, n, \mathbf{c}\right), \gamma\mathbf{I}\right), \tag{1}$$

where M_ψ is a denoising network and γ is a hyperparameter for variance. The diffusion model directly accesses the context $\mathbf{c} = (\mathbf{x}^{0:T}, \bar{\mathbf{y}}^{0:T})$, and is implicitly conditioned on $\mathbf{x}^{0:T}$ via concatenation of feature maps from the downscaler module. As in the downscaler, we bicubically upsample $\mathbf{x}^{0:T}$ before channel-wise concatenation with $\bar{\mathbf{y}}^{0:T}$ to match the dimensions when forming $\mathbf{c}$.

2.1.3 Loss Function

To train our model, we use the angular parametrization suggested by [59]. Specifically, this results in the diffusion loss of the form

$$L\left(\psi, \phi\right) = \mathbb{E}_{\mathbf{x}^{0:T}, \mathbf{y}^{0:T}, n, \epsilon} \sum_{t=0}^{T} \left\|\mathbf{v} - M_\psi\left(\mathbf{r}_n^{0:T}, n, \mathbf{c}\right)\right\|^2 \tag{2}$$

where $\epsilon \sim \mathcal{N}(0, I)$, n is sampled uniformly from $\{1, 2, \ldots, N\}$, and the sequences $\mathbf{x}^{0:T}$, $\mathbf{y}^{0:T}$ are sampled from the training distribution. Here, $\mathbf{c} = (\mathbf{x}^{0:T}, \bar{\mathbf{y}}^{0:T})$ where $\bar{\mathbf{y}}^{0:T} = \mu_\phi\left(\mathbf{x}^{0:T}\right)$. The scalars $\alpha_n^2 = \prod_{i=1}^{n}(1 - \beta_i)$ and $\sigma_n^2 = 1 - \alpha_n^2$ are used to define $\mathbf{v} \equiv \alpha_n\epsilon - \sigma_n\mathbf{r}_0^{0:T}$. Training and inference are concurrent across multiple frames due to spatio-temporal attention. Alg. 1 and 2 demonstrate the training and sampling strategy under the angular parametrization. We use DDIM sampling [62] to generate frame residuals with fewer diffusion steps.

2.1.4 Network Architecture

Both the downscaler and the conditional diffusion model employ a UNet backbone with similar architectures and key adaptations to the attention mechanism (see A.1). The downscaler takes the multi-channel input frames ($\mathbf{x}^{0:T}$), yielding an initial estimate ($\bar{\mathbf{y}}^{0:T}$). The diffusion UNet conditions on diffusion step n and concatenates feature maps from the downscaler with its own. The concatenated input to the diffusion UNet ($\mathbf{x}^{0:T}$, $\bar{\mathbf{y}}^{0:T}$, and $\mathbf{r}_n^{0:T}$), along with the conditioning variables (diffusion step n and the feature maps from downscaler), yields the output $\mathbf{v}$.

Computing full attention for temporal coherence across the entire video data cube is very expensive for processing long sequences or high-resolution inputs. To optimize efficiency, we decouple attention between spatial and temporal dimensions, use a linear variant of self-attention [26] for non-bottleneck layers (where the effective number of "tokens" for attention is relatively large), focus spatial attention on localized patches (instead of the entire feature map, which could be wasteful), and calculate per-channel temporal attention in large spatial dimensions (namely, the ultimate and penultimate expansion and contraction layers of UNet). These modifications dramatically reduce the time complexity and memory footprint of these transformer blocks.
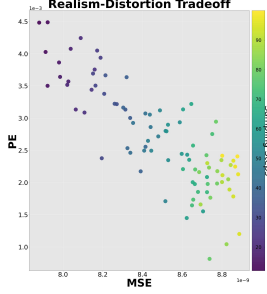
Figure 4: Tradeoff between mean square error and percentile error (see Sec. 3). Inference at Himalayan region (see Figs. 6 and 12).

Algorithm 1: Training STVD

while *not converged* **do**
  Sample $\mathbf{x}^{0:T}$ and $\mathbf{y}^{0:T}$;
  $n \sim \mathcal{U}(0, 1, 2, .., N)$;
  $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$;
  $\bar{\mathbf{y}}^{0:T} = \mu_\phi\left(\mathbf{x}^{0:T}\right)$;
  $\mathbf{r}_0^{0:T} = \mathbf{y}^{0:T} - \bar{\mathbf{y}}^{0:T}$;
  $\mathbf{v} = \alpha_n\epsilon - \sigma_n\mathbf{r}_0^{0:T}$;
  $\mathbf{r}_n^{0:T} = \alpha_n\mathbf{r}_0^{0:T} + \sigma_n\epsilon$;
  $\mathbf{c} = \left(\mathbf{x}^{0:T}, \bar{y}^{0:T}\right)$;
  $\hat{\mathbf{v}} = M_\psi\left(\mathbf{r}_n^{0:T}, n, \mathbf{c}\right)$;
  $L = ||\mathbf{v} - \hat{\mathbf{v}}||^2$;
  $(\psi, \phi) = (\psi, \phi) - \nabla_{\psi,\phi}L$;

Algorithm 2: Sampling STVD

Get an equally spaced increasing sub-sequence τ of length $K \ll N$;
$\bar{\mathbf{y}}^{0:T} = \mu_\phi\left(\mathbf{x}^{0:T}\right)$;
$\mathbf{c} = \left(\mathbf{x}^{0:T}, \bar{y}^{0:T}\right)$;
$\mathbf{r}_K^{0:T} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$;
for *n in reversed(τ)* **do**
  $\hat{\mathbf{v}} = M_\psi\left(\mathbf{r}_n^{0:T}, n, \mathbf{c}\right)$;
  $\hat{\mathbf{r}} = \alpha_n\mathbf{r}_n^{0:T} - \sigma_n\hat{\mathbf{v}}$;
  $\hat{\epsilon} = \frac{\sigma_n}{\alpha_n}\left(\mathbf{r}_n^{0:T} - \hat{\mathbf{r}}\right)$;
  $\mathbf{r}_{n-1}^{0:T} = \alpha_{n-1}\hat{\mathbf{r}} + \sigma_{n-1}\hat{\epsilon}$;
$\hat{\mathbf{y}}^{0:T} = \bar{\mathbf{y}}^{0:T} + \mathbf{r}_0^{0:T}$;

3 Experiments

We conduct a comprehensive evaluation of our proposed method, SpatioTemporal Video Diffusion (STVD), against six contemporary state-of-the-art baselines. The first two baselines are image super-resolution models based on the Swin Vision Transformer (Swin-IR)[35] and its residual diffusion variant (Swin-IR-Diff). The next two baselines are video super-resolution models grounded in vision transformer architecture (VRT) [34] and its recurrent variant (RVRT) [36]. The latter incorporates guided deformable attention for clip alignment, enhancing its temporal modeling capabilities. We compare against another video-super-resolution baseline (PSRT) [60] which also relies on the transformer architecture but uses multi-frame attention groups. Finally, we compare against a video diffusion baseline (VDM) [21]. Fig. 1 shows a global view of the input and predicted precipitation.

We perform ablation studies in three configurations. In the first two, we experiment with the input sequence length. While our proposed model uses a context length of 5 frames, we also conduct experiments with 3 frames and 1 frame (STVD-3 and STVD-1). Note that using a single context frame ablates for the temporal attention block as well. The third ablation (STVD-Single) involves removing the additional input channels (i.e. only providing the model with the low-resolution precipitation sequence) to assess their impact on performance metrics. In summary, our experiments demonstrate that our method outperforms all baselines across all metrics considered. Additionally, our ablation studies highlight the importance of temporal context and additional climate inputs.

Dataset Our dataset derives from an 11-member initial condition ensemble of 13-month simulations using a global atmosphere model, FV3GFS, run at 25 km resolution and forced by climatological sea surface temperatures and sea ice. The first month of each simulation is discarded to allow the simulations to spin up and meteorologically diverge, effectively providing 11 years of reference data (of which first 10 years are used for training and the last year for validation). FV3GFS, developed by the National Oceanic and Atmospheric Administration (NOAA), is a version of NOAA's operational global weather forecast model ([77, 10]).

Three-hourly average data were saved from this entire simulation, which used a 25 km horizontal "fine grid". We further coarsened the selected fields by a factor of 8 to create a 200 km horizontal "coarse grid", resulting in paired data (x_t, y_t), where x_t is the coarse-grid global state and y_t is the corresponding fine-grid global state. Our goal is to apply video downscaling to the coarse-grid precipitation field to obtain temporally smooth fine-grid precipitation estimates that are statistically similar to the true data. This approach is attractive because many fine-grid precipitation features, such as cold fronts and tropical cyclones, are poorly resolved on the coarse grid but are temporally coherent across periods much longer than 3 hours. We use 12 coarse-grid input fields, including precipitation, topography, and horizontal vector wind at various levels. See A.2 for the list of included atmospheric variables. FV3GFS uses a cubed-sphere grid, where the surface of the globe is divided into six tiles, each of which is covered by an $S \times S$ array of points. Our data fields reflect this structure with $S = 48$ for the 200 km coarse grid and $S = 384$ for the 25 km fine grid.
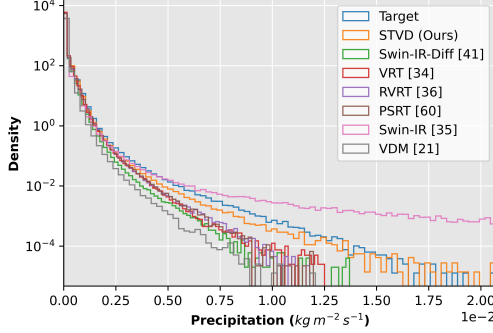
Figure 5: Distributions of the fine-grid three-hourly average precipitation, for all gridpoints around the globe. The Swin-IR baseline overestimates large precipitation events, whereas all other baselines underestimate key extreme and rare precipitation events. Our model aligns best with the fine-grid ground truth than any the other model. This is also evident with the the EMD and PE metrics discussed in Tab. 1 and Sec. 3.

Table 1: Quantitative comparison between our method and other competitive baselines. EMD represents the Earth-Mover Distance, PE denotes the 99.999th percentile error and SAE is the spatial-autocorrelation error. Overall, our proposed method (STVD) outperforms the baselines across all metrics. In our ablation study, the exclusion of additional side information (STVD-single) or decrement in context length (STVD-3 and STVD-1) appreciably degrades performance.

	CRPS (10^{-5})	MSE (10^{-8})	EMD (10^{-6})	PE (10^{-3})	SAE (10^{-6})
STVD (ours)	**1.85**	**0.59**	**2.49**	**1.2**	**4.00**
PSRT [60]	2.15	0.66	4.21	3.8	6.24
RVRT [36]	3.55	1.73	4.33	3.6	7.39
VRT [34]	3.58	1.74	4.61	4.0	7.39
Swin-IR-Diff [41]	2.29	1.94	6.38	4.4	7.70
VDM [21]	2.21	0.73	12.70	6.4	8.84
Swin-IR [35]	2.36	2.29	17.40	23.40	18.9
STVD-single	1.81	0.62	4.64	2.3	6.09
STVD-3	1.96	0.68	4.94	2.6	4.99
STVD-1	2.05	0.72	7.19	4.1	6.87

The application presented here serves as a pilot for broader uses of our methodology. Fine-grid simulations are significantly more computationally expensive than coarse-grid simulations (an 8-fold reduction in grid spacing requires almost 1000x more computation), so a coarse-grid simulation with super-resolved details in desired regions could be highly cost-effective for many applications.

During training, our model randomly selects data from one of the six tiles. This strategy ensures that the model learns from the diverse spatial contexts and weather regimes that produce precipitation worldwide. Post-training, for localized analysis, we selectively sample super-resolved precipitation channels from regions with complex terrain, such as California (Fig. 6). These regions can systematically pattern the precipitation on fine scales. This analysis helps us to see how well the super-resolution can learn the time-mean spatial patterns (e.g. precipitation enhancement on the windward side of mountain ranges and lee rain shadows) in the fine-grid reference data.

Training and Testing Details We downscale a sequence of precipitation frames from FV3GFS output by a factor of 8. Our approach (STVD) trains on 5 consecutive frames that are downscaled jointly. We optimize our model end-to-end with a single diffusion loss using Adam [28] with an initial learning rate of 1×10^{-4}, decaying to 5×10^{-7} with cosine annealing during training, executed on an NVidia RTX A6000 GPU. The diffusion model is trained using v-parametrization [59], with a fixed diffusion depth (N = 1400). Random tiles extracted from the cube-sphere representation of Earth, with dimensions 384 in high-resolution and 48 in low-resolution, are used during training. We train for one million steps, requiring approximately 7 days on a single node (slightly less for ablations). We use a batch size of one, apply a logarithmic transformation to precipitation states, and normalize to the range $[-1, 1]$. During testing, we employ DDIM sampling with 30 steps on an Exponential Moving Average (EMA) variant of our model (for full frame size), with a decay rate of 0.995.

Baseline Models We compare our generative setup against several recent high-performing transformer-based video super-resolution models. These models are trained deterministically. The first, Video Restoration Transformer (VRT) [34], allows for parallel frame prediction and long-range temporal dependency modeling. The second, recurrent VRT (RVRT) [36], incorporates guided deformable attention for effective clip alignment, enhancing its temporal modeling capabilities. The third, PSRT [60], removes the alignment module and modifies the attention window. We also compare against the recent Video Diffusion Model (VDM) [21] which employs global quadratic attention.

To assess the benefits of multi-frame downscaling, we compare with Swin-IR [35], a popular image super-resolution model that harnesses Swin Transformer blocks. However, Swin-IR is trained in a supervised fashion. Thus, as a generative baseline, we compare to Swin-IR-Diff. This model generates a deterministic prediction using Swin-IR [35], followed by modeling a stochastic residual
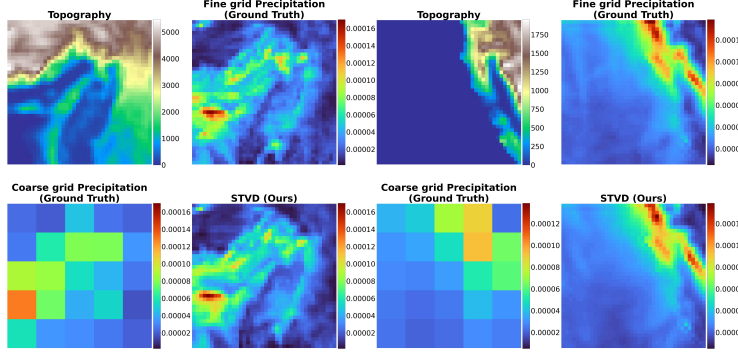
Figure 6: Precipitation over two regions (left: Himalayas; right: Northern California coast, same region as Fig. 3), averaged across a year, for our STVD model and the ground-truth. For each half, the topography of the region is shown in the corresponding top-left whereas the predicted annual average is shown in the corresponding bottom-right. Annually-averaged precipitation is an important indicator of water availability in a region. STVD successfully captures many details of the precipitation that are tied to local topography and are too fine to be resolved the coarse-grid data.

using diffusion. This baseline is inspired by concurrent work on single-image radar-reflectivity downscaling [41], where a UNet is used instead of Swin-IR. See A.2 for details.

Evaluation Metrics We evaluate our model differently from standard vision tasks. In addition to the Mean Square Error (MSE), which measures the average squared difference between predicted and actual values but lacks full distributional information, we use several distribution-level metrics for a more meaningful comparison. One such metric is the Continuous Ranked Probability Score (CRPS) [8, 66], which assesses the discrepancy between the predicted cumulative distribution function and the observed data. We compute CRPS over 10 stochastic realizations of our predictions. Fig. 7 visualizes several of these samples.

Furthermore, given the distinctive light-tailed exponential distribution of the precipitation climate state, it is crucial to ensure that downscaling does not significantly alter the distribution of precipitation rates. This necessitates two additional metrics. First, we compute the Earth Mover (or 1-Wasserstein) Distance [54] to quantify the agreement between the target and predicted global precipitation distributions, which are strongly affected by high-resolution details. Second, we focus on tail events and extreme precipitation by considering the 99.999th percentile error (PE), providing a nuanced understanding of the model's performance on rare and extreme precipitation events.

To further assess the spatial fidelity of our downscaling approach, we use the Spatial Autocorrelation Error (SAE) [68]. This metric calculates the mean absolute error between the spatial autocorrelation of the predictions and ground truth. Low SAE ensures that the spatial patterns and the fine structure in precipitation data are preserved during downscaling, which is critical for accurate climate modeling.

Qualitative and Quantitative Analysis Tab. 1 provides a quantitative evaluation comparing our method with state-of-the-art baselines and ablations. Our model (STVD) performs strongly across all metrics, outperforming all baselines. We highlight the distributional characteristics in Fig. 5. Swin-IR overestimates precipitation, while all other baselines underestimate it. This discrepancy is undesirable, as poor performance on rare and extreme precipitation events can negatively imapct disaster mitigation policies. In contrast, our method closely matches the precipitation distribution, as measured by PE and EMD.

Using only precipitation as an input (STVD-single) results in slightly worse performance across all metrics, indicating the predictive value of additional inputs. In contrast, our ablation model STVD-1, which lacks full sequence information, performs significantly worse, highlighting the importance of temporal attention in our approach (which decays as a function of time lag as shown in Fig. 14).

Figs. 3, 11 and 12 depict the performance of our model compared to other baselines on examples of a precipitation feature interacting with mountainous terrain. Our model generates high quality results
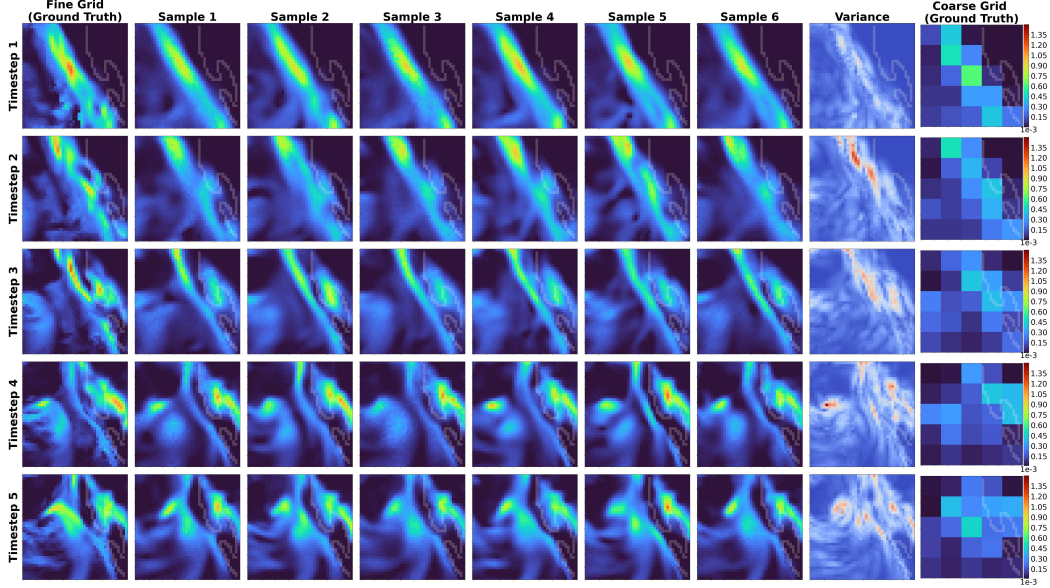
8

Figure 7: A visualization of the stochastic samples predicted by STVD for a given coarse-grid data. The precipitation event is the same as depicted in Fig. 3. Additionally, we also plot a variance map over the set of these samples to analyze the stochasticity better. Red regions correspond to high variance whereas blue regions correspond to low variance. Model stochasticity seems to be meaningful since the variance is large where mean precipitation is large.

which preserves most patterns with a high degree of similarity. PSRT and RVRT produce slightly more diffuse precipitation features, while Swin-IR produces slightly more pixelated features.

Fig. 6 shows annually-averaged precipitation from the patches in Figs. 3 and 12. Accurately capturing the fine-grid structure of time-mean precipitation is crucial for assessing long-term water availability. Our method (which includes fine-grid topography as a training input) effectively replicates the ground truth. This includes the strength and narrow spatial structure of high precipitation bands along the Northern California coastal mountains and the Sierras. These features are not resolved by the coarse-grid inputs to the super-resolution. See A.3 for full-page high-resolution samples and spectral analysis. Fig. 13 reveals that the spectra for baselines decay more rapidly than for STVD.

Realism-Distortion Tradeoff   Distortion metrics such as MSE often conflict with perceptual quality, where reducing distortion typically degrades perceptual realism [7]. In our context, this tradeoff translates to balancing MSE and PE. While MSE captures the average accuracy of predictions, PE represents the model's ability to reproduce extreme events, thereby serving as a proxy for realism. Realism in climate modeling refers to the accurate representation of extreme weather patterns, which are crucial for applications like flood forecasting and disaster mitigation. PE is a distributional criterion that effectively captures these tail events, offering a robust measure of realism. Fig. 4 illustrates this tradeoff, with darker colors corresponding to fewer STVD sampling steps. As the number of diffusion sampling steps increases, MSE tends to rise slightly, but PE decreases significantly. Depending on the application, this tradeoff may potentially be exploited by practitioners. Essentially, the conditional mean minimizes MSE, so any deviation from it increases MSE—even if the deviation appears more realistic. As for sampling steps, fewer steps correspond to larger time increments in the diffusion process. At one extreme, a single step predicts the conditional mean, minimizing MSE. Conversely, more sampling steps more accurately simulate diffusion, generating diverse, realistic samples that increase MSE while reducing PE.

4   Related Work

Diffusion Models   Diffusion models [61, 20, 64, 43, 45, 40] are a class of generative models based on an iterative denoising process. Closely related to our work are diffusion models for video. Recent

models [73] generate deterministic next-frame predictions autoregressively with additional residuals generated by a diffusion model, or generate videos directly in pixel space [18, 69, 21] or in a latent space [6, 5]. While some works on video diffusion [6, 21] employ video super-resolution as a step in the overall modeling process, our work focuses exclusively on the video super-resolution task, particularly within the context of precipitation downscaling.

Super-Resolution   Within the computer vision community, the paradigm for single image super-resolution has shifted from classical approaches [4, 13] to deep learning based methods [71]. Generative approaches, like cascaded diffusion [49, 55], SR3 [56], and DiffPIR [78] employ diffusion models for image super-resolution. However, these are unable to leverage temporal context. On the other hand, many approaches for video super-resolution have been proposed [9, 14, 22]. For a more comprehensive overview, see [38]. Recent models of note include the transformer-based models PSRT [60] and VRT [34], as well as the recurrent variant RVRT [37], which focuses on parallel decoding and guided clip alignment. We emphasize that these state-of-the-art approaches are deterministic, where our approach is generative. This allows us to preventing mode averaging and to produce more realistic samples, which is particularly critical in the context of precipitation modeling.

Data-driven weather and climate modeling   Recent years have seen advancements in data-driven climate and weather modeling [51, 42], with models like GraphCast [29], GenCast [48], and FourCastNet [46] providing forecasts that are competitive with meteorological methods while being significantly faster. Rather than replacing numerical forecasting methods, our approach seeks to augment their capabilities by downscaling coarse-grid predictions.

While downscaling for climate and weather has been approached using techniques based on domain knowledge [25, 24], we focus here on data-driven approaches. [68] draw inspiration from FRVSR [57], adopting an iterative approach that uses the high-resolution frame estimated in the previous step as input for subsequent iterations. [72] employed Fourier neural operators for downscaling at arbitrary resolutions. [16] generate physically consistent downscaled climate states, using a softmax layer to enforce conservation laws. These approaches, though, are deterministic and trained by minimizing the MSE, thus lacking the realism and uncertainty quantification provided by a generative approach.

In terms of generative approaches, concurrent work [41] employs diffusion models for downscaling climate states. The use of GANs has also been pervasive in downscaling and precipitation prediction [32, 47, 17, 50, 15, 70]. However, these GAN-based approaches inherit the mode collapse and training difficulties present in all GAN-based models [3]. Here, we highlight that these approaches are applied at each frame and cannot incorporate temporal information as is done in our model.

Beyond downscaling, [1] demonstrate the diffusion model's efficacy in synthesizing full rain density from vorticity inputs. Additionally, [19] uses a diffusion model for downscaling solar irradiance. These models are also used for probabilistic weather forecasting and nowcasting [31, 33].

5   Conclusion

We propose a video super-resolution method for probabilistic precipitation downscaling. Our model, STVD, deterministically super-resolves a given low-resolution frame sequence and then stochastically models the residual details via diffusion. Our model effectively resolves how fine-grid precipitation features, generated as weather systems, interact with complex topography based on temporally coherent coarse-grid information. Our method outperforms several competitive baselines on a range of quantitative metrics. This is an important step towards designing effective statistical downscaling methods, providing highly localized information for planning against extreme weather events, such as floods or hurricanes in a warming climate, using tractable coarse-grid atmospheric models.

Limitations and Broader Impacts   A limitation of our approach is the necessity of paired low-resolution and high-resolution images for training. While this can be done once prior to training, designing methods that only require the low-resolution states is an interesting challenge. In terms of broader impacts, our approach could potentially have harmful consequences if adopted blindly to a new dataset, where distribution shift could cause the model performance to degrade, potentially leading to underestimation of extreme weather risks such as droughts or floods. To mitigate these risks, the model should be re-trained and rigorously evaluated on the dataset of interest.

Acknowledgments and Disclosure of Funding

References

[1] Addison, H., Kendon, E., Ravuri, S., Aitchison, L., Watson, P.A.: Machine learning emulation of a local-scale uk climate model. arXiv preprint arXiv:2211.16116 (2022)

[2] Agustsson, E., Minnen, D., Johnston, N., Balle, J., Hwang, S.J., Toderici, G.: Scale-space flow for end-to-end optimized video compression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8503–8512 (2020)

[3] Arora, S., Risteski, A., Zhang, Y.: Do gans learn the distribution? some theory and empirics. In: International Conference on Learning Representations (2018)

[4] Bascle, B., Blake, A., Zisserman, A.: Motion deblurring and super-resolution from an image sequence. In: Computer Vision—ECCV'96: 4th European Conference on Computer Vision Cambridge, UK, April 15–18, 1996 Proceedings Volume II 4. pp. 571–582. Springer (1996)

[5] Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)

[6] Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22563–22575 (2023)

[7] Blau, Y., Michaeli, T.: The perception-distortion tradeoff. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6228–6237 (2018)

[8] Brown, T.A.: Admissible scoring systems for continuous distributions. (1974)

[9] Chan, K.C., Wang, X., Yu, K., Dong, C., Loy, C.C.: Basicvsr: The search for essential components in video super-resolution and beyond. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4947–4956 (2021)

[10] Community, U.: Ufs weather model (Jan 2021). doi:10.5281/zenodo.4460292, https://doi.org/10.5281/zenodo.4460292

[11] Dai, T., Cai, J., Zhang, Y., Xia, S.T., Zhang, L.: Second-order attention network for single image super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11065–11074 (2019)

[12] Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(2), 295–307 (2015)

[13] Farsiu, S., Robinson, M.D., Elad, M., Milanfar, P.: Fast and robust multiframe super resolution. IEEE transactions on image processing **13**(10), 1327–1344 (2004)

[14] Fuoli, D., Gu, S., Timofte, R.: Efficient video super-resolution through recurrent latent space propagation. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). pp. 3476–3485. IEEE (2019)

[15] Gong, A., Li, R., Pan, B., Chen, H., Ni, G., Chen, M.: Enhancing spatial variability representation of radar nowcasting with generative adversarial networks. Remote Sensing **15**(13), 3306 (2023)

[16] Harder, P., Yang, Q., Ramesh, V., Sattigeri, P., Hernandez-Garcia, A., Watson, C., Szwarcman, D., Rolnick, D.: Generating physically-consistent high-resolution climate data with hard-constrained neural networks. arXiv preprint arXiv:2208.05424 (2022)

[17] Harris, L., McRae, A.T., Chantry, M., Dueben, P.D., Palmer, T.N.: A generative deep learning approach to stochastic downscaling of precipitation forecasts. Journal of Advances in Modeling Earth Systems **14**(10), e2022MS003120 (2022)

[18] Harvey, W., Naderiparizi, S., Masrani, V., Weilbach, C.D., Wood, F.: Flexible diffusion modeling of long videos. In: Advances in Neural Information Processing Systems (2022)

[19] Hatanaka, Y., Glaser, Y., Galgon, G., Torri, G., Sadowski, P.: Diffusion models for high-resolution solar forecasts. arXiv preprint arXiv:2302.00170 (2023)

[20] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)

[21] Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. arXiv preprint arXiv:2204.03458 (2022)

[22] Huang, Y., Wang, W., Wang, L.: Bidirectional recurrent convolutional networks for multi-frame super-resolution. Advances in neural information processing systems **28** (2015)

[23] Kappeler, A., Yoo, S., Dai, Q., Katsaggelos, A.K.: Video super-resolution with convolutional neural networks. IEEE transactions on computational imaging **2**(2), 109–122 (2016)

[24] Karger, D.N., Nobis, M.P., Normand, S., Graham, C.H., Zimmermann, N.E.: Chelsa-trace21k–high-resolution (1 km) downscaled transient temperature and precipitation data since the last glacial maximum. Climate of the Past **19**(2), 439–456 (2023)

[25] Karger, D.N., Wilson, A.M., Mahony, C., Zimmermann, N.E., Jetz, W.: Global daily 1 km land surface precipitation based on cloud cover-informed downscaling. Scientific Data **8**(1), 307 (2021)

[26] Katharopoulos, A., Vyas, A., Pappas, N., Fleuret, F.: Transformers are rnns: Fast autoregressive transformers with linear attention. In: International conference on machine learning. pp. 5156–5165. PMLR (2020)

[27] Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1646–1654 (2016)

[28] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

[29] Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., et al.: Learning skillful medium-range global weather forecasting. Science **382**(6677), 1416–1421 (2023)

[30] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4681–4690 (2017)

[31] Leinonen, J., Hamann, U., Nerini, D., Germann, U., Franch, G.: Latent diffusion models for generative precipitation nowcasting with accurate uncertainty quantification. arXiv preprint arXiv:2304.12891 (2023)

[32] Leinonen, J., Nerini, D., Berne, A.: Stochastic super-resolution for downscaling time-evolving atmospheric fields with a generative adversarial network. IEEE Transactions on Geoscience and Remote Sensing **59**(9), 7211–7223 (2020)

[33] Li, L., Carver, R., Lopez-Gomez, I., Sha, F., Anderson, J.: Seeds: Emulation of weather forecast ensembles with diffusion models. arXiv preprint arXiv:2306.14066 (2023)

[34] Liang, J., Cao, J., Fan, Y., Zhang, K., Ranjan, R., Li, Y., Timofte, R., Van Gool, L.: Vrt: A video restoration transformer. arXiv preprint arXiv:2201.12288 (2022)

[35] Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1833–1844 (2021)

[36] Liang, J., Fan, Y., Xiang, X., Ranjan, R., Ilg, E., Green, S., Cao, J., Zhang, K., Timofte, R., Gool, L.V.: Recurrent video restoration transformer with guided deformable attention. Advances in Neural Information Processing Systems 35, 378–393 (2022)

[37] Liang, J., Fan, Y., Xiang, X., Ranjan, R., Ilg, E., Green, S., Cao, J., Zhang, K., Timofte, R., Van Gool, L.: Recurrent video restoration transformer with guided deformable attention. In: Advances in Neural Information Processing Systems (2022)

[38] Liu, H., Ruan, Z., Zhao, P., Dong, C., Shang, F., Liu, Y., Yang, L., Timofte, R.: Video super-resolution based on deep learning: a comprehensive survey. Artificial Intelligence Review 55(8), 5981–6035 (2022)

[39] Mahecha, M.D., Gans, F., Brandt, G., Christiansen, R., Cornell, S.E., Fomferra, N., Kraemer, G., Peters, J., Bodesheim, P., Camps-Valls, G., et al.: Earth system data cubes unravel global multivariate dynamics. Earth System Dynamics 11(1), 201–234 (2020)

[40] Manduchi, L., Pandey, K., Bamler, R., Cotterell, R., Däubener, S., Fellenz, S., Fischer, A., Gärtner, T., Kirchler, M., Kloft, M., et al.: On the challenges and opportunities in generative ai. arXiv preprint arXiv:2403.00025 (2024)

[41] Mardani, M., Brenowitz, N., Cohen, Y., Pathak, J., Chen, C.Y., Liu, C.C., Vahdat, A., Kashinath, K., Kautz, J., Pritchard, M.: Generative residual diffusion modeling for km-scale atmospheric downscaling. arXiv preprint arXiv:2309.15214 (2023)

[42] Mooers, G., Pritchard, M., Beucler, T., Srivastava, P., Mangipudi, H., Peng, L., Gentine, P., Mandt, S.: Comparing storm resolving models and climates via unsupervised machine learning. Scientific Reports 13(1), 22365 (2023)

[43] Pandey, K., Mandt, S.: A complete recipe for diffusion generative models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4261–4272 (2023)

[44] Pandey, K., Pathak, J., Xu, Y., Mandt, S., Pritchard, M., Vahdat, A., Mardani, M.: Heavy-tailed diffusion models. arXiv preprint arXiv:2410.14171 (2024)

[45] Pandey, K., Rudolph, M., Mandt, S.: Efficient integrators for diffusion generative models. In: International Conference on Learning Representations (2024)

[46] Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., Kurth, T., Hall, D., Li, Z., Azizzadenesheli, K., et al.: Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. arXiv preprint arXiv:2202.11214 (2022)

[47] Price, I., Rasp, S.: Increasing the accuracy and resolution of precipitation forecasts using deep generative models. In: International conference on artificial intelligence and statistics. pp. 10555–10571. PMLR (2022)

[48] Price, I., Sanchez-Gonzalez, A., Alet, F., Ewalds, T., El-Kadi, A., Stott, J., Mohamed, S., Battaglia, P., Lam, R., Willson, M.: Gencast: Diffusion-based ensemble forecasting for medium-range weather. arXiv preprint arXiv:2312.15796 (2023)

[49] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with CLIP latents. arXiv preprint arXiv:2204.06125 (2022)

[50] Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., Fitzsimons, M., Athanassiadou, M., Kashem, S., Madge, S., et al.: Skilful precipitation nowcasting using deep generative models of radar. Nature 597(7878), 672–677 (2021)

[51] Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat, f.: Deep learning and process understanding for data-driven earth system science. Nature **566**(7743), 195–204 (2019)

[52] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)

[53] Rota, C., Buzzelli, M., Bianco, S., Schettini, R.: Video restoration based on deep learning: a comprehensive survey. Artificial Intelligence Review pp. 1–48 (2022)

[54] Rubner, Y., Tomasi, C., Guibas, L.J.: A metric for distributions with applications to image databases. In: Sixth international conference on computer vision (IEEE Cat. No. 98CH36271). pp. 59–66. IEEE (1998)

[55] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Gontijo-Lopes, R., Ayan, B.K., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding. In: Advances in Neural Information Processing Systems (2022)

[56] Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–14 (2022)

[57] Sajjadi, M.S., Vemulapalli, R., Brown, M.: Frame-recurrent video super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6626–6634 (2018)

[58] Salathé, E.P., Steed, R., Mass, C.F., Zahn, P.H.: A high-resolution climate model for the us pacific northwest: Mesoscale feedbacks and local responses to climate change. Journal of climate **21**(21), 5708–5726 (2008)

[59] Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. ArXiv **abs/2202.00512** (2022)

[60] Shi, S., Gu, J., Xie, L., Wang, X., Yang, Y., Dong, C.: Rethinking alignment in video super-resolution transformers. Advances in Neural Information Processing Systems **35**, 36081–36093 (2022)

[61] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning. pp. 2256–2265 (2015)

[62] Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)

[63] Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. Advances in Neural Information Processing Systems **32** (2019)

[64] Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: International Conference on Learning Representations (2021)

[65] Stevens, B., Satoh, M., Auger, L., Biercamp, J., Bretherton, C.S., Chen, X., Düben, P., Judt, F., Khairoutdinov, M., Klocke, D., et al.: Dyamond: the dynamics of the atmospheric general circulation modeled on non-hydrostatic domains. Progress in Earth and Planetary Science **6**(1), 1–17 (2019)

[66] Taillardat, M., Fougères, A.L., Naveau, P., De Fondeville, R.: Evaluating probabilistic forecasts of extremes using continuous ranked probability score distributions. International Journal of Forecasting **39**(3), 1448–1459 (2023)

[67] Tang, J., Niu, X., Wang, S., Gao, H., Wang, X., Wu, J.: Statistical downscaling and dynamical downscaling of regional climate in china: Present climate evaluations and future climate projections. Journal of Geophysical Research: Atmospheres **121**(5), 2110–2129 (2016)

[68] Teufel, B., Carmo, F., Sushama, L., Sun, L., Khaliq, M., Bélair, S., Shamseldin, A., Kumar, D.N., Vaze, J.: Physics-informed deep learning framework to model intense precipitation events at super resolution. Geoscience Letters **10**(1), 19 (2023)

[69] Voleti, V., Jolicoeur-Martineau, A., Pal, C.: Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. Advances in neural information processing systems **35**, 23371–23385 (2022)

[70] Vosper, E., Watson, P., Harris, L., McRae, A., Santos-Rodriguez, R., Aitchison, L., Mitchell, D.: Deep learning for downscaling tropical cyclone rainfall to hazard-relevant spatial scales. Journal of Geophysical Research: Atmospheres p. e2022JD038163 (2023)

[71] Wang, Z., Chen, J., Hoi, S.C.: Deep learning for image super-resolution: A survey. IEEE transactions on pattern analysis and machine intelligence **43**(10), 3365–3387 (2020)

[72] Yang, Q., Hernandez-Garcia, A., Harder, P., Ramesh, V., Sattegeri, P., Szwarcman, D., Watson, C.D., Rolnick, D.: Fourier neural operators for arbitrary resolution climate data downscaling. arXiv preprint arXiv:2305.14452 (2023)

[73] Yang, R., Srivastava, P., Mandt, S.: Diffusion probabilistic modeling for video generation. Entropy **25**(10), 1469 (2023)

[74] Yang, R., Yang, Y., Marino, J., Mandt, S.: Insights from generative modeling for neural video compression. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)

[75] Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European conference on computer vision (ECCV). pp. 286–301 (2018)

[76] Zhao, S., Song, J., Ermon, S.: Towards deeper understanding of variational autoencoding models. arXiv preprint arXiv:1702.08658 (2017)

[77] Zhou, L., Lin, S.J., Chen, J.H., Harris, L.M., Chen, X., Rees, S.L.: Toward convective-scale prediction within the next generation global prediction system. Bulletin of the American Meteorological Society **100**(7), 1225–1243 (2019)

[78] Zhu, Y., Zhang, K., Liang, J., Cao, J., Wen, B., Timofte, R., Van Gool, L.: Denoising diffusion models for plug-and-play image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1219–1229 (2023)

Precipitation Downscaling
with Spatiotemporal Video Diffusion

Supplementary Materials

A.1 Model Architecture

Our architecture is a conditional extension of the DDPM [20] and SR3 [56] models. As discussed in Sec. 2, Fig. 8 presents the architecture of the proposed *denoising* and *downscaling* networks, while Fig. 9 provides detailed description of each component. Before elaborating on the specifics, we define the naming conventions for the parameter choices adopted in this section:

- `ChannelDim`: `ResBlock` channel dimension for first contractive layer of UNet.

- `ChannelMultipliers`: channel dimension multipliers for subsequent contractive layers (including the first layer) in both the downscaling and denoising modules. The expansive layer multipliers follow in reverse.

- `ResBlock`: a standard ResBlock implementation consisting of two blocks, each with a weight-standardised convolution using a 3×3 kernel, Group Normalization over groups of 8 and SiLU activation, followed by a channel-adjusting 1×1 convolution.

- `Attention`: a standard implementation of quadratic or linear attention, incorporating 4 attention heads, each with a 32-dimensional representation. Using a 3×3 convolutional kernel, query, key, and value feature maps are generated, resulting in feature maps of size [B $\times$ T, H $\times$ C, X, Y], where B, T, C, X, and Y denote batch, time, channel, height, and width, respectively, with H representing the number of heads. These feature maps are rearranged to [BATCH, HEAD, CHANNEL, TOKEN]), facilitating self-attention between TOKENs of CHANNEL dimensions. The rearrangement order determines whether spatial or temporal self-attention is performed. Subsequently, the feature maps revert to their original format, upon which a separate convolution operation is conducted, followed by layer normalization to project the feature maps back to their original dimensions, akin to their state before the initial convolution within the attention block. Further elaboration on the rearrangement choices for the variants is discussed, adhering to *einops*[3] notation:

    - `Q-Spatial`: quadratic variant applied in the bottleneck layer; self attends between every pixel of the feature map with the following rearrangement : [B $\times$ T, H $\times$ C, X, Y] -> [B $\times$ T, H, C, X $\times$ Y]
    - `Q-Temporal`: quadratic variant applied in the bottleneck layer; self attends between feature maps across time with the following rearrangement : [B $\times$ T, H $\times$ C, X, Y] -> [B, H, C $\times$ X $\times$ Y, T]
    - `L-Spatial`: linear variant applied in expansive and contractive layers of UNet; self attends between every pixel within a patch of the feature map with the following rearrangement : [B $\times$ T, H $\times$ C, X $\times$ P, Y $\times$ P] -> [B $\times$ T, H $\times$ X $\times$ Y, C, P $\times$ P] where P is patch size, starting with 192, halving at each contractive layer and doubling at each expansive layer.
    - `L-Temporal`: linear variant applied in expansive and contractive layers of UNet; self attends between feature maps across time in a channel factorised manner with the following rearrangement : [B $\times$ T, H $\times$ C, X, Y] -> [B, H $\times$ X $\times$ Y, C, T]

- `MLP`: conditioning on the denoising step n is achieved through this block, which uses 32-dimensional random Fourier features, followed by a linear layer, GELU activation and another linear layer to transform the noise step to a higher dimension.

- `Cov/TransCov`: these are convolutional (3×3 kernel) downsampling and upsampling blocks that change the spatial size by a factor of 2.

Fig. 8 illustrates the interaction between the U-Net architecture of the denoising and downscaling networks. The top U-Net depicts the denoising network, while the bottom U-Net depicts the
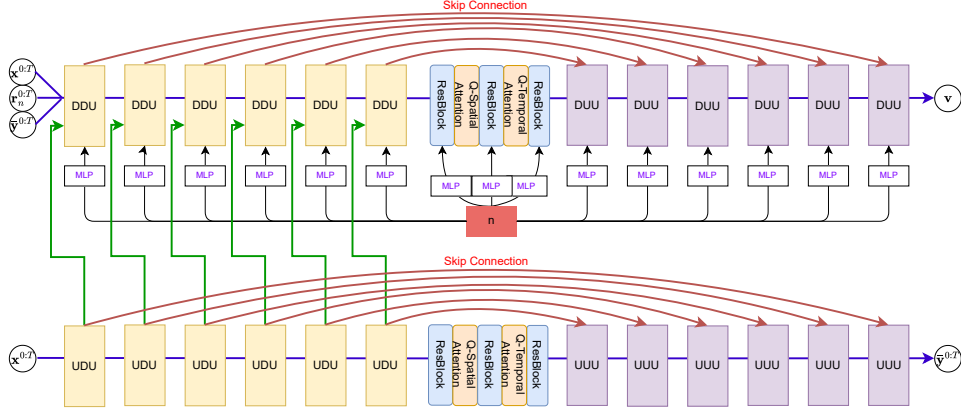
---

[3]`https://einops.rocks/1-einops-basics`

Figure 8: The figure depicts the overall model architecture where the top UNet performs diffusion on the residual, conditioned on the noise step n, $\mathbf{x}^{0:T}$, $\bar{\mathbf{y}}^{0:T}$ and the context of the bottom U-Net feature map (refer to Eq. (1) in Sec. 2.1). The bottom UNet is the deterministic downscaler. Details of each block are shown in Figure 9.

Figure 9: The details of the components of the modules shown in Figure 8; the colored arrows in the modules correspond to the arrows with the same color in Figure 8.

downscaling network. The denoising network is conditioned in three ways. First, the network is conditioned on the bicubically downscaled low-resolution frames $\mathbf{x}^{0:T}$, concatenated with both the noisy residual $\mathbf{r}_n^{0:T}$ the downscaler output and $\bar{\mathbf{y}}^{0:T}$ along the channel dimension. Second, the network is also conditioned on the feature maps generated by the downscaler network as indicated by the green arrows connecting the downsampling units of both the networks. The L-(Spatial/Temporal) Attention blocks from contractive layers of the downscaler network yield a feature map that gets concatenated with the inputs of both ResBlocks of the contractive layer of the denoising network. Finally, each contractive and expansive layer of the diffusion UNet gets conditioned on the denoising step n, shown via the **black** arrows. This conditional embedding for the step is generated through MLP and received by both ResBlocks. The information flows from the noisy residual through the network as shown by the blue arrows to predict the angular parameter $\mathbf{v}$. Both U-Nets have skip connections, indicated by the red arrows, between both ResBlocks of contractive and expansive layers of the same UNet.

Table 2: FV3GFS uses a cubed-sphere grid, in which the surface of the globe is divided into six tiles. Each high-resolution tile covers 25 km and is 384×384. Our UNet Encoder and Decoder have 6 layers with a base channel dimension of 64 and multipliers as stated above.

TileSize	ChannelDim	ChannelMultipliers
384×384	64	1,1,2,2,3,4

Figure 10: An illustration of the training and inference pipelines of Swin-IR-Diff. Similar to Fig. 2, blue blocks represent operations common to both training and inference phases. Red blocks signify operations exclusive to training, while green blocks indicate inference-only processes. However, in contrast, it is an image-only downscaler. This model takes in a current low-resolution frame which is deterministically downscaled via Swin-IR, followed by modeling of the residual details via conditional diffusion. The model details remain similar to what is described in A.1 with the absence of downscaler-conditioning and temporal attention in the diffusion model.

Table 3: Additional Variables in FV3GFS dataset.

Short Name	Long Name	Units
CPRATsfc	Surface convective precip. rate	$kg/m^2/s$
DSWRFtoa	Top of atmos. down shortwave flux	W/m^2
TMPsfc	Surface temperature	K
UGRD10m	10-meter eastward wind	m/s
VGRD10m	10-meter northward wind	m/s
ps	Surface pressure	Pa
u700	700-mb eastward wind	m/s
v700	700-mb northward wind	m/s
liq_wat	Vert. integral of cloud water mix ratio	$kg/kg \ kg/m^2$
sphum	Vert. integral of specific humidity	$kg/kg \ kg/m^2$
zsurf	Topography	–

A.2 On Swin-IR-Diff and Multiple Channels

Here, we discuss SwinIR-Diff, which expands on one of our robust baselines. Sec. 3 provides a concise overview of Swin-IR-Diff. Shown as a sketch in Fig. 10, this model opts to downscale each precipitation state individually, akin to an image super-resolution model. Resembling SR3 in its foundation of a conditional diffusion model, Swin-IR-Diff adopts a residual pipeline. It involves a deterministic prediction corrected by a residual generated from the conditional diffusion model, with the Swin-IR model serving as the deterministic downscaler in this context.

We conducted an ablation, focusing on the incorporation of additional climate states as input to our precipitation downscaling model STVD. The rationale for including these states is drawn from the insights of [17], who justified a similar selection for the task of precipitation forecast based on domain science. While Tab. 3 provides detailed information on the various states employed, the utility of these states is closely examined in Tab. 1, with specific attention to STVD (multiple input states) and STVD-single (only precipitation state as input). Clearly, the introduction of additional channels yields a notable improvement in performance.

A.3 Additional Samples

In addition to re-illustrating precipitation downscaling in the Sierras and Central California from Fig. 3 in Fig. 11), we present our model's output for another unique region—the Himalayas. Fig. 12 mirrors Fig. 11, displaying outputs from different models. Note that for the same regional topography, Fig. 6 (left) compares the annual precipitation time average. We also provide video samples corresponding to Fig. 11 (`california.gif`) and Fig. 12 (`himalaya.gif`) in the supplementary zip file.
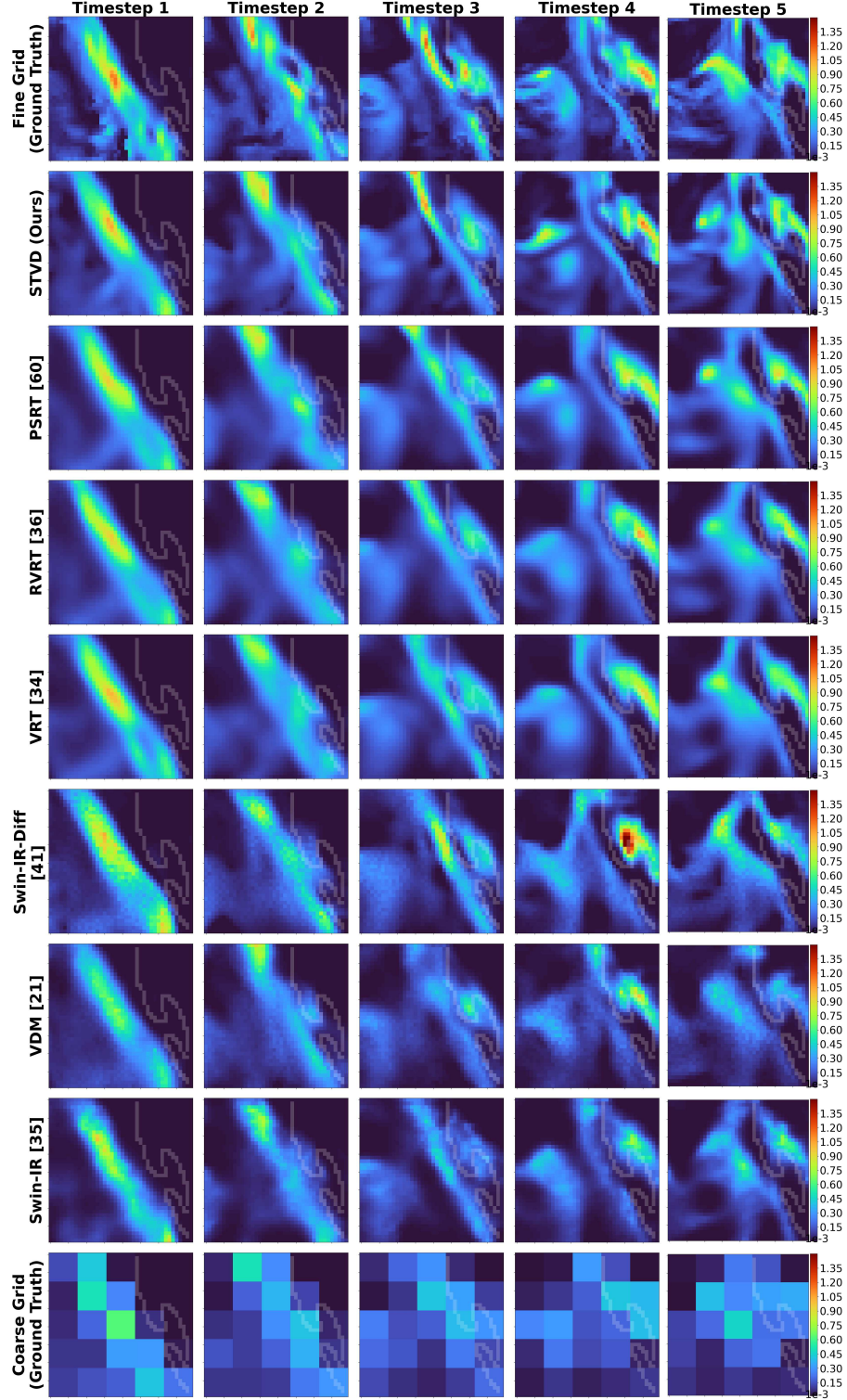
Figure 11: Qualitative comparison between our proposed model and all baselines for a specific precipitation event in the Sierra mountain range. This figure is a repetition of Fig. 3 for better visual overview. The first row represents the ground truth fine-grid precipitation state sequence, and the last row represents the coarse-grid precipitation that is being downscaled. All other rows correspond to our model and the baseline outputs. The time interval between adjacent frames is 3 hours; the plotted region is 1000×1000 km.
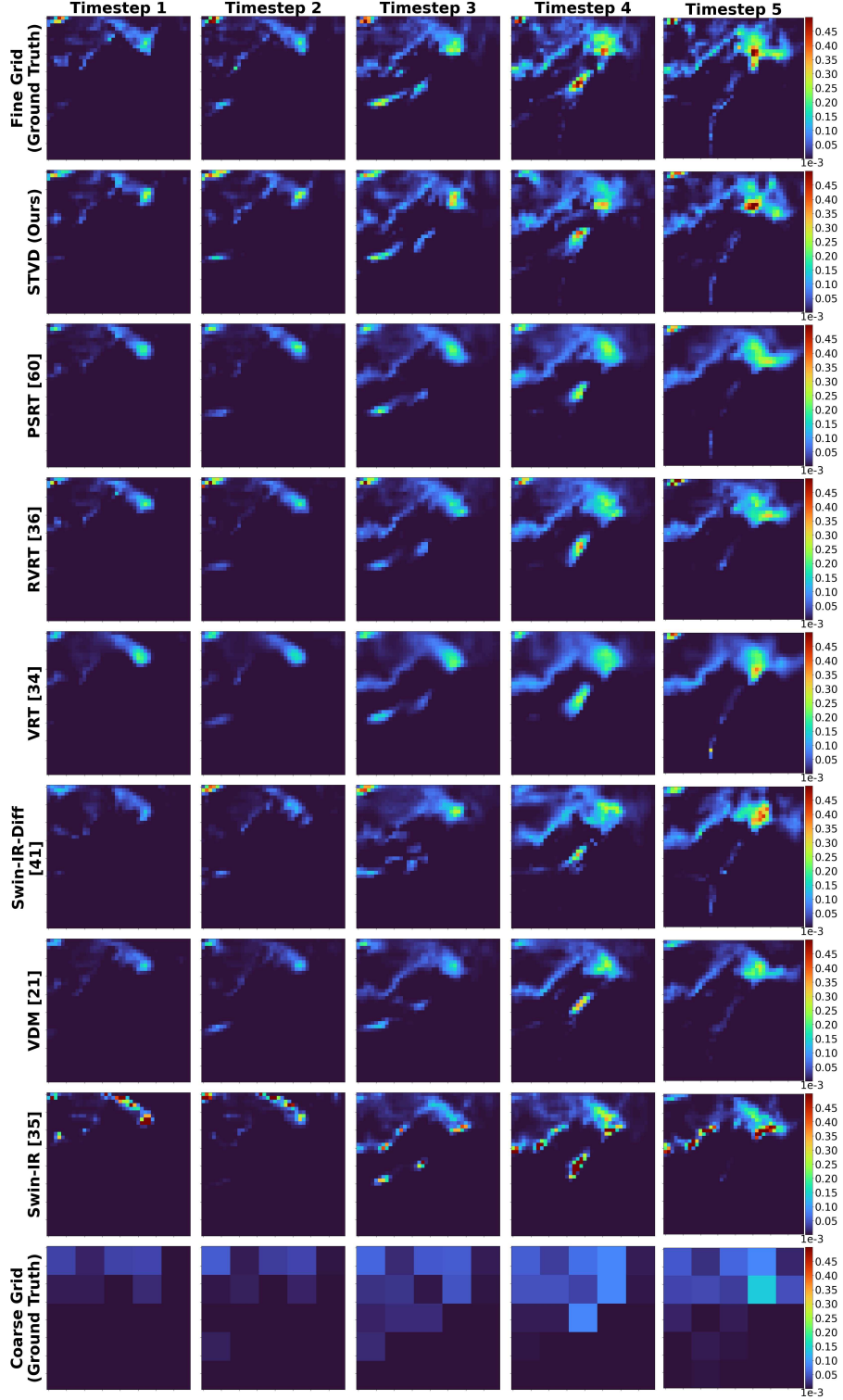
Figure 12: Another qualitative comparison between our proposed model and baselines for a specific precipitation event in the Himalayan mountain range. Fig. 6 (left) plots the regional topography. Similar to Fig. 3, the first row represents the ground truth fine-grid precipitation state sequence, and the last row represents the coarse-grid precipitation that is being downscaled. All other rows correspond to our model and the baseline outputs. The time interval between adjacent frames is 3 hours; the plotted region is 1000×1000 km.
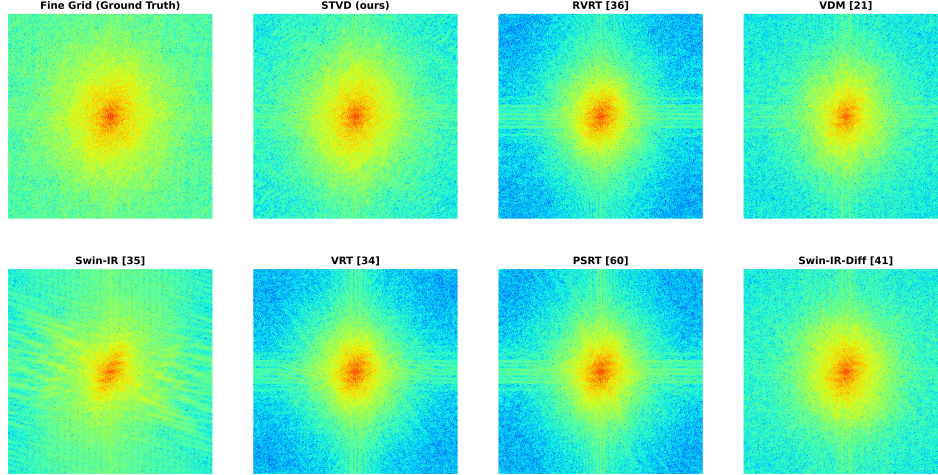
Figure 13: RVRT, VRT, PSRT and Swin-IR-Diff show a spectra which decays too rapidly, i.e. placing too little energy on the high-frequency components. The Swin-IR baseline exhibits artifacts in the spectra. Overall, the spectra of samples from our method most closely match the ground-truth spectra.

A.3.1 Spectra

In Fig. 13, we plot (in log-scale) the squared-magnitude of the complex-valued FFT applied to an image in our evaluation set. Overall, we see that the samples from STVD closely match the ground-truth high resolution spectrum. The baselines RVRT, VSRT and PSRT demonstrate a spectrum which decays too rapidly, placing too little energy in the high-frequency components. Additionally, we see a banding in these spectra. This indicates that these baselines are overly smooth compared to the ground-truth and follow similar banding patterns of the corresponding low resolution image. For the Swin-IR baseline, we observed outliers of large magnitudes in the generated precipitation maps, which we hypothesize leads to the observed checkerboard pattern seen in the spectrum. Swin-IR-Diff and VDM seem to decay the spectra more rapidly than STVD.

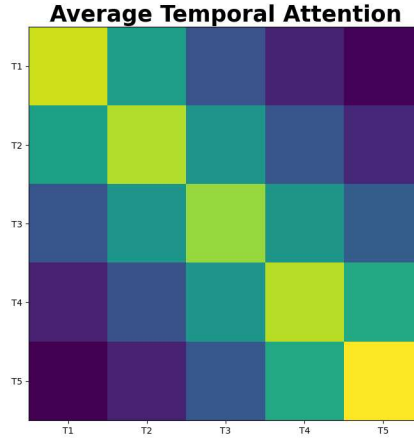### A.3.2 Temporal Attention Behaviour

Figure 14: A visualization of the temporal attention weights averaged over the entire validation set and attention heads for the bottleneck layer of the deterministic downscaler. $T1 - T5$ denotes the temporal sequence. The weights evidently decay as a function of temporal distance which makes physical sense. For example, feature map at position T2 attends the most towards itself along with immediate temporal neighbors at T1 and T3. Lighter colors correspond to larger weight.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: Sec. 3 outlines the experiments and Tab. 1 clearly demonstrates that the performance metrics for our model are better than the state-of-the-art baselines.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We discuss some limitations of our approach in Section 5.

   Guidelines:

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

3. **Theory Assumptions and Proofs**

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

   Answer: [NA]

Justification: This paper doesn't carry any theoretical results.

Guidelines:

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

4. **Experimental Result Reproducibility**

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

    Answer: [Yes]

    Justification: The architecture details have been discussed in Supplementary Materials A.1 and the experimental setup is discussed in Sec. 3.

    Guidelines:

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

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

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

8

Answer: [Yes]

Justification: The code for our model is available at `https://github.com/mandt-lab/STVD`

Guidelines:

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

6. **Experimental Setting/Details**

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

Answer: [Yes]

Justification: Experimental details can be found in Sec. 3.

Guidelines:

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

7. **Experiment Statistical Significance**

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

Answer: [No]

Justification: Training the models takes a long time and reporting meaningful error bars is computationally expensive.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

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

   Answer: [Yes]

   Justification: Details can be found in Sec. 3.

   Guidelines:

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

9. **Code Of Ethics**

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

   Answer: [Yes]

   Justification: Yes, our paper adheres to the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader Impacts**

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

    Answer: [Yes]

    Justification: The broader impacts our work are discussed in Section 5.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

Answer: [NA]

Justification: Our proposed model poses no such risk.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: All the reported baselines have been cited in Tab. 1 and Sec. 3.

Guidelines:

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

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

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

Answer: [NA]

Justification: We do not release any new assets. The details of the data and architecture can be found in Sec. 3 along with Supplementary Material A.1.

Guidelines:

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

14. **Crowdsourcing and Research with Human Subjects**

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

Answer: [NA]

Justification: No crowdsourcing or human subjects were involved.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

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

Answer: [NA]

Justification: No crowdsourcing or human subjects were involved.

Guidelines:

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