Probabilistic Weather Forecasting with Deterministic Guidance-based Diffusion Model

Donggeun Yoon*[1], Minseok Seo*[2], Doyi Kim[2], Yeji Choi[2], and Donghyeon Cho**[3]

[1] Korea Electronics Technology Institute, Seongnam, South Korea
[2] SI Analytics, Daejeon, South Korea
[3] Department of Computer Science, Hanyang University, Seoul, South Korea

Abstract. Weather forecasting requires both deterministic outcomes for immediate decision-making and probabilistic results for assessing uncertainties. However, deterministic models may not fully capture the spectrum of weather possibilities, and probabilistic forecasting can lack the precision needed for specific planning, presenting significant challenges as the field aims for enhance accuracy and reliability. In this paper, we propose the Deterministic Guidance-based Diffusion Model (DGDM) to exploit the benefits of both deterministic and probabilistic weather forecasting models. DGDM integrates a deterministic branch and a diffusion model as a probabilistic branch to improve forecasting accuracy while providing probabilistic forecasting. In addition, we introduce a sequential variance schedule that predicts from the near future to the distant future. Moreover, we present a truncated diffusion by using the result of the deterministic branch to truncate the reverse process of the diffusion model to control uncertainties. We conduct extensive analyses of DGDM on the Moving MNIST. Furthermore, we evaluate the effectiveness of DGDM on the Pacific Northwest Windstorm (PNW)-Typhoon satellite dataset for regional extreme weather forecasting, as well as on the WeatherBench dataset for global weather forecasting dataset. Experimental results show that DGDM achieves state-of-the-art performance not only in global forecasting but also in regional forecasting scenarios. The code is available at: `https://github.com/DongGeun-Yoon/DGDM`.

Keywords: Weather forecasting · Video prediction · Diffusion model

1 Introduction

Weather is a critical variable that impacts various aspects of daily life, including aviation, logistics, agriculture, and transportation. To provide accurate weather forecasting, the numerical weather prediction (NWP) model has been predominantly used for most weather forecasting since the 1950s [4, 7]. NWP employs a simulation-centric framework with vertical and horizontal grids that divide

* equal contribution
** corresponding author

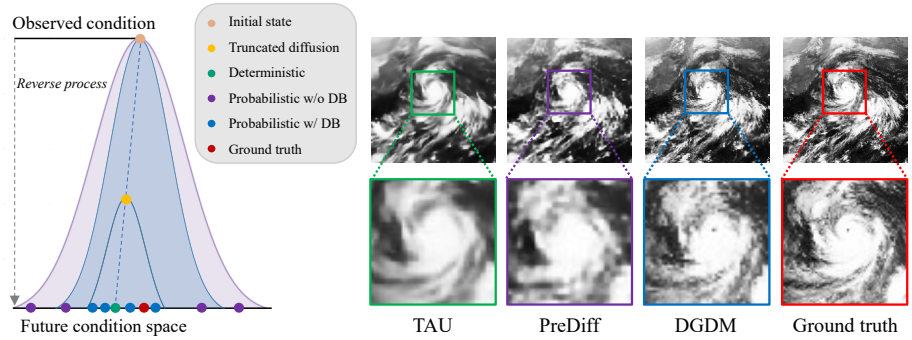

Fig. 1: Probabilistic models exhibit high diversity, posing challenges in selecting samples close to GT. In contrast, deterministic models produce a single output, restricting their capacity to capture weather possibility. DGDM provides precise probabilistic forecasting with deterministic guidance.

the Earth's atmosphere. Each grid cell translates the governing atmospheric behavior into partial differential equations (PDEs), which are then solved using numerical integral methods. Even a single 10-day forecasting simulation with the NWP model requires hours of computation across hundreds of supercomputer nodes. Despite these efforts, phenomena such as turbulent motion and tropical cumulus convection, which occur on a horizontal scales smaller than a few kilometers, cannot be captured by a single deterministic forecasting because they are smaller than the grid of the NWP model. Also, the nonlinearity and randomness inherent in atmospheric phenomena pose significant challenges to conducting accurate simulations [18].

To address the inherent uncertainty resulting from nonlinear and random atmospheric phenomena, NWP models add small random perturbations to the observed weather conditions and perform multiple simulations to consider the possible scenarios. This ensemble forecasting is effective in representing uncertain events. However, configuring numerous ensemble members for the NWP model is challenging because it requires a significant amount of computation and time to generate a single forecasting.

Several data-driven weather forecasting methods have been proposed [3, 19] to address the temporal and spatial resolution problems in NWP models. While all have surpassed the performance of NWP in 10-day global weather forecasting, they are fundamentally deterministic models that are difficult to account for uncertainties and therefore fail to capture the spectrum of weather possibilities. Recently, probabilistic weather forecasting models [10, 20] using the latent diffusion model [26] have been proposed for precipitation nowcasting. These models have shown success in data-driven short-term weather forecasting. However, as depicted in Fig. 1, probabilistic models tend to generate multiple outcomes with varying degrees of accuracy, which may result in the selection of samples that different significantly from the ground truth. Therefore, data-driven weather

forecasting encounters a trade-off problem where deterministic models have high performance but lack ensemble capabilities, and probabilistic models generate diverse samples but with lower accuracy.

In this paper, we introduce a Deterministic Guidance-based Diffusion Model (DGDM) for probabilistic weather forecasting. DGDM addresses the limitation that data-driven weather forecasting models are deterministic, and solves the problem that probabilistic weather forecasting models produce plausible futures rather than accurate forecasting. DGDM is structured into two branches: the deterministic branch and the probabilistic branch. In the training phase, the deterministic branch takes the observed weather condition as its input and aims to minimize discrepancies with the future weather condition. Simultaneously, the probabilistic branch uses both the observed weather condition and the future weather condition to train a direct mapping between the domains, using a Brownian bridge stochastic process [21]. To improve the quality of the video, we introduce a sequential variance schedule (SVS) that adjusts the diffusion step for each video frame. Since forecasting the immediate future is easier than forecasting the distant future, SVS shortens the diffusion steps for near-term forecasting while maintaining longer steps for more distant forecasting. In reverse process, the near-term futures, which have shorter diffusion steps become the conditions for later frames. In the inference phase, we can use the results of deterministic branch to obtain a pseudo intermediate state. This pseudo intermediate state is then employed to truncate the diffusion process, starting reverse process from the pseudo intermediate state instead of the initial state. Truncated diffusion not only allows DGDM to manage the inherent uncertainty in weather forecasting, allowing control over the spectrum of possible future weather, but also to reduce the number of diffusion steps without degrading performance.

For comparison with other baselines and analysis, we experiment with the synthetic dataset Moving MNIST [28]. We release the Pacific Northwest Windstorm (PNW)-Typhoon weather satellite dataset to verify the effectiveness of DGDM in regional extreme weather forecasting. Then, we validate DGDM for global weather forecasting using the WeatherBench [25] dataset. Our evaluation results show that DGDM achieves state-of-the-art performance in both low-resolution global and high-resolution regional weather forecasting.

Our contributions are summarized as follows:

- We present DGDM, a diffusion model that integrates a deterministic model for high-accuracy probabilistic weather forecasting.
- We propose SVS that generates from near-future to distant-future frames in order, with the near-futures acting as a condition of the later frames to improve the video quality.
- We present a truncated diffusion that controls the diversity of potential future weather conditions using the results of the deterministic branch.
- We introduce PNW-Typhoon dataset designed to evaluate regional extreme weather forecasting. We also rigorously evaluate the effectiveness of DGDM using the Moving MNIST, PNW-Typhoon, and WeatherBench datasets.

2 Related Work

2.1 Data-driven Weather Forecasting

Data-driven weather forecasting is gaining significant attention for its ability to provide accuracy comparable to NWP models even without supercomputing resources. In particular, in many countries where supercomputers cannot be operated, data-driven weather forecasting is a new paradigm that can perform weather forecasting with a single GPU server.

Shi *et al.* [27] introduced a ConvLSTM to predict precipitation via autoregressive inference. Furthermore, Ayzel *et al.* [1] and Trebing *et al.* [31] presented a data-driven short-term precipitation forecasting leveraging the U-net architecture. Beyond precipitation, studies such as [3, 6, 19] have proposed data-driven models for forecasting global climate variables. The Fourier-based neural network model [6] was designed to generate global data-driven forecasting for atmospheric variables. Subsequently, models such as GraphCast [19] and Pangu-Weather [3] have outperformed NWP in 10-day forecasting. While there have been significant advancements, these data-driven forecasting models are deterministic. This means that they are difficult to represent nonlinear and uncertainty phenomena.

2.2 Video Frame Prediction

Video frame prediction has various applications such as weather forecasting [11], human motion prediction [17], traffic flow prediction [37] and human robot interaction [8]. Video frame prediction is primarily divided into two main categories: autoregressive and non-autoregressive methods. Traditionally, autoregressive methods, employing architectures such as ConvLSTM [27] and RNN [12, 33, 34], have been foundational in the field of video frame prediction. Notably, PhyDNet was introduced in [12], which consists of a two-branch deep architecture designed to disentangle physical dynamics from unknown factors. PhyDNet achieved state-of-the-art performance across multiple datasets through a recurrent physical cell (PhyCell) that executes PDE-constrained prediction in a latent space. Despite various advances, autoregressive methods inherently suffer from deteriorating performance due to error accumulation with increasing lead time[4].

To tackle the problem of error accumulation inherent in autoregressive models, recent research has turned towards non-autoregressive approaches [9, 16, 23, 29,38]. For instance, Gao *et al.* [9] designed a non-autoregressive model featuring a multi-in-multi-out structure called SimVP to prevent error accumulation over specified target lead times. However, most video frame prediction models were developed based on minimizing the mean squared error (MSE) between predictions and ground truth, which is unsuitable for probabilistic weather forecasting.

[4] **Lead time**: The time interval between the beginning and end of weather forecast.

2.3 Video Generation

Video generation technology has made great progress by applying a variety of approaches, including GAN [22], VAE [2, 24], and diffusion models [14, 15, 32]. Babaeizadeh *et al*. [2] argued for the importance of incorporating uncertainty in real-world video prediction, utilizing VAE to effectively manage this uncertainty. Chatterjee *et al*. [5] introduced a Bayesian network and hierarchical framework designed to address real-world uncertainty. After denoising diffusion probabilistic models (DDPM) [13], there has been notable progress in the video generation field. Ho *et al*. [14] proposed a video diffusion model (VDM) that generates videos by gradually denoising noisy videos. VDM is trained to maximize the variational lower bound of the log-likelihood of the data and achieves impressive results on video generation. Afterward, RaMViD [15], which proposed random-mask video diffusion, and MCVD [32], which integrated video prediction, generation, and interpolation into one framework, were proposed. While these probabilistic models can generate plausible future frames, they are not optimized for accurate forecasting.

Therefore, the direct use of video generation models for weather forecasting is unsuitable. Gao *et al*. [10] and Leinonen *et al*. [20] have attempted precipitation nowcasting through latent diffusion model [26]. Nevertheless, the use of ensemble methods poses significant challenges due to their high diversity. The proposed DGDM addresses this issue by effectively merging probabilistic and deterministic models, offering a truncated diffusion method that can still control the range of possible future weather scenarios.

3 Method

The objective of DGDM is to forecast future weather condition $y \in \mathbb{R}^{C\times\hat{L}\times H\times W}$ given the observed weather condition $x \in \mathbb{R}^{C\times L\times H\times W}$. Where H, W, and C represent the height, width, and channels, respectively. L and $\hat{L}$ denote the lengths of the observed and target frames. As shown in Fig. 2, DGDM consists of two branches: a deterministic branch and a probabilistic branch. Furthermore, the results of the deterministic branch are used to modulate uncertainties, which helps to control the diversity of possible future weather in the probabilistic branch.

3.1 Deterministic Branch

The deterministic branch (DB) adopts a non-autoregressive structure, which is beneficial for achieving high reconstruction accuracy as it prevents error accumulation at fixed lead times. Following previous architecture of video frame prediction models [9, 29], we comprise DB as an encoder $e(\cdot)$, a translator $st(\cdot)$, and a decoder $d(\cdot)$ structure. Given an input x, the loss function of DB is formulated as:

$$L_{DB} = \|y - d(st(e(x)))\|^2. \tag{1}$$

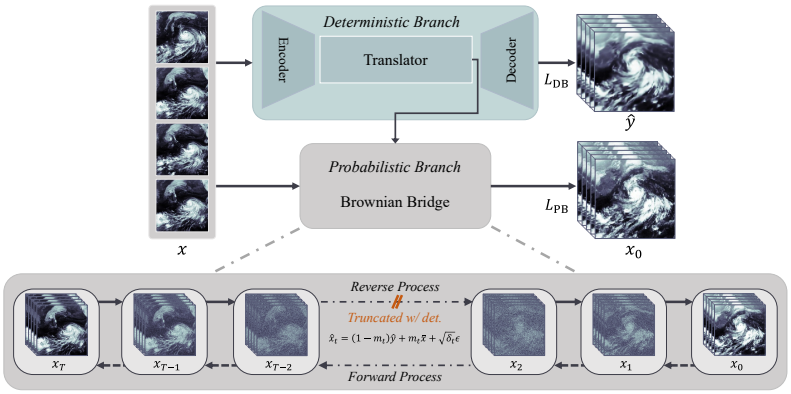

Fig. 2: Overview of DGDM: During training, DGDM simultaneously trains both the deterministic and probabilistic models. During inference, the results of the deterministic model are used to truncate the reverse process of the probabilistic model.

3.2 Probabilistic Branch

The probabilistic forecasting method in NWP is to apply different small random perturbations to the observed conditions each time the model is run, thereby generating diverse outcomes. Inspired by the NWP process, which produces probabilistic predictions through deterministic start and end points coupled with perturbations to the observed conditions, we adopt a diffusion model using Brownian bridge as the probabilistic branch (PB) of the DGDM.

Brownian Bridge Diffusion Process When we consider the problem as one of predicting the stochastic trajectories between an observed weather condition x and a future weather condition y, a Brownian bridge can be applied as a continuous-time stochastic model where the probability distribution during the diffusion process is conditional on the starting and ending states. Specifically, the state distribution at each time step of a Brownian bridge process starting from point $y = x_0$ with $x_0 \sim q_{\mathrm{data}}(x_0)$ at $t = 0$ and ending at point $x = x_T$ at $t = T$ can be formulated as:

$$p(x_t|x_0, x_T) = \mathcal{N}\left((1 - \frac{t}{T})x_0 + \frac{t}{T}x_T, \frac{t(T-t)}{T}\mathbf{I}\right), \tag{2}$$

where $\mathcal{N}$ denotes the normal distribution and $\mathbf{I}$ is the identity matrix that scales the variance of the distribution. We can define the Brownian bridge diffusion process in the simplified notation of DDPM as follows:

$$\begin{aligned} q(x_t|x_0, x_T) &= N(x_t; (1 - m_t)x_0 + m_t x_T, \delta_t I), \\ x_0 = y, m_t &= \frac{t}{T}. \end{aligned} \tag{3}$$

Here, m_t denotes a time weight, linearly increasing from 0 to 1 with respect to t. Concurrently, δ_t is a variance, calculated as $2(m_t - m_t^2)$. Note that, as in the

case of image to image translation tasks where the Brownian bridge diffusion process has been successfully applied BBDM [21].

Deterministic Guidance-based Diffusion Model Compared to general diffusion models that generate from noise, Brownian bridge diffusion process generates from input. Therefore, using input x directly leads to two main issues. First, the target length $\hat{L}$ must match the input length L for the diffusion model to function properly, which complicates the control of output length $\hat{L}$. Second, there is a large time interval between the first frame of the input and the first frame of the target, and the other frames have the same time interval. This makes the direct use of x to perform Eq. (3) inefficient. To address these issues, we use only the last frame (LF) of x for PB. This approach proves to be efficient because the amount of change from the last frame to the future frame is minimal, and it eliminates the constraint on the length of outputs by using the last frame as many times as $\hat{L}$. Consequently, the forward process of PB is defined as follows:

$$\begin{aligned} x_t &= (1 - m_t)x_0 + m_t\bar{x} + \sqrt{\delta_t}\epsilon, \\ x_{t-1} &= (1 - m_{t-1})x_0 + m_{t-1}\bar{x} + \sqrt{\delta_{t-1}}\epsilon. \end{aligned} \tag{4}$$

Here, x_0 is the future frame y and $\bar{x}$ denotes the input of PB, which replicate LF of x as many as L times. Concurrently, ϵ corresponds to Gaussian noise, specifically $N \sim (0, 1)$. However, as $\bar{x}$ consists of identical frames, spatial-temporal modeling is not feasible. To resolve this issue, we employ the feature z extracted from $st(e(x))$ in the DB to perform cross-attention [26] as a condition in PB. Therefore, PB is trained using following function.

$$L_{\text{PB}} = \mathbb{E}\left|\left|m_t(\bar{x} - x_0) + \sqrt{\delta_t}\epsilon - \epsilon_\theta(x_t, t, z)\right|\right|^2. \tag{5}$$

The DB and PB are jointly trained in an end-to-end manner, and the total objective function for DGDM is defined as follows:

$$L_{\text{total}} = \lambda_{PB}L_{\text{PB}} + \lambda_{DB}L_{\text{DB}}, \tag{6}$$

where λ_{PB} and λ_{DB} are both set to 1.

Sequential Variance Schedule Forecasting the immediate future is easier due to the minimal change from the given frame, while uncertainty increases with longer lead times. We therefore introduce a sequential variance schedule (SVS), which varies the diffusion step for each frame accordingly. As depicted in Fig. 3, the near future is relatively predictable, thus we set a short diffusion step to complete it first, and then sequentially complete the far future. Since the completed near frames are subsequently used as a condition for generating the far future, SVS contribute to improve the quality of the video. Given the total number of diffusion steps T, the output length $\hat{L}$, and stride of diffusion steps per frame S, the diffusion step for each frame index i is defined by the equation:

$$\text{SVS} = \{T - (\hat{L} - i) \cdot S : i = 1, \ldots, \hat{L}\}. \tag{7}$$

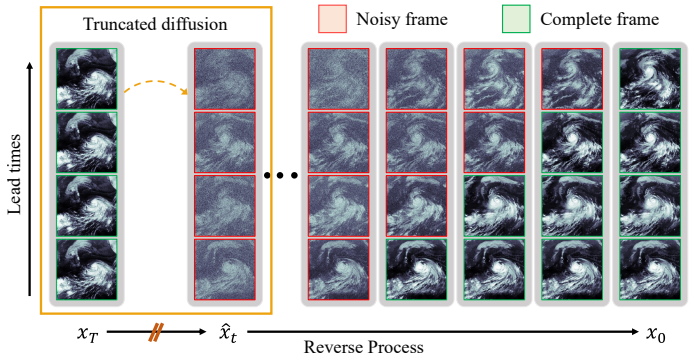

Fig. 3: Reverse process of DGDM. The frames with shorter lead time are generated before the frames with longer lead time and become the condition of the later frames. The reverse process is truncated by the pseudo intermediate state $\hat{x}_t$.

3.3 Inference

For probabilistic forecasting, the reverse process of the Brownian bridge starts at $x_T = \bar{x}$. Similar to the reverse process of DDPM, we then proceed through the reverse process that incrementally moves from x_t to x_{t-1} with the trained model.

$$p_\theta(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t, t, z), \tilde{\delta}_t I), \tag{8}$$

where $\mu_\theta(x_t, t, z)$ denotes the estimated mean value of the noise, and $\tilde{\delta}_t$ is the variance of noise. We reach the ending point of the diffusion process, where $x_0 = y$. For acceleration, the inference process integrates a non-Markovian chain, as detailed in BBDM [21].

Truncated Diffusion with Deterministic By incorporating the results of DB $\hat{y}$ into the reverse process, we truncate the diffusion process, as illustrated in Fig. 3. Since DB is trained to minimize the different between prediction and ground truth, $\hat{y}$ closely resembles y. By substituting $\hat{y}$ for x_0 in Eq. (4), we obtain a pseudo intermediate state $\hat{x}_t$.

$$\hat{x}_t = (1 - m_t)\hat{y} + m_t\bar{x} + \sqrt{\delta_t}\epsilon. \tag{9}$$

Note that $\hat{x}_t$ is not precisely identical to x_t, but the Brownian bridge is a stochastic process, thus we use $\hat{x}_t$ in the reverse process without additional training. By starting the reverse process with $\hat{x}_t$ instead of x_T and reducing the number of diffusion steps, we not only control the diversity but also improve the inference speed.

4 Experiments

4.1 Experimental Setting

Dataset We experiment on two weather datasets: PNW-Typhoon and Weatherbench [25] datasets. For comparison with other baselines and analysis, we experiment with the synthetic dataset Moving MNIST [28].

- **Moving MNIST [28]** The Moving MNIST dataset contains 10,000 video sequences in which two digits move independently around the frame. The digits frequently intersect with each other and bounce off the edges of the frame. The values of $C, L, \hat{L}, H, W$ are 1, 10, 10, 64, and 64 respectively.
- **PNW-Typhoon** The Pacific Northwest Windstorm (PNW)-Typhoon dataset is a collection of typhoons observed from 2019 to 2023 using GK2A (GEO-KOMPSAT-2A) satellite. PNW-Typhoon dataset uses observations at 1-hour intervals over the East Asia regions with a 2 km spatial resolution. All data preprocessing strictly adhered to the official GK2A user manual[5]. In our experiments, we use channels of the infrared ray (IR) at 10.5 μm, short wave (SW) at 0.38 μm, and water vapor (WV) at 0.69 μm. Typhoons from 2019 to 2022 are used for training, while the ones from 2023 are used for testing. The values for $C, L, \hat{L}, H, W$ are 3, 10, 10, 128, and 128, respectively.
- **Weatherbench [25]** The WeatherBench dataset is a comprehensive weather data consisting of various climatic factors. We employ the WeatherBench-S framework from [30], where each climatic factor is trained individually. Raw data is re-gridded to a 5.625° resolution, corresponding to a 32×64 grid. We use from 2010 to 2015 for training, 2016 for validation, and 2017 and 2018 for test. The values for $C, L, \hat{L}, H, W$ are 1, 12, 12, 32, and 64, respectively.

Evaluation Metric For evaluating DGDM, we use mean squared error (MSE) to heavily penalize larger errors and mean absolute error (MAE) for a linear error penalty. We utilize peak signal-to-noise ratio (PSNR) to evaluate the quality of signal representation against corrupting noise, and structural similarity index measure (SSIM) [36] to assess perceptual results. Additionally, we employ Frechet video distance (FVD) to measure the similarity between the generated videos and the ground truth videos in feature space.

Implementation Details For our implementation, the Adam optimizer is employed with a learning rate of $1e^{-4}$ for the probabilistic branch and $3e^{-4}$ for the deterministic branch. We adopt a forwarding process of 1,000 steps and a reverse process of 200 steps. And, the reverse process is truncated after 100 steps. The number of training epochs varies by dataset: 2,000 for the Moving MNIST, 200 for the PNW-Typhoon dataset, and 50 for the WeatherBench dataset. All experiments are conducted using the PyTorch framework on a single A100 GPU. Detailed network architectures are provided in the supplementary material.

[5] `https://nmsc.kma.go.kr/enhome/html/base/cmm/selectPage.do?page=satellite.gk2a.intro`

Table 1: Performance comparison results from the Moving MNIST. For diffusion models, the amount of computation is indicated by the number of steps required in the reverse process. **Bold** indicates the best performance, and underline indicates the second-best performance.

Model	Diversity	#Param.	#Flops	Evaluation metric				
				MAE↓	MSE↓	PSNR↑	SSIM↑	FVD↓
ConvLSTM [27]	×	15.0M	56.79G	90.63	29.79	22.14	0.928	79.193
PredRNN [34]	×	23.8M	0.12T	72.82	23.96	23.28	0.946	50.407
PredRNN++ [33]	×	38.6M	0.17T	69.58	22.05	23.65	0.950	45.731
MIM [35]	×	38.0M	179.20G	70.67	22.92	23.53	0.948	47.530
PhyDNet [12]	×	3.1M	15.30G	61.47	20.35	24.21	0.955	38.752
SimVP [9]	×	58.0M	9.43G	89.04	32.14	21.83	0.926	72.969
TAU [29]	×	46.8M	15.95G	51.46	**15.68**	**25.71**	**0.966**	28.169
VDM [14]	✓	35.7M	1000 × 77.45G	123.12	86.33	18.71	0.879	8.800
MCVD [32]	✓	28.0M	100 × 9.92G	172.47	64.68	19.23	0.565	8.161
RaMViD [15]	✓	235.1M	1000 × 1.05T	123.76	81.26	18.87	0.878	12.059
PreDiff [10]	✓	129.4M	1000 × 0.70T	81.16	42.20	19.05	0.931	7.889
DGDM-DB	×	27.0M	11.46G	56.45	17.88	24.94	0.961	19.216
DGDM-PB	✓	63.3M	100 × 77.29G	50.21	20.96	25.08	0.962	7.461
DGDM-Best	×	63.3M	10 × 100 × 77.29G	**47.31**	19.14	25.59	**0.966**	**7.427**
DGDM-Average	✓	63.3M	10 × 100 × 77.29G	48.54	20.52	25.22	**0.966**	9.617

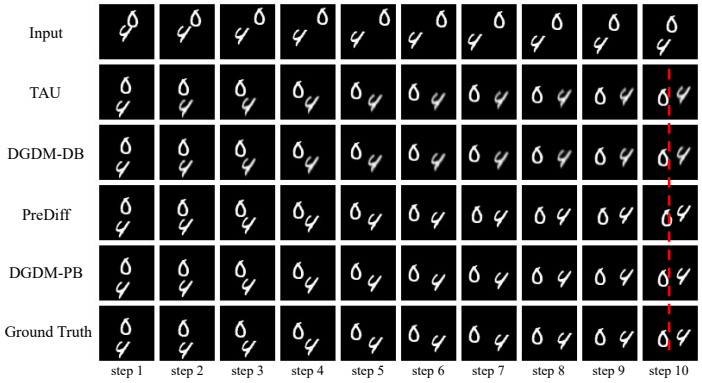

Fig. 4: Visualization of Moving MNIST. The red dash line guides the location of digits.

4.2 Experimental Results

Moving MNIST We compare the performance with existing data-driven deterministic models [9,12,27,29,33–35] and probabilistic models [10,14,15,32]. As shown in Table 1, deterministic models without diversity achieve better scores in MSE, PSNR, and SSIM, but probabilistic models obtain better FVD. This suggests that while deterministic models are adept at predicting future movements in the Moving MNIST, they struggle to produce finer details. On the other hand, despite being a probabilistic model, DGDM demonstrates superior performance in MAE, SSIM, and FVD by incorporating the deterministic branch. DGDM-

Table 2: Performance comparison results from the PNW-Typhoon dataset include IR, WV, and SW channels are infrared red, water vapor, and short wave, respectively.

Model	PNW-Typhoon														
	MSE↓			MAE↓			PSNR↑			SSIM↑			FVD↓		
	IR	SW	WV	IR	SW	WV	IR	SW	WV	IR	SW	WV	IR	SW	WV
ConvLSTM [27]	937.80	1075.43	409.65	2973.50	3221.82	1952.39	13.06	14.31	17.94	0.399	0.400	0.642	1689.39	1321.36	993.59
PredRNN [34]	774.73	804.28	253.28	2342.76	2933.85	1756.32	14.23	14.57	18.34	0.401	0.403	0.629	1640.11	1293.09	1033.27
PredRNN++ [33]	667.18	739.01	213.51	2372.64	3088.18	1552.26	13.51	14.55	20.29	0.414	0.410	0.630	1340.61	1040.10	926.51
MIM [35]	625.99	683.51	195.14	2398.68	3127.97	1567.28	13.85	13.95	17.55	0.392	0.392	0.613	1262.59	969.19	954.53
PhyDNet [12]	655.14	728.93	189.02	2460.73	3144.94	1675.00	14.42	14.10	18.96	0.408	0.398	0.576	1305.54	1007.79	944.65
SimVP [9]	643.06	706.72	204.33	2452.73	3079.17	1609.57	14.01	14.21	19.17	0.401	0.406	0.598	1283.45	991.53	940.42
TAU [29]	565.47	664.50	166.11	2117.84	2842.39	1493.33	15.12	14.82	19.49	0.404	0.418	0.603	997.47	880.41	891.90
VDM [14]	881.08	1128.37	401.52	2794.82	3355.41	2063.66	13.43	12.31	17.07	0.371	0.383	0.607	830.29	727.09	573.47
MCVD [32]	605.51	904.93	472.19	2166.98	2799.92	2300.42	14.89	13.57	16.74	0.430	0.433	0.647	737.99	439.97	481.70
RaMViD [15]	770.72	1152.08	341.68	2781.69	3355.09	1803.71	13.96	12.55	18.39	0.392	0.392	0.624	1091.97	935.57	**395.73**
PreDiff [10]	568.93	683.72	141.51	2050.45	2299.12	1064.25	14.64	14.01	20.78	0.418	0.414	0.632	760.73	487.65	548.54
DGDM-DB	495.52	576.78	133.16	1949.46	2162.01	1031.67	15.89	15.43	21.79	0.442	0.436	0.653	819.72	710.32	799.30
DGDM-PB	461.90	540.79	121.27	1875.08	2065.37	1003.18	16.08	15.57	21.91	0.478	0.457	0.678	705.56	422.03	508.36
DGDM-Best	459.85	533.35	120.75	1871.08	2047.80	1000.82	16.10	15.63	21.93	0.480	0.460	0.679	**698.93**	**414.07**	502.26
DGDM-Average	**456.11**	**516.66**	**120.34**	**1863.26**	**2004.71**	**998.31**	**16.15**	**15.83**	**21.96**	**0.482**	**0.483**	**0.682**	733.03	482.80	531.14

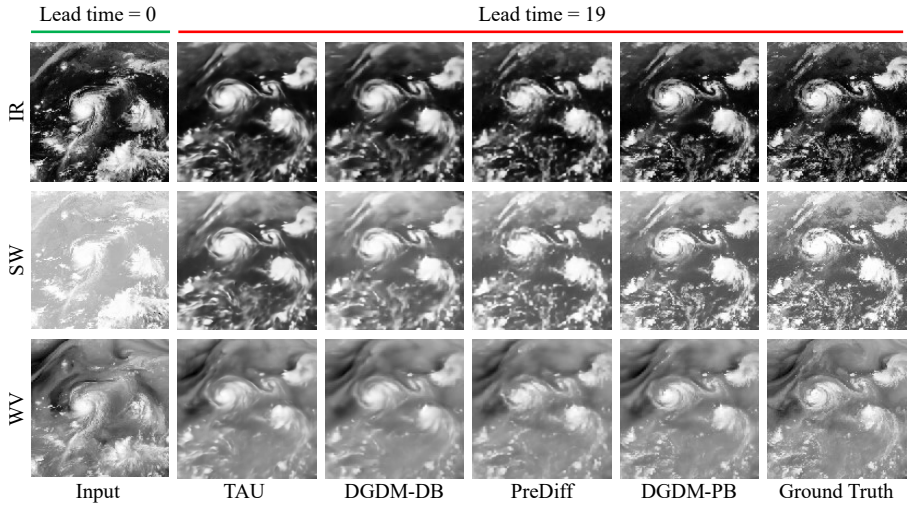

Fig. 5: Qualitative comparison results on the PNW-Typhoon. TAU and PreDiff are the best-performing approaches in deterministic and probabilistic models, respectively.

DB and DGDM-PB represent the results from the deterministic and probabilistic branches of DGDM, respectively. Notably, when selecting the best performance from several samples, DGDM-Best shows impressive results. DGDM-Average, which averages multiple samples to simulate an ensemble method, achieves improved scores in MAE, MSE, PSNR, and SSIM. However, the performance of FVD declined when samples are selected randomly. These experimental results indicate that DGDM is a model that meets our accuracy expectations while producing diverse outputs.

As illustrated in Fig. 4, these experimental analyses provide insights that confirm our observations: deterministic models precisely predict locations but

Table 3: Quantitative comparison results on the WeatherBench dataset.

Model	WeatherBench					
	Temperature (t2m)		Humidity (r)		Wind (uv10)	
	MSE	RMSE	MSE	RMSE	MSE	RMSE
ConvLSTM [27]	1.521	1.233	35.146	5.928	1.898	1.378
PredRNN [34]	1.331	1.154	37.611	6.133	1.881	1.372
PredRNN++ [33]	1.634	1.278	35.146	5.928	1.873	1.369
MIM [35]	1.784	1.336	36.534	6.044	3.140	1.772
SimVP [9]	1.238	1.113	34.355	5.861	1.999	1.414
TAU [29]	<u>1.162</u>	<u>1.078</u>	31.831	5.642	<u>1.593</u>	<u>1.262</u>
VDM [14]	2.343	1.530	43.293	6.579	2.235	1.495
MCVD [32]	2.512	1.584	45.691	6.759	2.221	1.490
RaMViD [15]	1.908	1.381	39.028	6.247	2.764	1.662
DGDM-DB	1.177	1.085	30.624	5.533	1.742	1.320
DGDM-PB	1.155	1.075	29.529	5.434	1.772	1.331
DGDM-Best	**1.025**	**1.012**	**28.572**	**5.345**	**1.591**	**1.262**
DGDM-Average	1.183	1.087	<u>30.326</u>	<u>5.506</u>	1.644	1.282

lack the clarity of the GT, while probabilistic models, despite their clarity, do not predict locations as accurately. Since probabilistic model, PreDiff has a lot of diversity, the generated samples are different in a large gap, making it difficult to use the ensemble method. However, DGDM demonstrates a close resemblance to GT in both location accuracy and clarity.

PNW-Typhoon Since the typhoon is an extreme weather event that has many nonlinearities and changes rapidly, forecasting the typhoon 10-hours later is challenging. Table 2 shows that most of the deterministic models struggle to capture the uncertainty of typhoons, resulting in poor accuracy. On the other hand, the probabilistic models show similar accuracy to the deterministic models, particularly PreDiff [10] outperforms the deterministic models in WV. This suggests that probabilistic models have potential for accounting for uncertainty. Then, DGDM superior other models and the simple averaging ensemble method obtain best performance.

In typhoon analysis, factors like the eye of the typhoon, its size, and the way clouds disperse and coalesce are important. As depicted in Fig. 5, while TAU [29] forecasts the size of the typhoon well, the results are too blurry to definitively assess the cloud patterns. On the other hand, PreDiff provides clearer images but falls short in forecasting the size of the typhoon and cloud formation of left down size of figure. DGDM most accurately simulates the size of a typhoon and is also the most precise in forecasting the formation and dissipation of clouds.

WeatherBench Table 3 shows the results of quantitative experiments in a global weather forecasting dataset WeatherBench. DGDM-Best, a method of selecting the sample with the highest performance among all 20 samples, achieves the best performance in the WeatherBench dataset. In addition, DGDM-Average,

Table 4: Quantitative comparisons on components of DGDM.

Components					Deterministic		Probabilsitic	
DB	PB	BB	LF	SVS	MAE	FVD	MAE	FVD
✓					58.13	18.50	-	-
	✓				-	-	123.20	8.80
✓	✓				95.09	58.83	110.07	14.74
	✓	✓			-	-	89.45	17.05
✓	✓	✓			56.71	20.51	52.04	10.06
✓	✓	✓	✓		56.18	18.72	50.35	9.31
✓	✓	✓	✓	✓	56.45	19.22	**50.22**	**8.28**

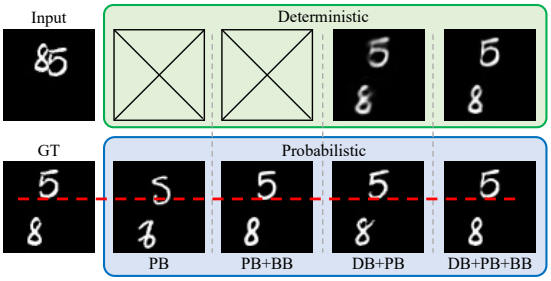

Fig. 6: Visualization the results by components.

which ensembles all models, shows the second-best performance in the Humidity modality. These experimental results indicate that DGDM has high usability even if a single output is not selected and simply ensembled with average.

4.3 Ablation Study

Effect of Each Component As shown in Table 4 and Fig. 6, using only PB produces plausible and clear results, but the location and shape is different from GT. When PB is combined with the BB, the shape of the results is more aligned with GT, and PB guided by DB corrects the location. Moreover, using last frame (LF) for PB is not only remove the output length constraint but also improves the performance. Note that, SVS enhances MAE and FVD performance and also accelerates inference speed for frames with shorter lead times. While the total diffusion step for the video remains constant, SVS generates frames earlier for shorter lead times due to the fewer diffusion steps allocated for them. These experimental results demonstrate that all components of DGDM are complementary to each other. Interestingly, it has been observed that when DB and PB are trained together, the performance of DB also improves.

Effect of Truncated Diffusion with Deterministic Table 5 shows the results of the reverse process based on the number of denoising steps. When using

Table 5: Quantitative results with different denoising steps.

Denoising steps		Evaluation metric					
		MSE		MAE		FVD	
		Mean	STD	Mean	STD	Mean	STD
w/o truncated diffusion	200	23.621	0.105	51.223	0.126	8.282	0.087
	175	23.326	0.066	50.561	0.079	8.600	0.121
	150	23.146	0.066	50.363	0.083	8.221	0.149
	125	23.453	0.145	51.340	0.190	8.487	0.122
	100	23.222	0.058	52.046	0.081	8.132	0.154
w/ truncated diffusion	175	21.947	0.046	49.543	0.080	8.326	0.079
	150	21.613	0.031	**49.498**	0.052	8.126	0.080
	125	21.370	0.028	49.792	0.034	7.913	0.072
	100	**20.963**	0.021	50.210	0.033	**7.461**	0.068

only a non-Markovian chain without truncated diffusion, there are no significant relationship between the number of denoising steps and MAE, MSE, and FVD. On the other hand, setting the reverse step to 200 and using deterministic results to truncate the reverse process to reduce the denoising step significantly improves accuracy. In addition, there is a trend where reducing the denoising steps results in a lower standard deviation (STD) of MAE, MSE and FVD. This suggests that truncated diffusion effectively modulate the range of the possible future. In addition, truncated diffusion improves computational efficiency without degrading performance.

5 Conclusion

In this paper, we present DGDM for high-accuracy probabilistic weather forecasting. DGDM bridges the gap between the accuracy of deterministic models and the diversity of probabilistic models, addressing the limitation of existing data-driven models. By using truncated diffusion, DGDM controls the range of possible future weather events. Furthermore, through the video frame conditioning method, SVS, we improve video quality. We prove the effectiveness of proposed methods on the synthetic dataset. DGDM demonstrates its effectiveness not only in reginal extreme weather, but also in global weather forecasting. We envision DGDM becoming a cornerstone in the domain of weather forecasting.

Limitations and Future work Our paper has not explored the use of multiple modalities. By leveraging work on modalities in diffusion models, we believe that DGDM can utilize multi-modality like other weather prediction methods. Due to the use of deterministic guidance, DGDM generates samples that are both closely aligned with future and diverse. However, a challenge persists for DGDM, as it is fundamentally a probabilistic model. This necessitates the careful selection of a sample from various possibilities to closely match future weather conditions. We will study methods for selecting the best sample from a variety of possibilities, rather than simply average ensembling.

Acknowledgement

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2020-II201373, Artificial Intelligence Graduate School Program(Hanyang University))

References

1. Ayzel, G., Scheffer, T., Heistermann, M.: Rainnet v1. 0: a convolutional neural network for radar-based precipitation nowcasting. Geoscientific Model Development **13**(6), 2631–2644 (2020)
2. Babaeizadeh, M., Finn, C., Erhan, D., Campbell, R.H., Levine, S.: Stochastic variational video prediction. arXiv preprint arXiv:1710.11252 (2017)
3. Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., Tian, Q.: Accurate medium-range global weather forecasting with 3d neural networks. Nature **619**(7970), 533–538 (2023)
4. Brown, A., Milton, S., Cullen, M., Golding, B., Mitchell, J., Shelly, A.: Unified modeling and prediction of weather and climate: A 25-year journey. Bulletin of the American Meteorological Society **93**(12), 1865–1877 (2012)
5. Chatterjee, M., Ahuja, N., Cherian, A.: A hierarchical variational neural uncertainty model for stochastic video prediction. In: ICCV. pp. 9751–9761 (2021)
6. Dabrowski, J.J., Zhang, Y., Rahman, A.: Forecastnet: a time-variant deep feed-forward neural network architecture for multi-step-ahead time-series forecasting. pp. 579–591. Springer (2020)
7. Dee, D.P., Uppala, S.M., Simmons, A.J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M., Balsamo, G., Bauer, d.P., et al.: The era-interim reanalysis: Configuration and performance of the data assimilation system. Quarterly Journal of the royal meteorological society **137**(656), 553–597 (2011)
8. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: A benchmark. In: CVPR. pp. 304–311 (2009)
9. Gao, Z., Tan, C., Wu, L., Li, S.Z.: Simvp: Simpler yet better video prediction. In: CVPR. pp. 3170–3180 (2022)
10. Gao, Z., Shi, X., Han, B., Wang, H., Jin, X., Maddix, D., Zhu, Y., Li, M., Wang, Y.: Prediff: Precipitation nowcasting with latent diffusion models. arXiv preprint arXiv:2307.10422 (2023)
11. Gruca, A., Serva, F., Lliso, L., Rípodas, P., Calbet, X., Herruzo, P., Pihrt, J., Raevskyi, R., Šimánek, P., Choma, M., Li, Y., Dong, H., Belousov, Y., Polezhaev, S., Pulfer, B., Seo, M., Kim, D., Shin, S., Kim, E., Ahn, S., Choi, Y., Park, J., Son, M., Cho, S., Lee, I., Kim, C., Kim, T., Kang, S., Shin, H., Yoon, D., Eom, S., Shin, K., Yun, S.Y., Le Saux, B., Kopp, M.K., Hochreiter, S., Kreil, D.P.: Weather4cast at neurips 2022: Super-resolution rain movie prediction under spatio-temporal shifts. In: Ciccone, M., Stolovitzky, G., Albrecht, J. (eds.) Proceedings of the NeurIPS 2022 Competitions Track. Proceedings of Machine Learning Research, vol. 220, pp. 292–313. PMLR (28 Nov–09 Dec 2022), `https://proceedings.mlr.press/v220/gruca22a.html`
12. Guen, V.L., Thome, N.: Disentangling physical dynamics from unknown factors for unsupervised video prediction. In: CVPR. pp. 11474–11484 (2020)

13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models **33**, 6840–6851 (2020)
14. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. arXiv:2204.03458 (2022)
15. Höppe, T., Mehrjou, A., Bauer, S., Nielsen, D., Dittadi, A.: Diffusion models for video prediction and infilling. Transactions on Machine Learning Research (2022)
16. Hu, X., Huang, Z., Huang, A., Xu, J., Zhou, S.: A dynamic multi-scale voxel flow network for video prediction. In: CVPR. pp. 6121–6131 (2023)
17. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE TPAMI **36**(7), 1325–1339 (2013)
18. Iorio, J., Duffy, P., Govindasamy, B., Thompson, S., Khairoutdinov, M., Randall, D.: Effects of model resolution and subgrid-scale physics on the simulation of precipitation in the continental united states. Climate Dynamics **23**, 243–258 (2004)
19. Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Pritzel, A., Ravuri, S., Ewalds, T., Alet, F., Eaton-Rosen, Z., et al.: Graphcast: Learning skillful medium-range global weather forecasting. arXiv preprint arXiv:2212.12794 (2022)
20. 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)
21. Li, B., Xue, K., Liu, B., Lai, Y.K.: Bbdm: Image-to-image translation with brownian bridge diffusion models. In: CVPR. pp. 1952–1961 (2023)
22. Luc, P., Clark, A., Dieleman, S., Casas, D.d.L., Doron, Y., Cassirer, A., Simonyan, K.: Transformation-based adversarial video prediction on large-scale data. arXiv preprint arXiv:2003.04035 (2020)
23. Ning, S., Lan, M., Li, Y., Chen, C., Chen, Q., Chen, X., Han, X., Cui, S.: Mimo is all you need: a strong multi-in-multi-out baseline for video prediction. In: AAAI. vol. 37, pp. 1975–1983 (2023)
24. Rakhimov, R., Volkhonskiy, D., Artemov, A., Zorin, D., Burnaev, E.: Latent video transformer. arXiv preprint arXiv:2006.10704 (2020)
25. Rasp, S., Dueben, P.D., Scher, S., Weyn, J.A., Mouatadid, S., Thuerey, N.: Weatherbench: a benchmark data set for data-driven weather forecasting. Journal of Advances in Modeling Earth Systems **12**(11), e2020MS002203 (2020)
26. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
27. Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.c.: Convolutional lstm network: A machine learning approach for precipitation nowcasting **28** (2015)
28. Srivastava, N., Mansimov, E., Salakhudinov, R.: Unsupervised learning of video representations using lstms. In: International conference on machine learning. pp. 843–852. PMLR (2015)
29. Tan, C., Gao, Z., Wu, L., Xu, Y., Xia, J., Li, S., Li, S.Z.: Temporal attention unit: Towards efficient spatiotemporal predictive learning. In: CVPR. pp. 18770–18782 (2023)
30. Tan, C., Li, S., Gao, Z., Guan, W., Wang, Z., Liu, Z., Wu, L., Li, S.Z.: Openstl: A comprehensive benchmark of spatio-temporal predictive learning. In: Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2023)
31. Trebing, K., Staczyk, T., Mehrkanoon, S.: Smaat-unet: Precipitation nowcasting using a small attention-unet architecture. Pattern Recognition Letters **145**, 178–186 (2021)

32. Voleti, V., Jolicoeur-Martineau, A., Pal, C.: Mcvd-masked conditional video diffusion for prediction, generation, and interpolation **35**, 23371–23385 (2022)
33. Wang, Y., Gao, Z., Long, M., Wang, J., Philip, S.Y.: Predrnn++: Towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning. In: International Conference on Machine Learning. pp. 5123–5132. PMLR (2018)
34. Wang, Y., Long, M., Wang, J., Gao, Z., Yu, P.S.: Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms **30** (2017)
35. Wang, Y., Zhang, J., Zhu, H., Long, M., Wang, J., Yu, P.S.: Memory in memory: A predictive neural network for learning higher-order non-stationarity from spatiotemporal dynamics. In: CVPR. pp. 9154–9162 (2019)
36. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE TIP **13**(4), 600–612 (2004)
37. Zhang, J., Zheng, Y., Qi, D.: Deep spatio-temporal residual networks for citywide crowd flows prediction. In: AAAI. vol. 31 (2017)
38. Zhong, Y., Liang, L., Zharkov, I., Neumann, U.: Mmvp: Motion-matrix-based video prediction. In: ICCV. pp. 4273–4283 (2023)