ICML International Conference On Machine Learning

Published as a **spotlight** paper at ICML 2025

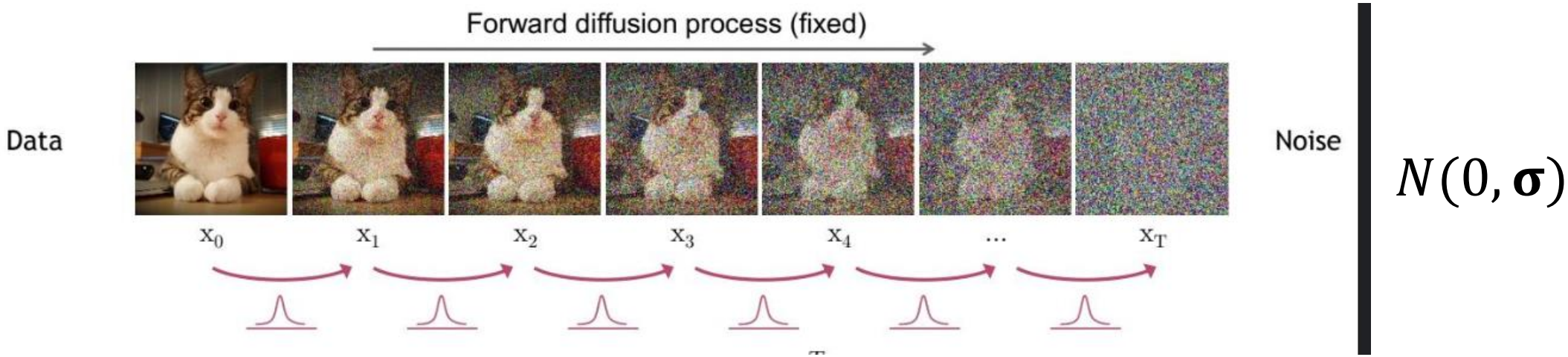
中南大學 CENTRAL SOUTH UNIVERSITY

Non-stationary Diffusion For Probabilistic Time Series Forecasting

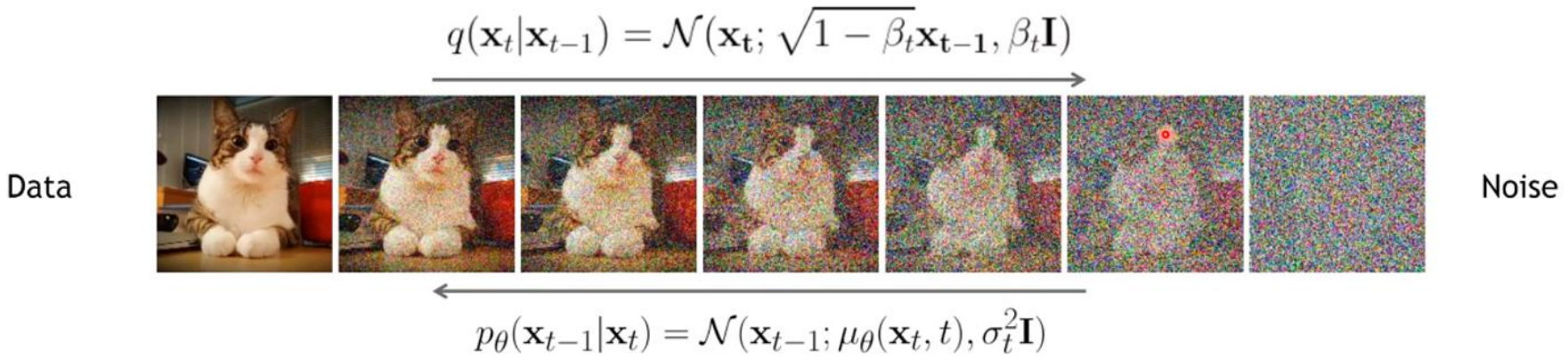
(Presenter)

Weiwei Ye*, Zhuopeng Xu*, Ning Gui

The origin DDPM model

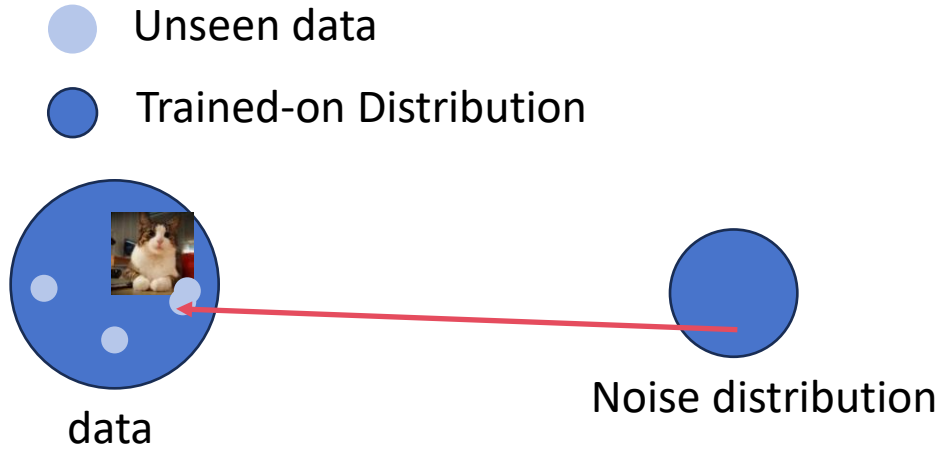


$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ **Parameterize the reverse denoising distribution**



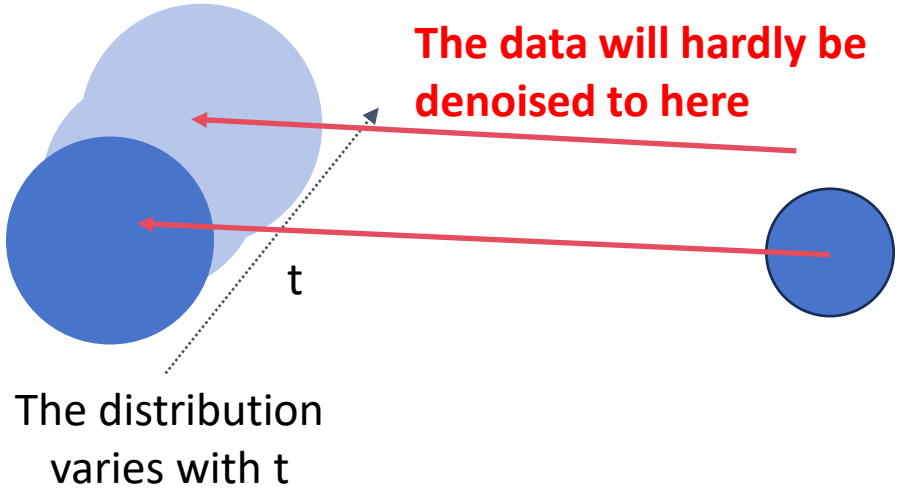
The origin DDPM model can not handle Non-stationarity

Computer Vision



To create an effective generation, all we need to do is interpolate between noise

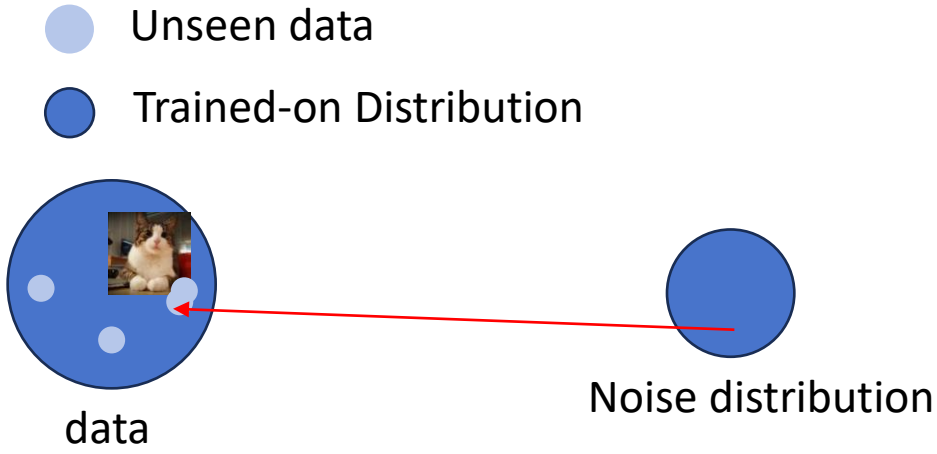
Time Series Forecasting



To accurately predict, a extrapolate is required, we need a time-varying noise distribution.

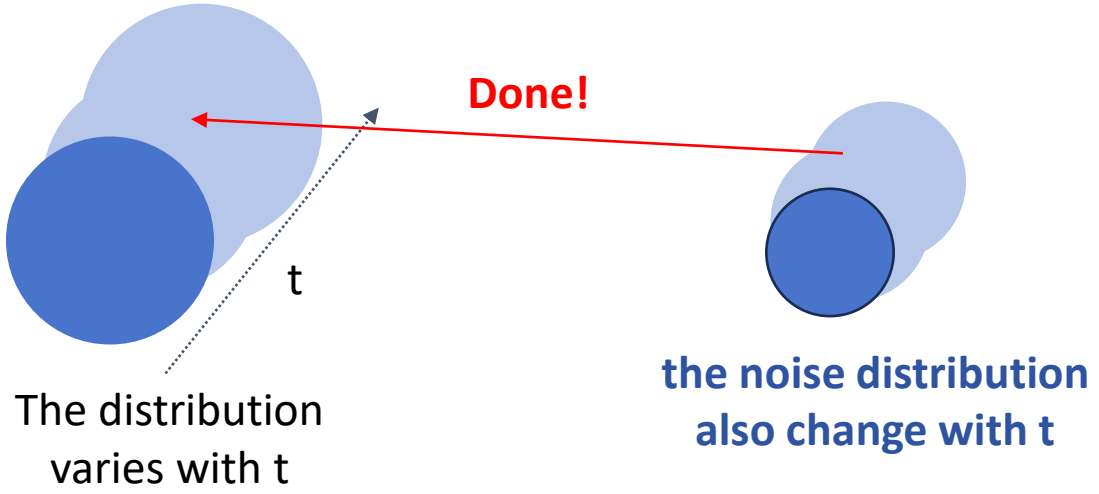
The origin DDPM model can not handle Non-stationarity

Additive Noise Model



$$\mathbf{Y}_t = f(\mathbf{X}_t) + \boldsymbol{\epsilon},$$
$$\text{where } \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma})$$

Location-Scale Noise Model

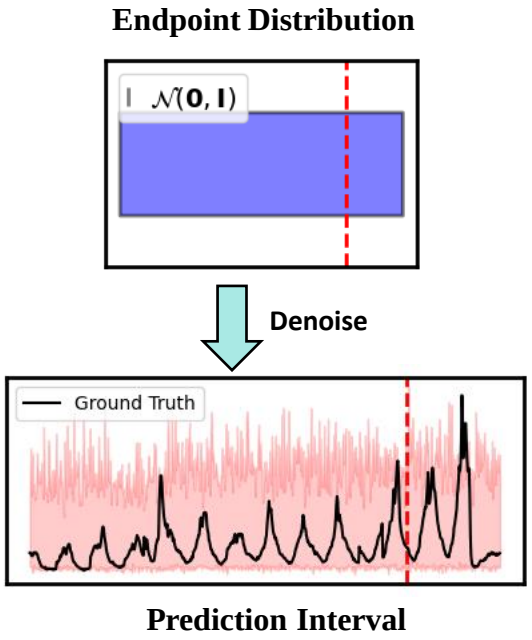


$$\mathbf{Y}_t = f(\mathbf{X}_t) + \sqrt{\mathrm{g}(\mathbf{X}_t)}\boldsymbol{\epsilon},$$

(Contribution 1)

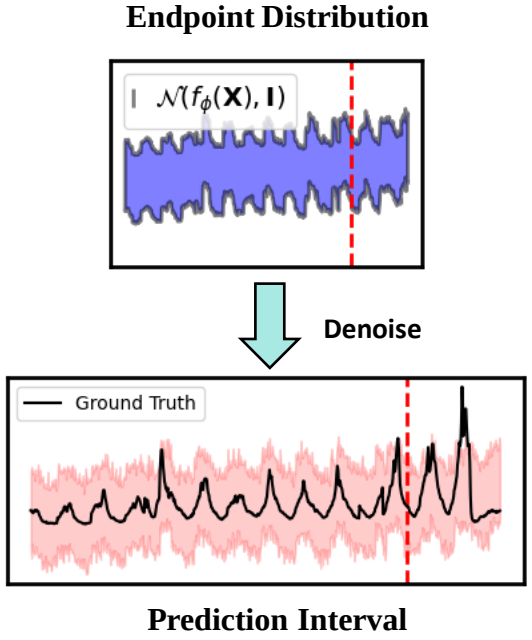
The origin DDPM model can not handle Non-stationarity

Additive Noise Model



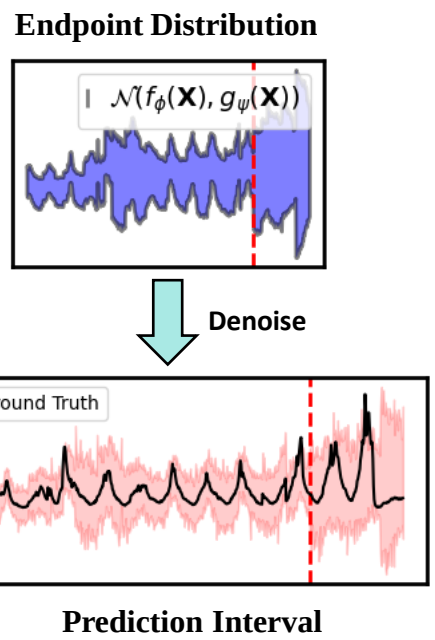
The distribution do not change

Location-Scale Noise Model (Mean only)



The mean varies with t

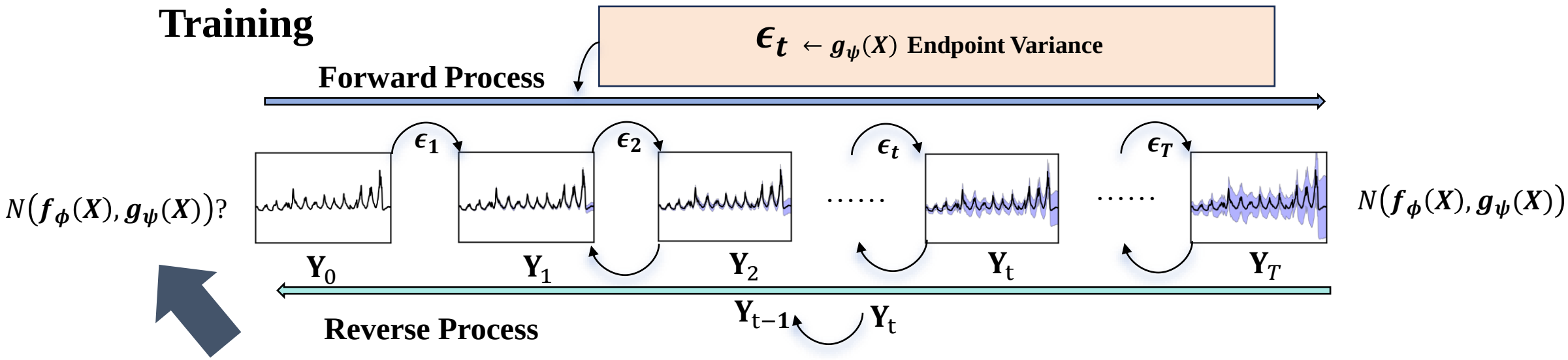
Location-Scale Noise Model



The whole distribution varies with t

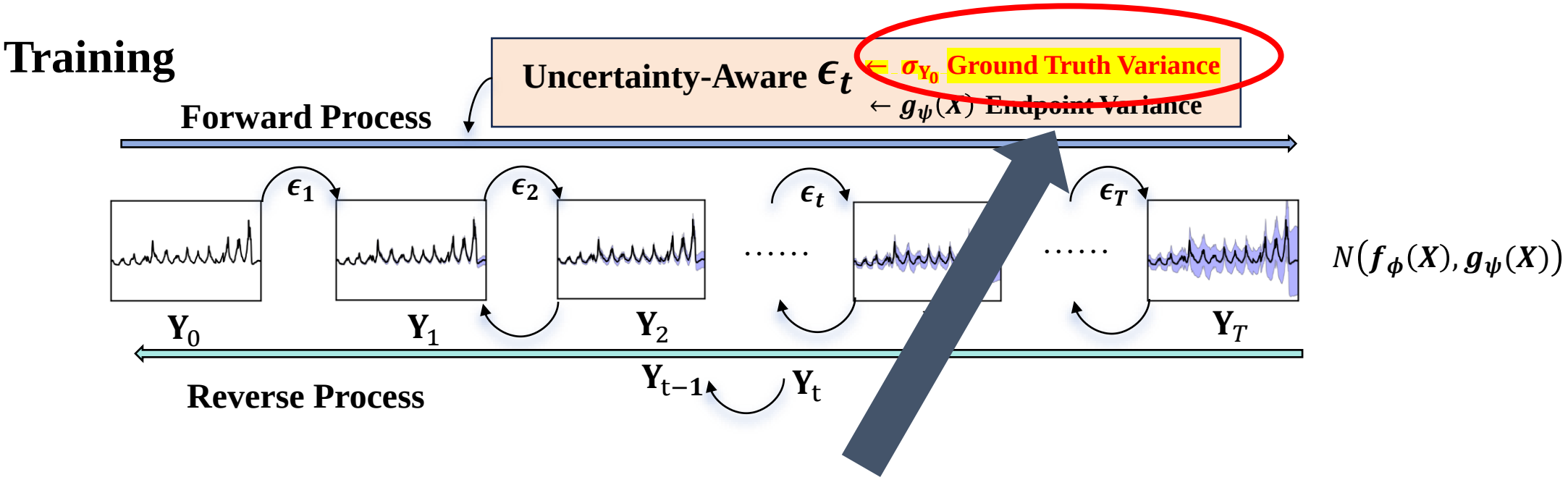
Obviously, this is the best!!!

The stiffness problem



the denoising process and de-noise ϵ_t relies only on g(x), after denoise the distribution will resemble highly with $N(f_\phi(X), g_\psi(X))$

The stiffness problem



To alleviate stiffness, we put the ground truth variance into the denoising process. (Contribution 2)

Design The Forward Distribution

So to handle the time-varying uncertainty, we use LSNM instead of ANM and redefine the endpoint of the original DDPM, so the uncertainty level can change according the data and time：

$$p(\mathbf{Y}_T \mid f_\phi(\mathbf{X}), g_\psi(\mathbf{X})) := \mathcal{N}(f_\phi(\mathbf{X}), g_\psi(\mathbf{X}))$$

$f_\phi(\mathbf{X})$ and $g_\psi(\mathbf{X})$ estimate the expectation $\mathbb{E}[\mathbf{Y} \mid \mathbf{X}]$ and $\mathrm{Var}[\mathbf{Y} \mid \mathbf{X}]$ ， and can be parameterized by any point prediction model 。

Additionally， to alleviate stiffness, we put the ground truth variance into the denoising process, the forward distribution can be formulated as：

$$\mathcal{N}(\mathbf{Y}_t; \sqrt{\alpha_t}\mathbf{Y}_{t-1} + (1-\sqrt{\alpha_t})f_\phi(\mathbf{X}), (\beta_t^2 g_\psi(\mathbf{X}) + \alpha_t\beta_t\boldsymbol{\sigma}_{\mathbf{Y}_0}))$$

The stiffness problem

LSNM endpoint

$$p\left(\mathbf{Y}_T \mid f_\phi(\mathbf{X}), g_\psi(\mathbf{X})\right) := \mathcal{N}\left(f_\phi(\mathbf{X}), g_\psi(\mathbf{X})\right)$$

Forward Distribution

$$q\left(\mathbf{Y}_t \mid \mathbf{Y}_{t-1}, f_\phi(\mathbf{X}), g_\psi(\mathbf{X}), \boldsymbol{\sigma}_{\mathbf{Y}_0}\right)$$
$$= \mathcal{N}\left(\mathbf{Y}_t;\ \sqrt{\alpha_t}\mathbf{Y}_{t-1} + (1-\sqrt{\alpha_t}) f_\phi(\mathbf{X}), (\beta_t^2 g_\psi(\mathbf{X}) + \alpha_t \beta_t \boldsymbol{\sigma}_{\mathbf{Y}_0})\right)$$

Closed Form

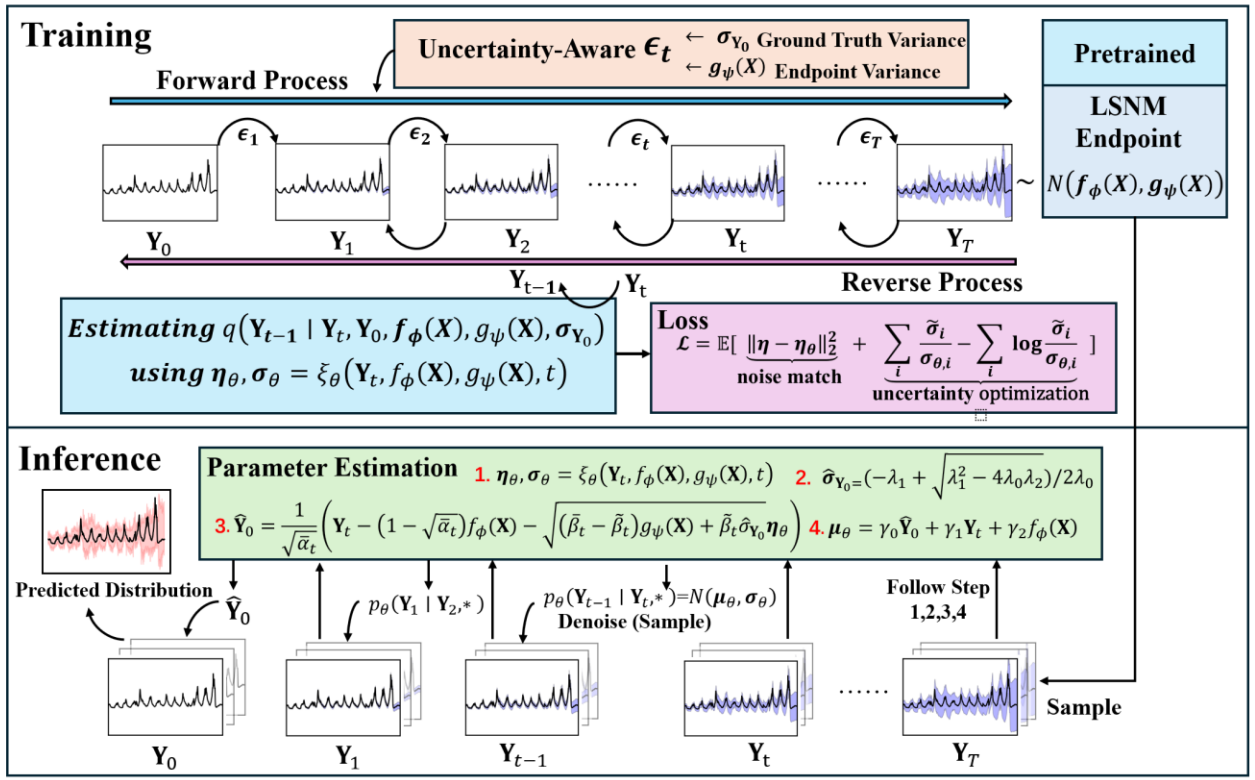
$$q\left(\mathbf{Y}_t \mid \mathbf{Y}_0, f_\phi(\mathbf{X}), g_\psi(\mathbf{X}), \boldsymbol{\sigma}_{\mathbf{Y}_0}\right)$$
$$= \mathcal{N}\left(\mathbf{Y}_t; \sqrt{\bar{\alpha}_t}\mathbf{Y}_0 + \left(1-\sqrt{\bar{\alpha}_t}\right) f_\phi(\mathbf{X}), \left(\bar{\beta}_t - \tilde{\beta}_t\right) g_\psi(\mathbf{X}) + \tilde{\beta}_t \boldsymbol{\sigma}_{\mathbf{Y}_0}\right)$$

Reverse Distribution

$$q\left(\mathbf{Y}_{t-1} \mid \mathbf{Y}_t, \mathbf{Y}_0, f_\phi(\mathbf{X}) g_\psi(\mathbf{X}) \boldsymbol{\sigma}_{\mathbf{Y}_0}\right) := \mathcal{N}\left(\mathbf{Y}_{t-1}; \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\sigma}}\right)$$

$$\tilde{\boldsymbol{\mu}} := \gamma_0 \mathbf{Y}_0 + \gamma_1 \mathbf{Y}_t + \gamma_2 f_\phi(\mathbf{X}) \quad \tilde{\boldsymbol{\sigma}} := \frac{\boldsymbol{\sigma}_t \overline{\boldsymbol{\sigma}}_{t-1}}{\alpha_t \overline{\boldsymbol{\sigma}}_{t-1} + \boldsymbol{\sigma}_t}$$

$$\gamma_0 := \frac{\sqrt{\bar{\alpha}_{t-1}} \boldsymbol{\sigma}_t}{\alpha_t \overline{\boldsymbol{\sigma}}_{t-1} + \boldsymbol{\sigma}_t} \quad \gamma_1 := \frac{\sqrt{\alpha_t}\, \overline{\boldsymbol{\sigma}}_{t-1}}{\alpha_t \overline{\boldsymbol{\sigma}}_{t-1} + \boldsymbol{\sigma}_t} \quad \gamma_2 := \frac{\sqrt{\alpha_t}(\alpha_t - 1)\overline{\boldsymbol{\sigma}}_{t-1} + \left(1-\sqrt{\bar{\alpha}_t}\right)\boldsymbol{\sigma}_t}{\alpha_t \overline{\boldsymbol{\sigma}}_{t-1} + \boldsymbol{\sigma}_t}$$



Note: Reverse distribution is derived using Bayes rule.

The closed form

The forward distribution $q\left(\mathbf{Y}_t \mid \mathbf{Y}_0, f_\phi(\mathbf{X}), g_\psi(\mathbf{X}), \sigma_{\mathbf{Y}_0}\right)$ has closed form:

$$\mathcal{N}(\mathbf{Y}_t; \sqrt{\bar{\alpha}_t}\mathbf{Y}_0 + (1-\sqrt{\bar{\alpha}_t})f_\phi(\mathbf{X}), (\bar{\beta}_t - \tilde{\beta}_t)g_\psi(\mathbf{X}) + \tilde{\beta}_t\boldsymbol{\sigma}_{\mathbf{Y}_0})$$

The parameters are: (detail proof in Appendix A)

$$\tilde{\alpha}_t := \sum_{k=0}^{t-1} \prod_{i=t-k}^{t} \alpha_i \qquad \bar{\beta}_t := 1 - \bar{\alpha}_t$$

$$\hat{\alpha}_t := \sum_{k=0}^{t-1} \left(\prod_{i=t-k}^{t} \alpha_i \right) \alpha_{t-k} \qquad \tilde{\beta}_t := \tilde{\alpha}_t - \hat{\alpha}_t$$

The Reverse Distribution

In the reverse distribution, the posterior of $\mathbf{Y}_{t-1}$ given $\mathbf{Y}_t$ can be derived under certain condition:

$$q\left(\mathbf{Y}_{t-1} \mid \mathbf{Y}_t, \mathbf{Y}_0, f_\phi(\mathbf{X}) g_\psi(\mathbf{X}) \boldsymbol{\sigma}_{\mathbf{Y}_0}\right) := \mathcal{N}\left(\mathbf{Y}_{t-1}; \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\sigma}}\right)$$

$$\tilde{\boldsymbol{\mu}} := \gamma_0 \mathbf{Y}_0 + \gamma_1 \mathbf{Y}_t + \gamma_2 f_\phi(\mathbf{X}) \qquad \tilde{\boldsymbol{\sigma}} := \frac{\boldsymbol{\sigma}_t \boldsymbol{\sigma}_{t-1}}{\alpha_t \boldsymbol{\sigma}_{t-1} + \boldsymbol{\sigma}_t}$$

$$\gamma_0 := \frac{\sqrt{\bar{\alpha}_{t-1}} \boldsymbol{\sigma}_t}{\alpha_t \boldsymbol{\sigma}_{t-1} + \boldsymbol{\sigma}_t} \qquad \gamma_1 := \frac{\sqrt{\alpha_t} \boldsymbol{\sigma}_{t-1}}{\alpha_t \boldsymbol{\sigma}_{t-1} + \boldsymbol{\sigma}_t} \qquad \gamma_2 := \frac{\sqrt{\alpha_t}(\alpha_t - 1)\boldsymbol{\sigma}_{t-1} + \left(1 - \sqrt{\bar{\alpha}_{t-1}}\right)\boldsymbol{\sigma}_t}{\alpha_t \boldsymbol{\sigma}_{t-1} + \boldsymbol{\sigma}_t}$$

Our target is to match the reverse distribution by parameterizing a model $p_\theta(\mathbf{Y}_{t-1} \mid \mathbf{Y}_t, f_\phi(\mathbf{X}), g_\psi(\mathbf{X}))$

Loss function

To parameterize $p_\theta(\mathbf{Y}_{t-1} \mid \mathbf{Y}_t, f_\phi(\mathbf{X}), g_\psi(\mathbf{X}))$, we optimize the KL divergence between $q(\mathbf{Y}_{t-1} \mid \mathbf{Y}_t, f_\phi(\mathbf{X}), g_\psi(\mathbf{X}), \boldsymbol{\sigma}_{\mathbf{Y}_0})$ and p_θ :

$$\mathcal{L} = \mathbb{E}\left[D_{\mathrm{KL}}\left(\mathcal{N}(\boldsymbol{x}; \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\sigma}}) \| \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\sigma}_{\boldsymbol{\theta}})\right)\right]$$

The final form:

$$\mathcal{L} \propto \mathbb{E}\left[\|\boldsymbol{\eta} - \boldsymbol{\eta}_\theta\|_2^2 + \sum_i \frac{\tilde{\boldsymbol{\sigma}}_i}{\boldsymbol{\sigma}_{\theta,i}} - \sum_i \log\left(\frac{\tilde{\boldsymbol{\sigma}}_i}{\boldsymbol{\sigma}_{\theta,i}}\right)\right]$$

The first term ensures the estimation of the distribution mean, the rest terms ensure the estimation of the posterior variance.

Experiment-results

Table 3. Experiment result on nine real-world datasets, **bold face** indicate best result.

Models	Datasets	ETTh1	ETTh2	ETTm1	ETTm2	ECL	EXG	ILI	Solar	Traffic
TimeGrad (2021)	CRPS	0.606	1.212	0.647	0.775	0.397	0.826	1.140	**0.293**	0.407
	QICE	6.731	9.488	6.693	6.962	7.118	9.464	6.519	7.378	4.581
CSDI (2022)	CRPS	0.492	0.647	0.524	0.817	0.577	0.855	1.244	0.432	1.418
	QICE	3.107	5.331	2.828	8.106	7.506	7.864	7.693	9.957	13.613
TimeDiff (2023)	CRPS	0.465	0.471	0.464	0.316	0.750	0.433	1.153	0.700	0.771
	QICE	14.931	14.813	14.795	13.385	15.466	14.556	14.942	14.914	15.439
DiffusionTS (2024)	CRPS	0.603	1.168	0.574	1.035	0.633	1.251	1.612	0.470	0.668
	QICE	6.423	9.577	5.605	9.959	8.205	10.411	10.090	6.627	5.958
TMDM (2024)	CRPS	0.452	0.383	0.375	0.289	0.461	0.336	0.967	0.350	0.557
	QICE	2.821	4.471	2.567	2.610	10.562	6.393	6.217	9.342	10.676
NsDiff **(ours)**	CRPS	**0.392**	**0.358**	**0.346**	**0.256**	**0.290**	**0.324**	**0.806**	0.300	**0.378**
	QICE	**1.470**	**2.074**	**2.041**	**2.030**	**6.685**	**5.930**	**5.598**	**6.820**	**3.601**

NsDiff achieves SOTA on most datasets

Visualized results

NsDiff aligns more with the ground truth

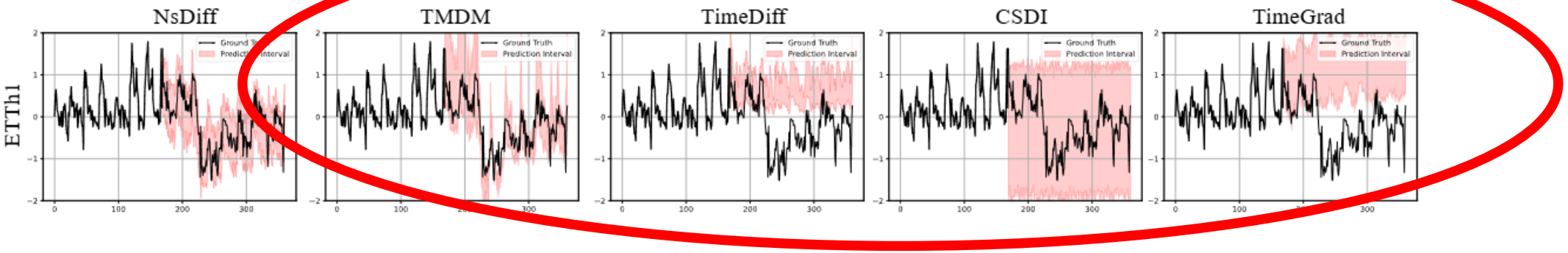


Figure 3. The 95% prediction intervals of a ETTh1 sample, the black line is the true values, the red area represents the prediction interval.

Other models can not handle non-stationarity

Thanks for your listening!!!

Training phase

During training time:

1. We first train two estimator f and g (or we can train together)

2. We train the reverse noise estimator.

Algorithm 1 Training

Input: Data $\mathbf{X}$, target $\mathbf{Y}$, model f_ϕ, noise and variance estimation model ξ_θ, total timesteps T

Pre-train $f_\phi(\mathbf{X})$ to predict $\mathbb{E}(\mathbf{Y}|\mathbf{X})$

Pre-train $g_\psi(\mathbf{X})$ to predict $\mathrm{Var}(\mathbf{Y}|\mathbf{X})$

repeat

Draw $\mathbf{Y}_0 \sim q(\mathbf{Y}_0 \mid \mathbf{X})$

Draw $t \sim \mathrm{Uniform}(\{1, \ldots, T\})$

Draw $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

Compute $\mathbf{Y}_t$:

$$\mathbf{Y}_t = \sqrt{\bar{\alpha}_t}\mathbf{Y}_0 + (1 - \sqrt{\bar{\alpha}_t})f_\phi(\mathbf{X}) + \sqrt{(\bar{\beta}_t - \tilde{\beta}_t)g_\psi(\mathbf{X}) + \tilde{\beta}_t \boldsymbol{\sigma}_{\mathbf{Y}_0}}\boldsymbol{\eta}$$ ▷ using Eq. 7

Compute estimated noise and variance:

$\boldsymbol{\eta}_\theta, \boldsymbol{\sigma}_\theta = \xi_\theta(\mathbf{Y}_t, f_\phi(\mathbf{X}), g_\psi(\mathbf{X}), t)$

Compute loss $\mathcal{L}$ ▷ using Eq. 13

Numerical optimization step on $\nabla_\theta \mathcal{L}$

until Convergence

Inference phase

During inference phase, to estimate $\tilde{\boldsymbol{\sigma}}$, we utilize the quadratic expansion to estimate the result:

$$\lambda_0 \boldsymbol{\sigma}_{\mathbf{Y}_0}^2 + \lambda_1 \boldsymbol{\sigma}_{\mathbf{Y}_0} + \lambda_2 = 0$$

$$\hat{\boldsymbol{\sigma}}_{\mathbf{Y}_0} = \frac{-\lambda_1 + \sqrt{\lambda_1^2 - 4\lambda_0\lambda_2}}{2\lambda_0}$$

which is constrained by the following:

$$g_\psi(\mathbf{X}) < \boldsymbol{\sigma}_\theta \left(\frac{\alpha_t}{\beta_t^2} + \frac{1}{\bar{\beta}_{t-1} - \tilde{\beta}_{t-1}} \right)$$

In the DDPM settings, ($\beta_t \in (0,1)$, this term is vary large!

Algorithm 2 Inference

Input: data $\mathbf{X}$, models f_ϕ, g_ψ, and ξ_θ
Initialize $\mathbf{Y}_T \sim \mathcal{N}(f_\phi(\mathbf{X}), g_\psi(\mathbf{X}))$
for $t = T$ **to** 1 **do**
  if $t > 1$ **then**
    Draw $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  end if
  Compute $\boldsymbol{\eta}_\theta, \boldsymbol{\sigma}_\theta = \xi_\theta(\mathbf{Y}_t, f_\phi(\mathbf{X}), g_\psi(\mathbf{X}), t)$
  Compute $\hat{\boldsymbol{\sigma}}_{\mathbf{Y}_0} = \frac{-\lambda_1 + \sqrt{\lambda_1^2 - 4\lambda_0\lambda_2}}{2\lambda_0}$ ▷ using Eq. 18
  Compute $\hat{\mathbf{Y}}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}} \left(\mathbf{Y}_t - (1 - \sqrt{\bar{\alpha}_t}) f_\phi(\mathbf{X}) - \sqrt{(\bar{\beta}_t - \tilde{\beta}_t) g_\psi(\mathbf{X}) + \tilde{\beta}_t \hat{\boldsymbol{\sigma}}_{\mathbf{Y}_0}} \boldsymbol{\eta}_\theta \right)$ ▷ using Eq. 7
  if $t > 1$ **then**
    Set $\mathbf{Y}_{t-1} = \gamma_0 \hat{\mathbf{Y}}_0 + \gamma_1 \mathbf{Y}_t + \gamma_2 f_\phi(\mathbf{X}) + \sqrt{\boldsymbol{\sigma}_\theta} \boldsymbol{z}$
  else
    Set $\mathbf{Y}_{t-1} = \hat{\mathbf{Y}}_0$
  end if
end for
Output: $\mathbf{Y}_0$

Experiment-Datasets

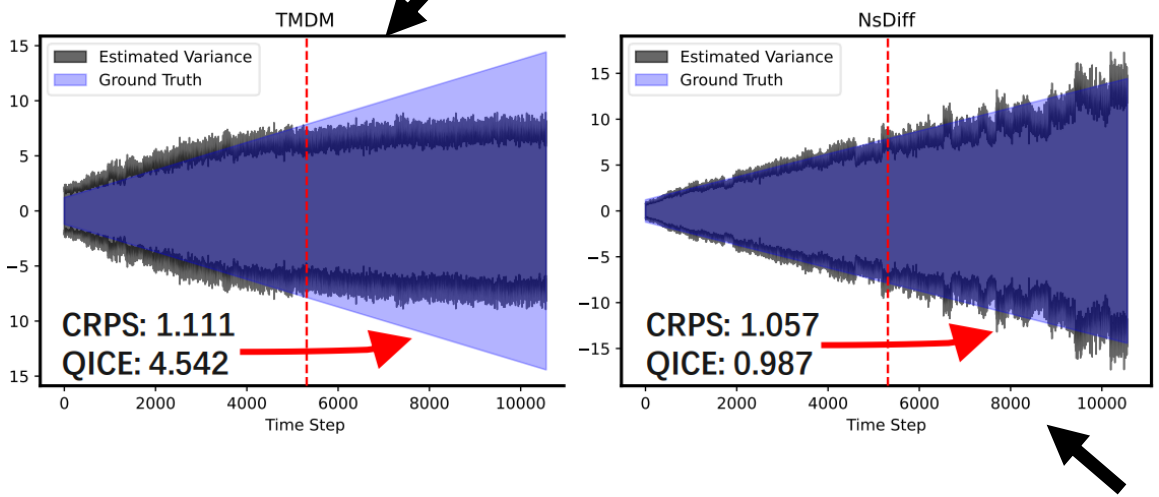
Dataset	Dimension	Timesteps	Pred.steps	Uncert.Var.
ETTm1	7	69680	192	2.53
ETTm2	7	69680	192	1.27
ETTh1	7	17420	192	2.50
ETTh2	7	17420	192	1.29
Exchange	8	7588	192	0.85
ILI	7	966	36	8.26
Electricity	321	26304	192	3.94
Traffic	862	17544	192	181.83
SolarEnergy	137	52560	192	0.92

The uncertainty varies across dataset, from 0.92(SolarEnergy) to 181.83(Traffic)

Synthetic dataset results

Variance	Linear		Quadratic	
Models	CRPS	QICE	CRPS	QICE
TimeGrad	1.129	3.669	2.204	10.740
CSDI	1.100	3.332	1.866	5.050
TimeDiff	1.274	10.314	2.495	14.670
DiffusionTS	1.454	9.290	2.123	11.273
TMDM	1.111	4.542	2.217	11.404
NsDiff	**1.057**	**0.987**	**1.777**	**1.336**

TMDM fail to estimate the linear-growing variance



NsDiff accurately estimates the variance

Two simple variants of NsDiff

Table 1. NsDiff Variants.

Variants	Endpoint	Forward Noise
w/o LSNM	$\mathcal{N} f_\phi(\mathbf{X}), \mathbf{I})$	$\beta_t \mathbf{I}$
w/o UANS	$\mathcal{N}(f_\phi(\mathbf{X}), g_\psi(\mathbf{X}))$	$\beta_t g_\psi(\mathbf{X})$
NsDiff	$\mathcal{N}(f_\phi(\mathbf{X}), g_\psi(\mathbf{X}))$	$\beta_t^2 g_\psi(\mathbf{X}) + \beta_t \alpha_t \boldsymbol{\sigma}_{\mathbf{Y}_0}$

（1）Without the uncertainty-aware noise schedule（w/o UANS）

$$\mathcal{N}(\mathbf{Y}_t; \sqrt{\alpha_t}\mathbf{Y}_{t-1} + (1-\sqrt{\alpha_t}) f_\phi(\mathbf{X}), (1-\alpha_t) g_\psi(\mathbf{X}))$$

（2）Without LSNM（w/o LSNM）

$$\mathcal{N}(\mathbf{Y}_t; \sqrt{\alpha_t}\mathbf{Y}_{t-1} + (1-\sqrt{\alpha_t}) f_\phi(\mathbf{X}), (1-\alpha_t)\mathbf{I})$$

Variants	Endpoint	QICE	CRPS
w/o LSNM	$\mathcal{N}(f_\phi(\mathbf{X}), \mathbf{I})$	2.821 ± 0.718	0.452 ± 0.027
w/o UANS	$\mathcal{N}(f_\phi(\mathbf{X}), g_\psi(\mathbf{X}))$	2.684 ± 0.787	0.413 ± 0.015
NsDiff	$\mathcal{N}(f_\phi(\mathbf{X}), g_\psi(\mathbf{X}))$	$\mathbf{1.470 \pm 0.207}$	$\mathbf{0.392 \pm 0.009}$

The complete NsDiff performs the best on both the performance and robustness.

Two simple variants of NsDiff

Table 1. NsDiff Variants.

Variants	Endpoint	Forward Noise
w/o LSNM	$\mathcal{N} f_\phi(\mathbf{X}), \mathbf{I})$	$\beta_t \mathbf{I}$
w/o UANS	$\mathcal{N}(f_\phi(\mathbf{X}), g_\psi(\mathbf{X}))$	$\beta_t g_\psi(\mathbf{X})$
NsDiff	$\mathcal{N}(f_\phi(\mathbf{X}), g_\psi(\mathbf{X}))$	$\beta_t^2 g_\psi(\mathbf{X}) + \beta_t \alpha_t \boldsymbol{\sigma}_{\mathbf{Y}_0}$

（1）Without the uncertainty-aware noise schedule（w/o UANS）

$$\mathcal{N}(\mathbf{Y}_t; \sqrt{\alpha_t}\mathbf{Y}_{t-1} + (1-\sqrt{\alpha_t}) f_\phi(\mathbf{X}), (1-\alpha_t) g_\psi(\mathbf{X}))$$

（2）Without LSNM（w/o LSNM）

$$\mathcal{N}(\mathbf{Y}_t; \sqrt{\alpha_t}\mathbf{Y}_{t-1} + (1-\sqrt{\alpha_t}) f_\phi(\mathbf{X}), (1-\alpha_t)\mathbf{I})$$

Variants	Endpoint	QICE	CRPS
w/o LSNM	$\mathcal{N}(f_\phi(\mathbf{X}), \mathbf{I})$	2.821 ± 0.718	0.452 ± 0.027
w/o UANS	$\mathcal{N}(f_\phi(\mathbf{X}), g_\psi(\mathbf{X}))$	2.684 ± 0.787	0.413 ± 0.015
NsDiff	$\mathcal{N}(f_\phi(\mathbf{X}), g_\psi(\mathbf{X}))$	$\mathbf{1.470 \pm 0.207}$	$\mathbf{0.392 \pm 0.009}$

The complete NsDiff performs the best on both the performance and robustness.