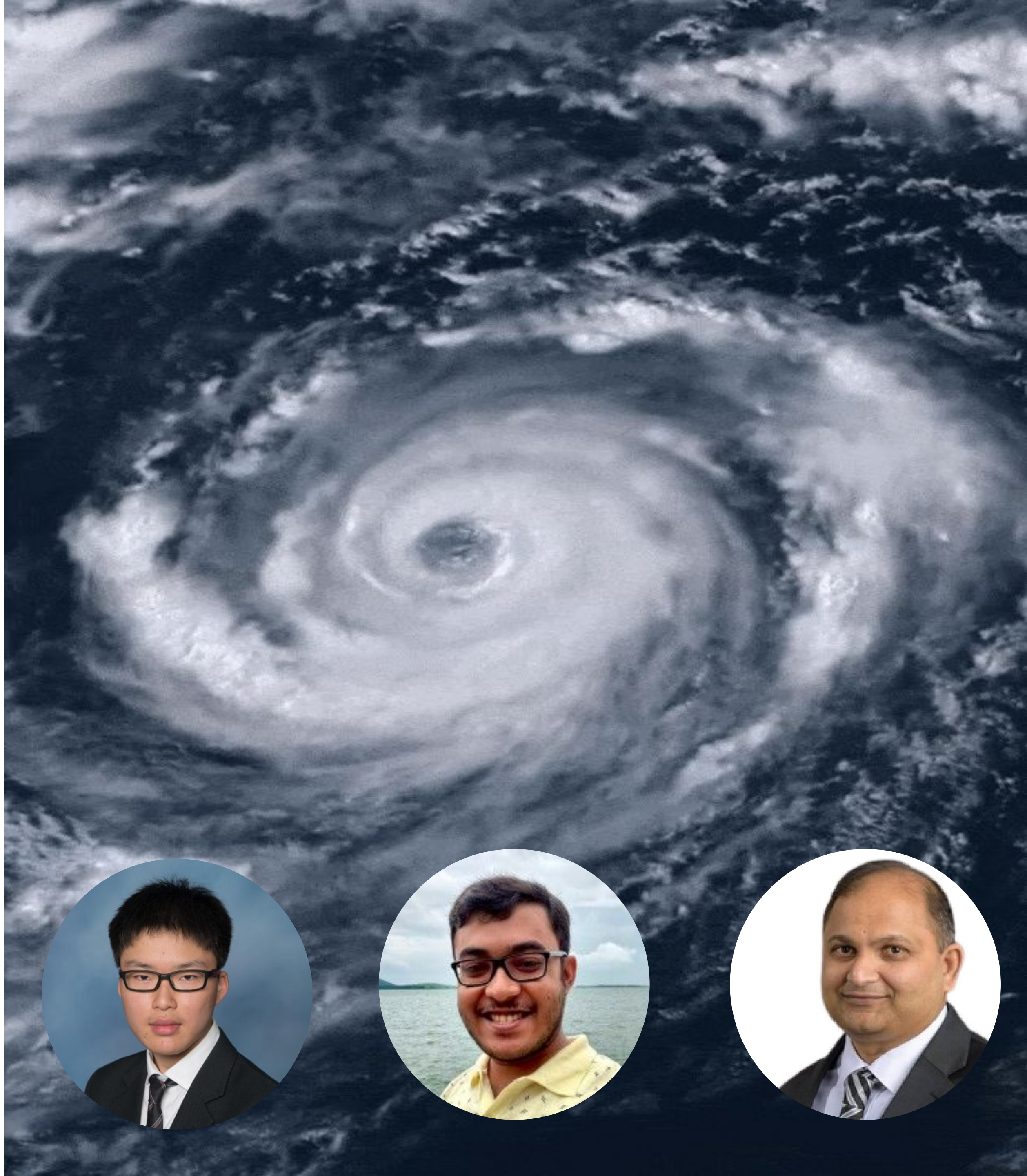
IMPERIAL

UNIVERSITY OF CAMBRIDGE

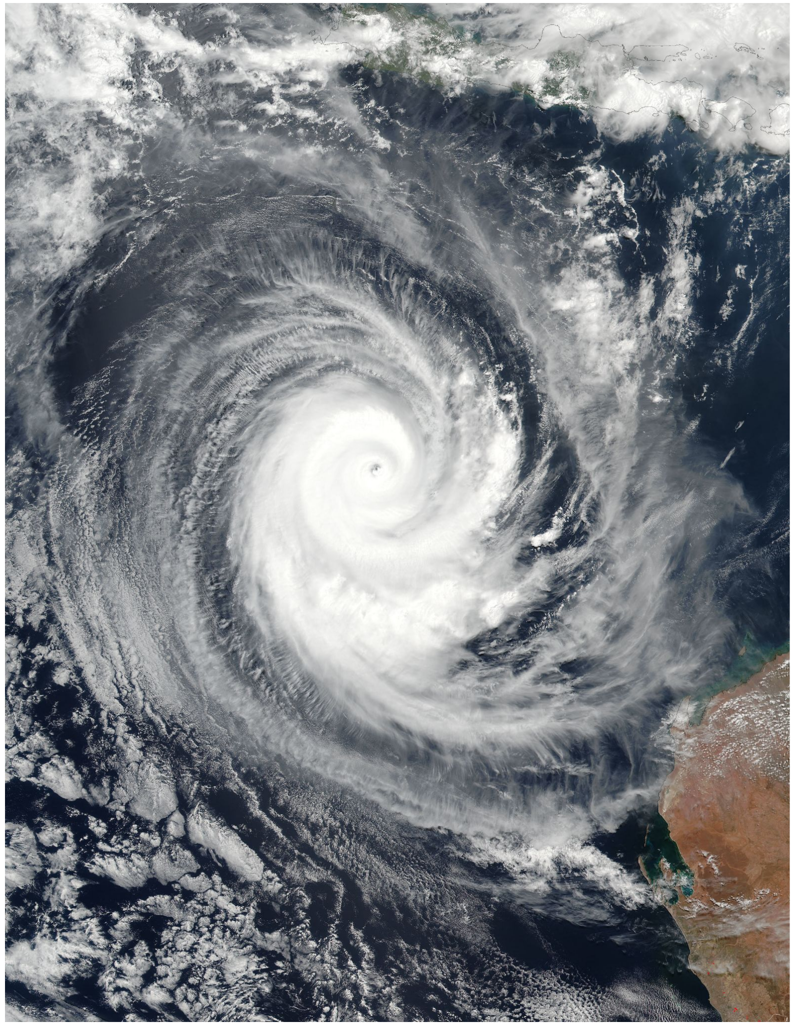
Improving Tropical Cyclone Forecasting With Video Diffusion Models

Zhibo Ren
Pritthijit Nath
Pancham Shukla



Introduction

Motivations



1. Tropical cyclones pose substantial risks to human life and infrastructure.

2. Climate change is driving unprecedented shifts in cyclone frequency, trajectory, and intensity.

3. While deep learning methods show promise, they often struggle to model temporal evolution effectively.

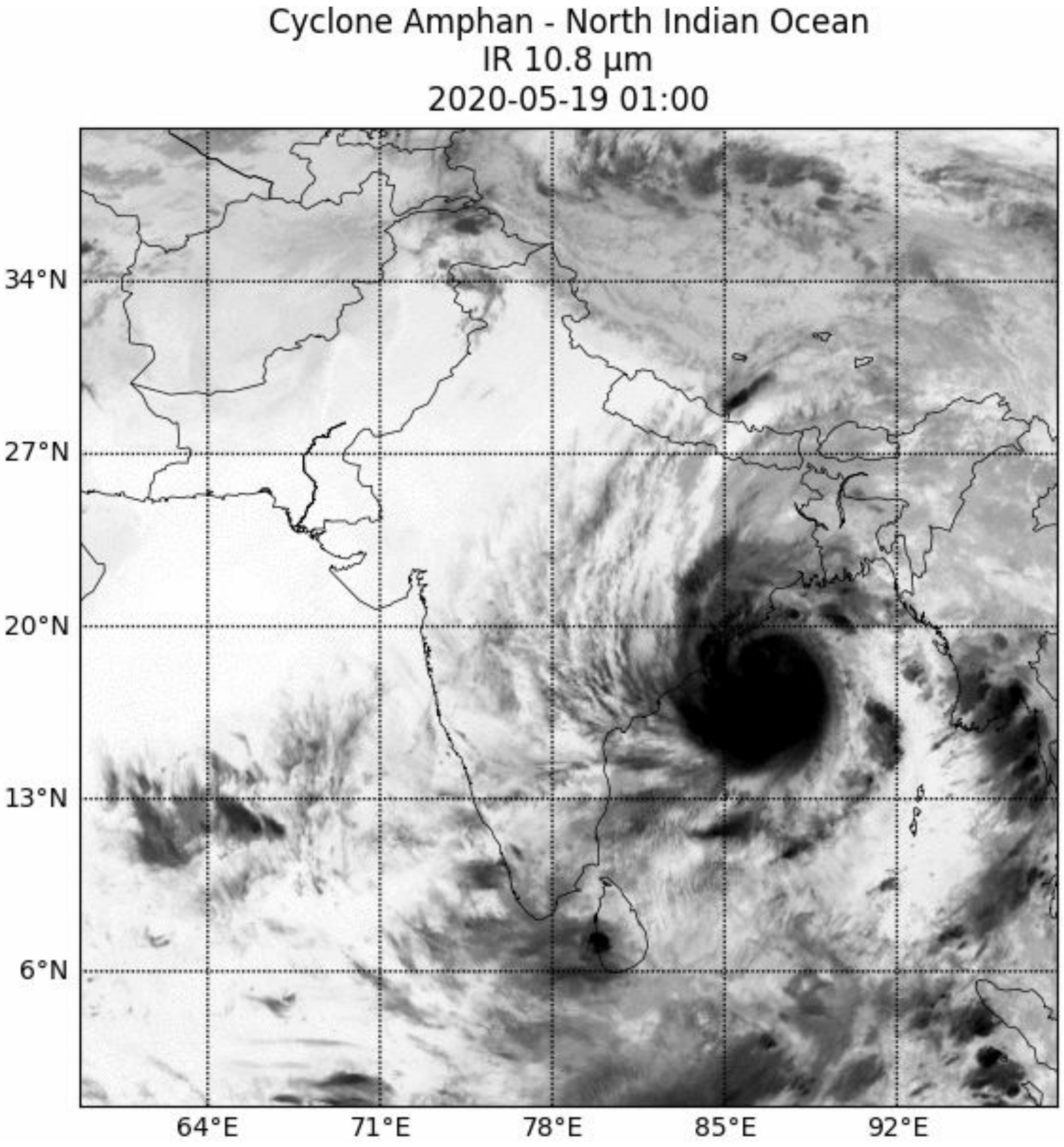
4. Video diffusion models provide a powerful alternative for capturing the spatiotemporal dynamics of cyclone development.

5. Transitioning from frame-wise prediction to coherent multi-frame generation enhances temporal consistency.

Introduction

Data

- **Dataset from Nath et al.[1]:** Infrared satellite imagery of tropical cyclones
- **Training data:** 1,092 video sequences, each comprising 10 consecutive frames
- **Test data:** 335 video sequences held out for evaluation
- **Corrupted data handling:** NaN values replaced with zeros
- **Data quality control:** Consistent masking applied during generation to preserve reliability



[1] Nath P, Shukla P, Wang S, Quilodrán-Casas C. Forecasting Tropical Cyclones with Cascaded Diffusion Models. arXiv; 2024. ArXiv:2310.01690. Available from: http://arxiv.org/abs/2310.01690.

Methods

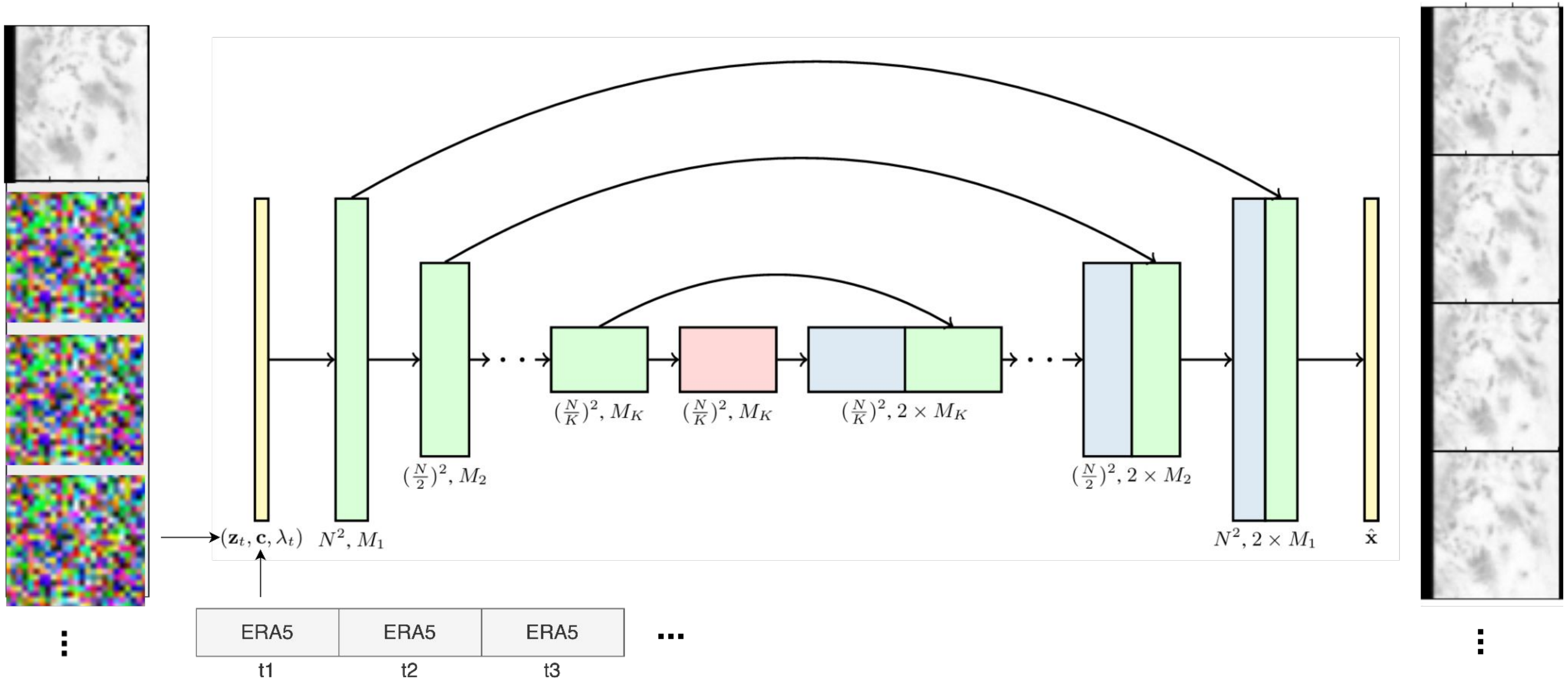
Video Diffusion Models

1. **3D U-Net architecture**, inspired by Ho et al.[2]
2. Leverages **temporal convolutions** and **attention mechanisms**
3. Generates **64×64 IR 10.8μm satellite imagery sequences**
4. **Conditioned** on the initial IR frame and **ERA5 meteorological variables**
5. **Simultaneously generates** 10 forecast frames
6. Incorporates **classifier-free guidance** and **dynamic thresholding** for improved generation quality

[2] Ho J, Salimans T, Gritsenko A, Chan W, Norouzi M, Fleet DJ. Video Diffusion Models. arXiv; 2022. ArXiv:2204.03458. Available from: http://arxiv.org/abs/2204.03458.

Methods

Video Diffusion Models



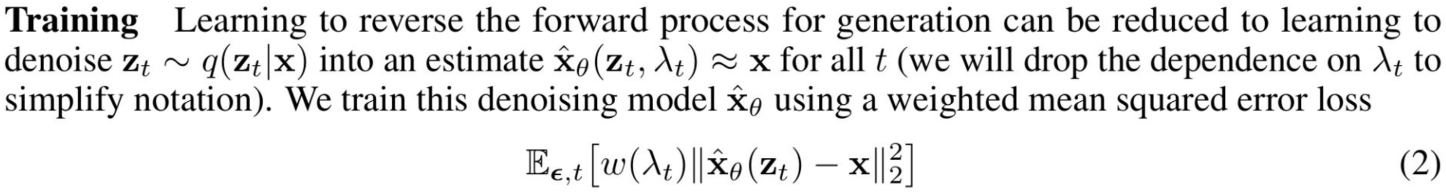
Architecture diagram adapted from Ho et al[2].

[2] Ho J, Salimans T, Gritsenko A, Chan W, Norouzi M, Fleet DJ. Video Diffusion Models. arXiv; 2022. ArXiv:2204.03458. Available from: http://arxiv.org/abs/2204.03458.

Methods

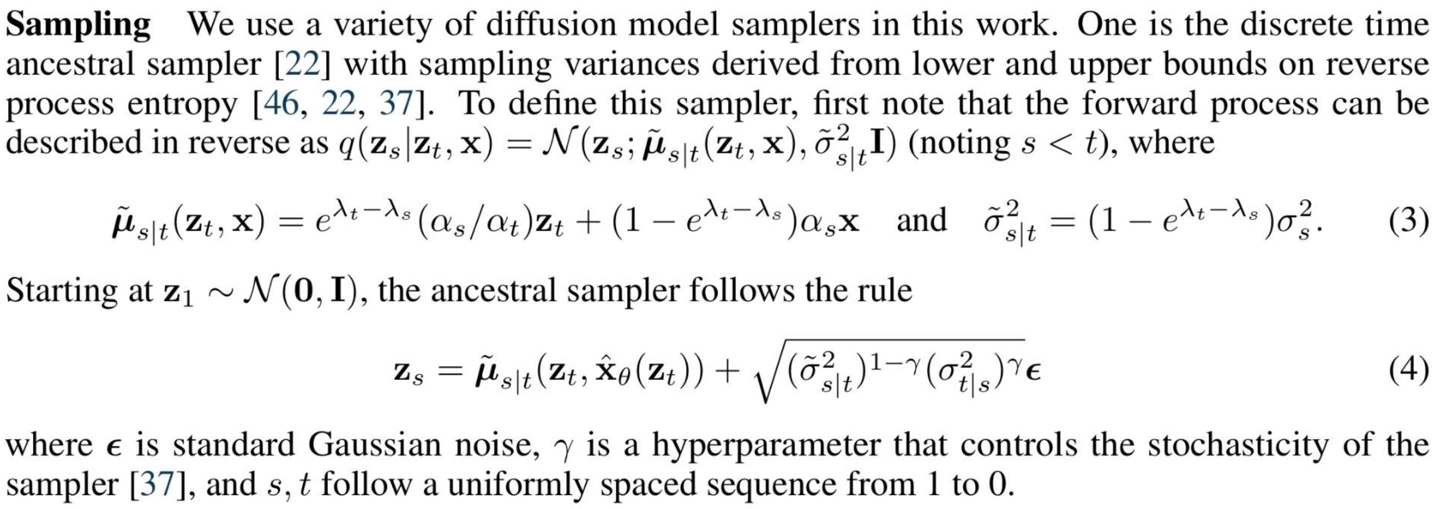
Video Diffusion Models

1. **Training objective:** Minimise a weighted mean squared error to learn to denoise latent inputs over time.
2. **Sampling strategy:** Use an ancestral sampler with adaptive variance and controlled noise injection.
3. **Classifier-free guidance:** Improve conditional generation by blending conditional and unconditional predictions with a tunable guidance strength.



Training Learning to reverse the forward process for generation can be reduced to learning to denoise $\mathbf{z}_t \sim q(\mathbf{z}_t|\mathbf{x})$ into an estimate $\hat{\mathbf{x}}_\theta(\mathbf{z}_t, \lambda_t) \approx \mathbf{x}$ for all t (we will drop the dependence on λ_t to simplify notation). We train this denoising model $\hat{\mathbf{x}}_\theta$ using a weighted mean squared error loss

$$\mathbb{E}_{\boldsymbol{\epsilon},t}\left[w(\lambda_t)\|\hat{\mathbf{x}}_\theta(\mathbf{z}_t) - \mathbf{x}\|_2^2\right] \tag{2}$$



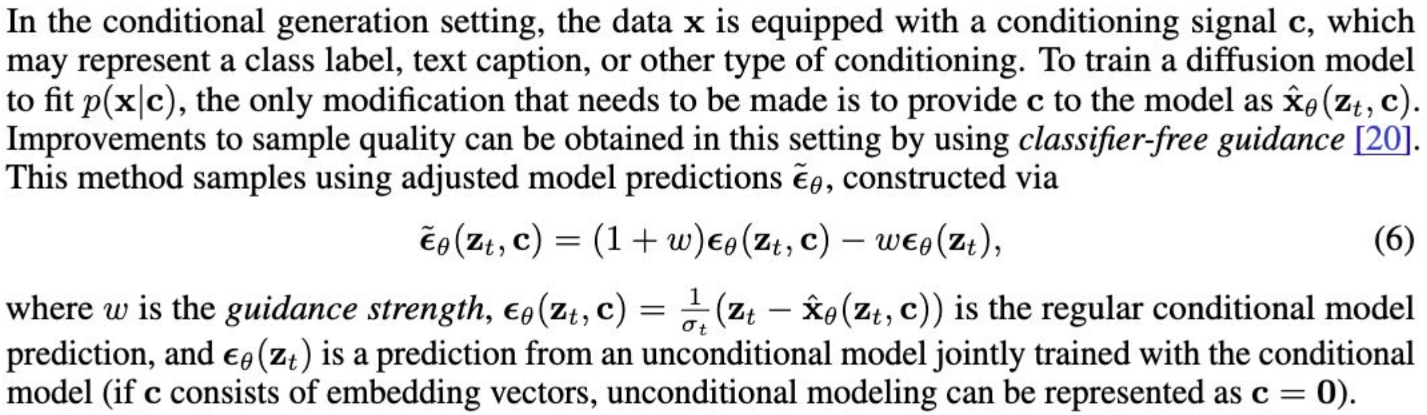
Sampling We use a variety of diffusion model samplers in this work. One is the discrete time ancestral sampler [22] with sampling variances derived from lower and upper bounds on reverse process entropy [46, 22, 37]. To define this sampler, first note that the forward process can be described in reverse as $q(\mathbf{z}_s|\mathbf{z}_t, \mathbf{x}) = \mathcal{N}(\mathbf{z}_s; \tilde{\boldsymbol{\mu}}_{s|t}(\mathbf{z}_t, \mathbf{x}), \tilde{\sigma}^2_{s|t}\mathbf{I})$ (noting $s < t$), where

$$\tilde{\boldsymbol{\mu}}_{s|t}(\mathbf{z}_t, \mathbf{x}) = e^{\lambda_t - \lambda_s}(\alpha_s/\alpha_t)\mathbf{z}_t + (1 - e^{\lambda_t - \lambda_s})\alpha_s\mathbf{x} \quad \text{and} \quad \tilde{\sigma}^2_{s|t} = (1 - e^{\lambda_t - \lambda_s})\sigma_s^2. \tag{3}$$

Starting at $\mathbf{z}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the ancestral sampler follows the rule

$$\mathbf{z}_s = \tilde{\boldsymbol{\mu}}_{s|t}(\mathbf{z}_t, \hat{\mathbf{x}}_\theta(\mathbf{z}_t)) + \sqrt{(\tilde{\sigma}^2_{s|t})^{1-\gamma}(\sigma^2_{t|s})^\gamma}\boldsymbol{\epsilon} \tag{4}$$

where $\boldsymbol{\epsilon}$ is standard Gaussian noise, γ is a hyperparameter that controls the stochasticity of the sampler [37], and s, t follow a uniformly spaced sequence from 1 to 0.



In the conditional generation setting, the data $\mathbf{x}$ is equipped with a conditioning signal $\mathbf{c}$, which may represent a class label, text caption, or other type of conditioning. To train a diffusion model to fit $p(\mathbf{x}|\mathbf{c})$, the only modification that needs to be made is to provide $\mathbf{c}$ to the model as $\hat{\mathbf{x}}_\theta(\mathbf{z}_t, \mathbf{c})$. Improvements to sample quality can be obtained in this setting by using *classifier-free guidance* [20]. This method samples using adjusted model predictions $\tilde{\boldsymbol{\epsilon}}_\theta$, constructed via

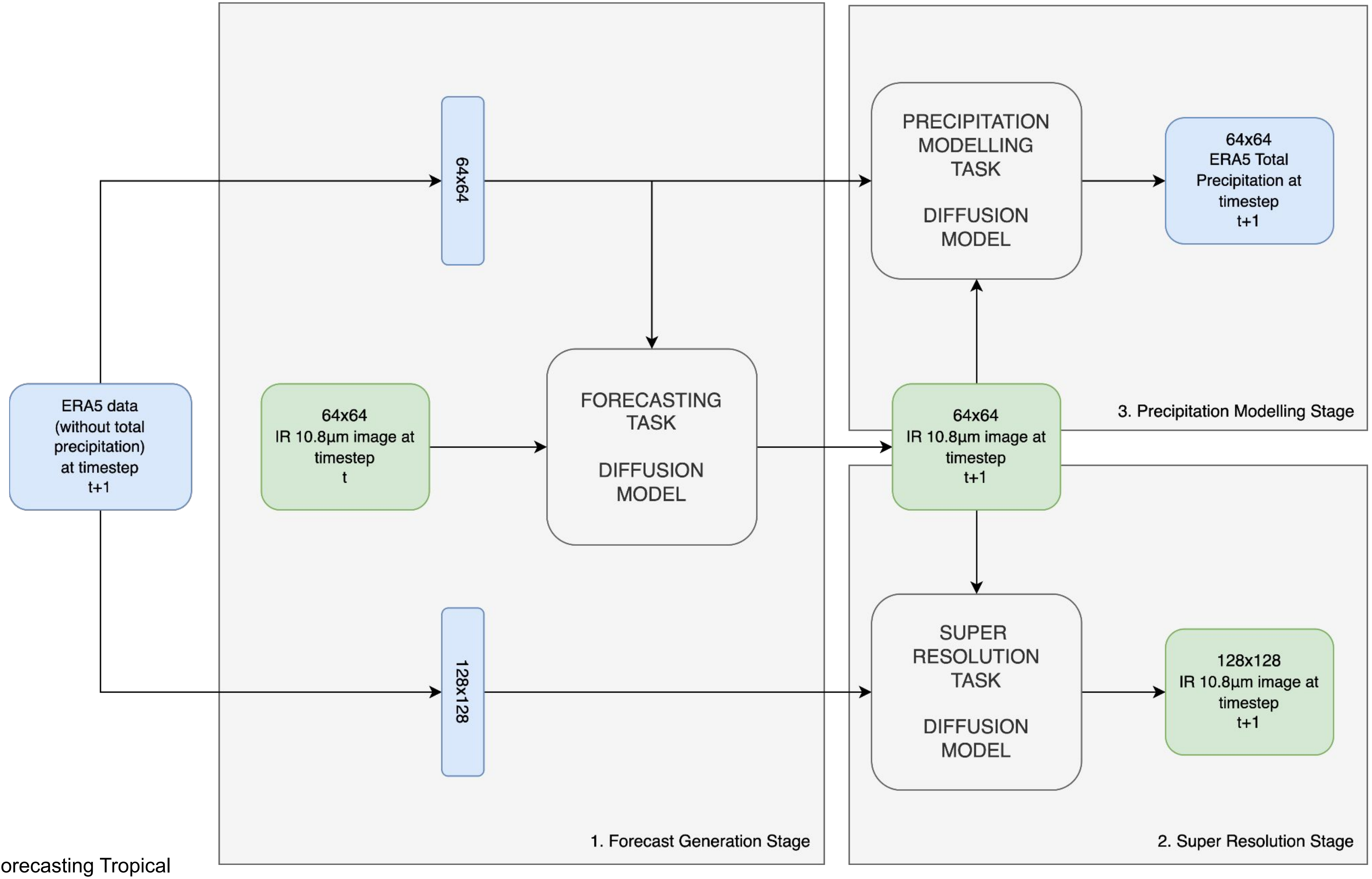
$$\tilde{\boldsymbol{\epsilon}}_\theta(\mathbf{z}_t, \mathbf{c}) = (1 + w)\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, \mathbf{c}) - w\boldsymbol{\epsilon}_\theta(\mathbf{z}_t), \tag{6}$$

where w is the *guidance strength*, $\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, \mathbf{c}) = \frac{1}{\sigma_t}(\mathbf{z}_t - \hat{\mathbf{x}}_\theta(\mathbf{z}_t, \mathbf{c}))$ is the regular conditional model prediction, and $\boldsymbol{\epsilon}_\theta(\mathbf{z}_t)$ is a prediction from an unconditional model jointly trained with the conditional model (if $\mathbf{c}$ consists of embedding vectors, unconditional modeling can be represented as $\mathbf{c} = \mathbf{0}$).

[2] Ho J, Salimans T, Gritsenko A, Chan W, Norouzi M, Fleet DJ. Video Diffusion Models. arXiv; 2022. ArXiv:2204.03458. Available from: http://arxiv.org/abs/2204.03458.

Methods
Baseline

Cascaded Diffusion Models by Nath et al.[1]



ERA5 data (without total precipitation) at timestep t+1
64x64
128x128
64x64 IR 10.8µm image at timestep t
FORECASTING TASK
DIFFUSION MODEL
1. Forecast Generation Stage
PRECIPITATION MODELLING TASK
DIFFUSION MODEL
64x64 ERA5 Total Precipitation at timestep t+1
3. Precipitation Modelling Stage
64x64 IR 10.8µm image at timestep t+1
SUPER RESOLUTION TASK
DIFFUSION MODEL
128x128 IR 10.8µm image at timestep t+1
2. Super Resolution Stage

[1] Nath P, Shukla P, Wang S, Quilodrán-Casas C. Forecasting Tropical Cyclones with Cascaded Diffusion Models. arXiv; 2024. ArXiv:2310.01690. Available from: http://arxiv.org/abs/2310.01690.

Methods

Two Stage Training

1. **Two-stage training:**
   Start with single-frame prediction, followed by multi-frame generation.

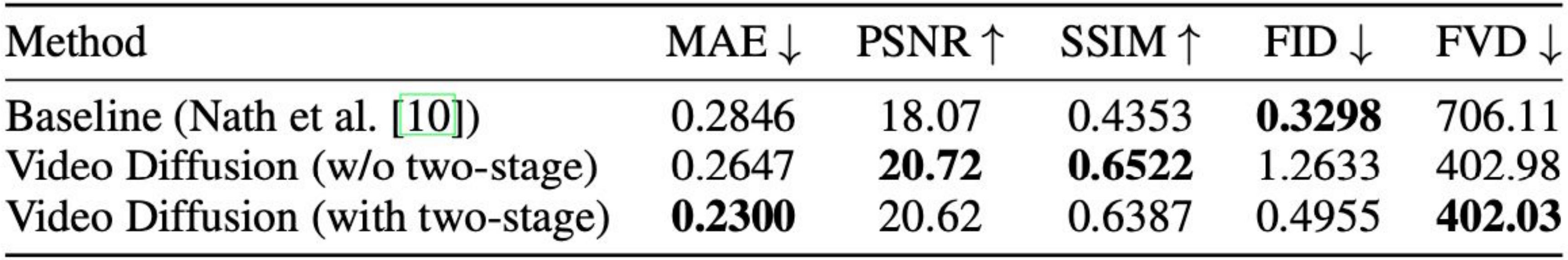
2. **Improves single-frame quality:**
   Especially important for capturing fine-grained spatial detail.

3. **Effective in low-data settings:**
   Reduces FID from 1.26 to 0.49, demonstrating enhanced fidelity.

Results

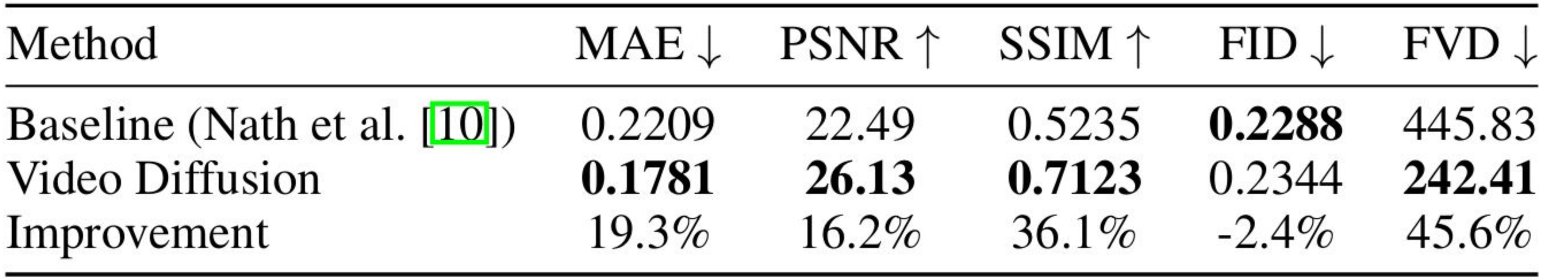
Performance Metrics

Table 1: Performance comparison in low-data regime



Method	MAE ↓	PSNR ↑	SSIM ↑	FID ↓	FVD ↓
Baseline (Nath et al. [10])	0.2846	18.07	0.4353	**0.3298**	706.11
Video Diffusion (w/o two-stage)	0.2647	**20.72**	**0.6522**	1.2633	402.98
Video Diffusion (with two-stage)	**0.2300**	20.62	0.6387	0.4955	**402.03**

Table 2: Comparison with baseline model on 10-frame prediction task



Method	MAE ↓	PSNR ↑	SSIM ↑	FID ↓	FVD ↓
Baseline (Nath et al. [10])	0.2209	22.49	0.5235	**0.2288**	445.83
Video Diffusion	**0.1781**	**26.13**	**0.7123**	0.2344	**242.41**
Improvement	19.3%	16.2%	36.1%	-2.4%	45.6%

Results

Visual Results

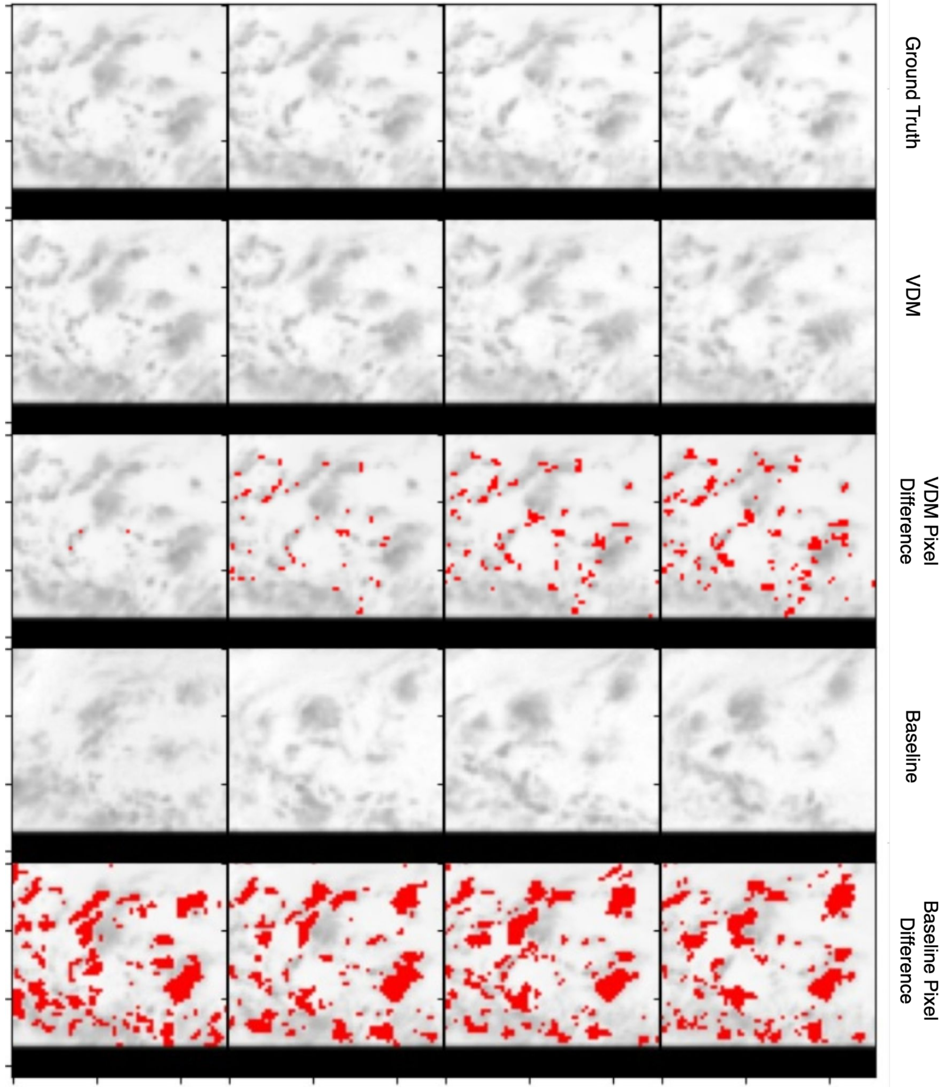
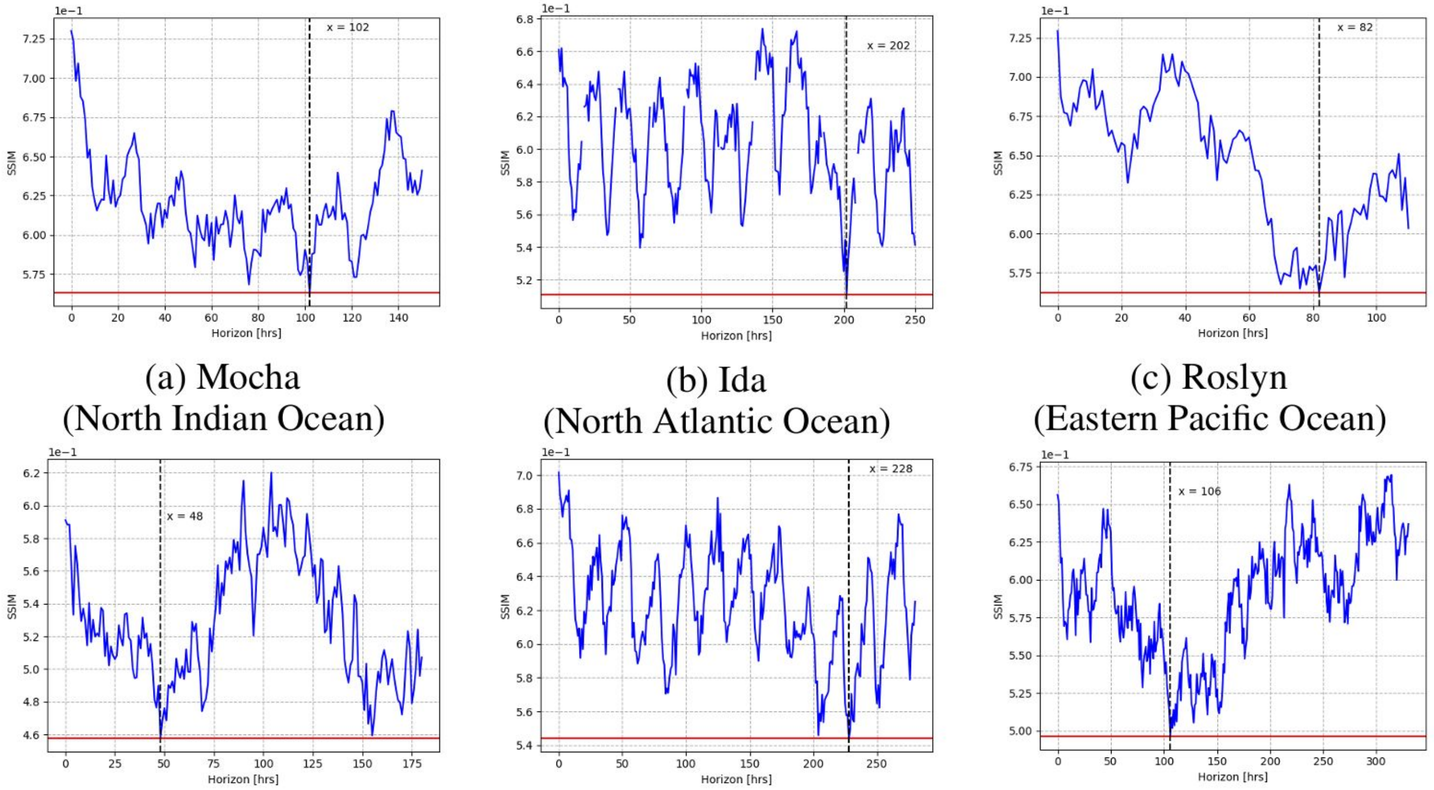


Figure 1: Qualitative comparison of TC forecasting results on the first four frames generated. From top to bottom: (1) Ground truth, (2) our VDM predictions, (3) the difference between VDM prediction and ground truth, (4) Nath et al.'s predictions, and (5) the difference between Nath et al.'s predictions and ground truth. Our VDM method demonstrates improved temporal consistency and more accurate TC evolution patterns.

Results

Long Horizon Forecasting

Comparison with the baseline: Superior long-horizon forecasting capability (from 36 to 50 hours)



(a) Mocha
(North Indian Ocean)

(b) Ida
(North Atlantic Ocean)

(c) Roslyn
(Eastern Pacific Ocean)

(d) Molave
(Western Pacific Ocean)

(e) Gombe
(SW Indian Ocean)

(f) Veronica
(Australia)

Figure B.1: SSIM values over the entire cyclonic duration. The dashed lines indicate the hourly marks at which the minimum SSIM values are obtained for each cyclone.

Conclusion

Key Takeaways

1. **First application of video diffusion models** for tropical cyclone (TC) forecasting
2. **Two-stage training** enhances model stability and single-frame quality
3. **Extended forecasting horizon:** from 36 to 50 hours with high reliability
4. **FVD introduced** as a more appropriate metric for evaluating TC forecast consistency
5. **45.6% improvement in temporal coherence** while maintaining competitive visual fidelity

Conclusion

Future Study

Given current limitations, several promising areas for future research include:

1. Integrating additional satellite channels beyond infrared (IR) channels
2. Extending generation beyond the current 10-frame limit
3. Investigating physics-informed loss functions for improved realism
4. Applying the framework to other weather forecasting tasks requiring temporal coherence
5. Exploring strategies for improving computational efficiency

IMPERIAL

UNIVERSITY OF CAMBRIDGE

THANKS!

Ren, Z. (2025). Improving Tropical Cyclone Forecasting With Video Diffusion Models. arXiv preprint arXiv:2501.16003