Resolving Endpoint Underfitting in Diffusion Bridges via Noise Alignment

Yurong Gao[1]; Zicheng Zhang[2]; Congying Han[3]; Tiande Guo[3]; Xinmin Qiu[3]
[1]School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Sciences
[2]JD.com
[3]School of Mathematical Sciences, University of Chinese Academy of Sciences

CVPR JUNE 3-7, 2026 DENVER COLORADO

Our Research Subject

Diffusion Bridge Models (e.g., I2SB and DDBM)

1. The diffusion bridge model is a type of generative model. Unlike flow matching and diffusion processes, it takes two data distributions as endpoints and can directly achieve the transformation between these data distributions. Its structure is naturally suitable for image restoration tasks.

$$X_t = \underbrace{\frac{\overline{\sigma_t^2}}{\sigma_t^2+\sigma_t^2}X_0 + \frac{\sigma_t^2}{\sigma_t^2+\sigma_t^2}X_1}_{\text{Mean term}} + \underbrace{\sqrt{\frac{\sigma_t^2\,\overline{\sigma_t^2}}{\sigma_t^2+\sigma_t^2}}Z}_{\text{Stochastic term}}.$$

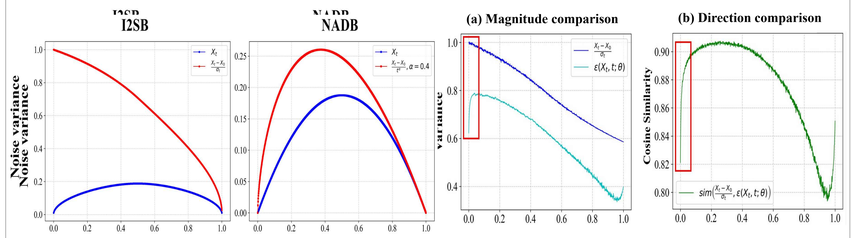
2. Mainstream diffusion bridge models are typically trained by mimicking the denoising paradigm of standard diffusion process:

$$\mathcal{L} = \mathbb{E}_{t,X_0,X_1,Z}\left[\left\|\epsilon(X_t,t;\theta) - \frac{X_t - X_0}{\sigma_t}\right\|^2\right].$$

Motivation & Problem

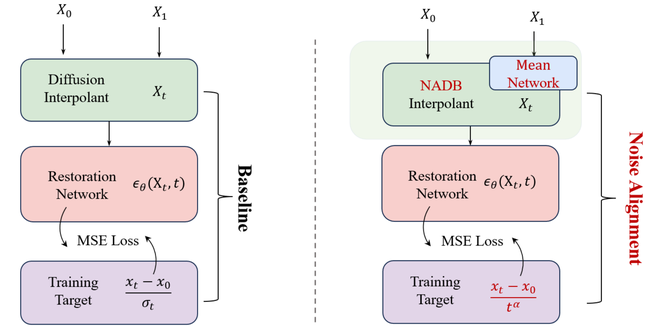
Noise Level Mismatch

1. To quantify the network's failure on this ill-conditioned task, we assess the difference between the network's prediction and the ground truth target along two dimensions:
   - **Magnitude:** We compare the Variance between the two. This measures whether the network predicts the correct amplitude of stochasticity.
   - **Direction:** We compute the Cosine Similarity between the two. This measures whether the direction of the prediction is pointing in the correct direction.
2. A significant discrepancy exists at the target endpoint $(t \to 0)$: For the network input (X_t), its noise coefficient approaches 0 as $t \to 0$, making the input become nearly deterministic X_0. For the training target, its noise coefficient approaches 1 as $t \to 0$. The target becomes highly stochastic Z. This mismatch imposes an ill-conditioned task on the network: it predicts random noise from a clean and deterministic input.



Method

Noise-Aligned Diffusion Bridge (NADB)



To address the issue of noise level mismatch, we utilized the mean network and reconfigured the transmission path and training objective.

Definition 3 (Magnitude-Aligned Stochastic Interpolant). *Let $\alpha \in (0,1)$ and k be a finite constant. We define the magnitude-aligned stochastic interpolant X_t connecting the target $X_0 \sim \rho_0$ and the degraded $X_1 \sim \rho_1$ as:*

$$X_t := (1-t^\alpha)X_0 + t^\alpha X_1 + kt(1-t)Z. \quad (5)$$

We define the corresponding magnitude-aligned training objective Y_t as the scaled displacement, which decomposes as:

$$Y_t := \frac{X_t - X_0}{t^\alpha} = (X_1 - X_0) + kt^{1-\alpha}(1-t)Z. \quad (6)$$

The training objective is to predict this target:

$$\mathcal{L}_{base} = \mathbb{E}_{t,X_0,X_1,Z}\left[\|\epsilon(X_t,t;\theta) - Y_t\|^2\right]. \quad (7)$$

Definition 4 (Mean Network). *Given the joint distribution of the paired data $(X_0,X_1) \sim (\rho_0,\rho_1)$, we define a mean network $M(\cdot;\phi)$ trained to approximate the posterior mean $\mathbb{E}[X_0|X_1]$. The network's output is defined as:*

$$\hat{X}_0 = M(X_1;\phi). \quad (10)$$

The parameters ϕ are optimized by minimizing the mean squared error (MSE) objective:

$$\mathcal{L}_{MSE}(\phi) = \mathbb{E}_{(X_0,X_1)}\left[\|M(X_1;\phi) - X_0\|^2\right]. \quad (11)$$

Algorithm 1 NADB Training

Input: Paired data distributions (ρ_0,ρ_1), pre-trained mean network M_ϕ

repeat
  $t \sim \mathcal{U}([0,1])$, $X_0 \sim \rho_0$, $X_1 \sim \rho_1$
  $\hat{X}_0 = M_\phi(X_1)$ ▷ Get cleaner endpoint via Eq. (10)
  $X_t \sim q(X_t|X_0,\hat{X}_0)$ ▷ Sample via Eq. (13)
  Take gradient descent step on $\epsilon(X_t,t;\theta)$ using $\mathcal{L}_{\text{NADB}}$ in Eq. (14)
until converges

Algorithm 2 NADB Generation

Input: Degraded input $X_1 \sim \rho_1$, time threshold d, steps N

$\hat{X}_0 = M_\phi(X_1)$ ▷ Compute clean endpoint once
for $n = N$ to 1 **do**
  Predict X_0^ϵ using $\epsilon(X_t,t;\theta)$
  if $t_n < d$ **then**
    Sample from $X_{n-1} \sim p(X_{n-1} \mid X_0^\epsilon, \hat{X}_0, X_n)$
  else
    Sample from $X_{n-1} \sim p(X_{n-1} \mid X_0^\epsilon, X_n)$
  end if
end for
return X_0

Results

Table 1. Comparison with I2SB on numerous image restoration tasks.

Method	JPEG Restoration						4× Super-Resolution						Deblurring					
	QF	NFE	FID↓	PSNR↑	SSIM↑	LPIPS↓	Filter	NFE	FID↓	PSNR↑	SSIM↑	LPIPS↓	Kernel	NFE	FID↓	PSNR↑	SSIM↑	LPIPS↓
I2SB	5	10	8.0	24.50	0.69	0.30	Pool	10	7.3	24.86	0.70	0.27	Uniform	10	10.3	24.19	0.65	0.32
NADB			6.9	24.45	0.69	0.30			5.3	24.75	0.71	0.23			4.8	27.70	0.81	0.18
I2SB		100	4.7	23.60	0.65	0.31		100	4.1	23.57	0.65	0.26		100	5.0	22.47	0.58	0.32
NADB			4.3	23.63	0.65	0.30			3.7	23.63	0.67	0.22			3.4	27.39	0.80	0.17
I2SB	10	10	6.1	26.27	0.76	0.24	Bicubic	10	8.1	25.07	0.70	0.27	Gaussian	10	7.4	25.42	0.71	0.27
NADB			5.7	26.55	0.77	0.23			6.8	25.39	0.73	0.24			4.2	30.03	0.87	0.15
I2SB		100	3.6	25.34	0.72	0.24		100	4.1	23.74	0.65	0.27		100	3.9	23.98	0.66	0.27
NADB			3.5	25.81	0.74	0.23			4.1	24.46	0.70	0.22			3.1	30.34	0.88	0.13

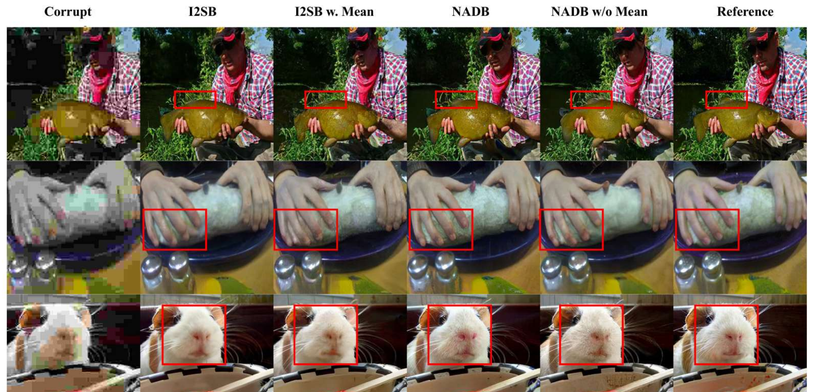


Table 2. Comparison of NADB with diffusion models (NFE=100). Reported results are taken from [21], and 4× Super-Resolution is evaluated on the full validation set, differing from Tab. 1.

JPEG Restoration			4× Super-Resolution			Deblurring		
QF	Method	FID↓	Filter	Method	FID↓	Kernel	Method	FID↓
5	DDRM	28.2	Pool	DDRM	14.8	Uniform	DDRM	9.9
	ΠGDM	8.6		DDNM	9.9		DDNM	3.0
	Palette	8.3		ΠGDM	3.8		Palette	4.1
	NADB	4.3		NADB	1.1		NADB	3.4
10	DDRM	16.7	Bicubic	DDRM	21.3	Gaussian	DDRM	6.1
	ΠGDM	6.0		DDNM	13.6		DDNM	2.9
	Palette	5.4		ΠGDM	3.6		Palette	3.1
	NADB	3.5		NADB	1.0		NADB	3.1

Table 4. Qualitative comparisons with four models on JPEG-5 restoration tasks.

QF	NFE	Method	FID ↓	PSNR ↑	SSIM↑	LPIPS ↓
5	10	I2SB	8.0	24.50	0.69	0.30
		NADB	6.9	24.45	0.69	0.30
		I2SB w. Mean	8.8	24.51	0.69	0.31
		NADB w/o Mean	7.0	24.36	0.69	0.30
	100	I2SB	4.7	23.60	0.65	0.31
		NADB	4.3	23.63	0.65	0.30
		I2SB w. Mean	4.4	23.64	0.65	0.31
		NADB w/o Mean	5.2	23.54	0.65	0.31

Our Research Subject

Diffusion Bridge Models (e.g., I2SB and DDBM)

1. The diffusion bridge model is a type of generative model. Unlike flow matching and diffusion processes, it takes two data distributions as endpoints and can directly achieve the transformation between these data distributions. Its structure is naturally suitable for image restoration tasks.

$$X_t = \underbrace{\frac{\overline{\sigma_t^2}}{\overline{\sigma_t^2}+\sigma_t^2} X_0 + \frac{\sigma_t^2}{\overline{\sigma_t^2}+\sigma_t^2} X_1}_{\text{Mean term}} + \underbrace{\sqrt{\frac{\sigma_t^2\,\overline{\sigma_t^2}}{\overline{\sigma_t^2}+\sigma_t^2}}\, Z}_{\text{Stochastic term}},$$

2. Mainstream diffusion bridge models are typically trained by mimicking the denoising paradigm of standard diffusion process:

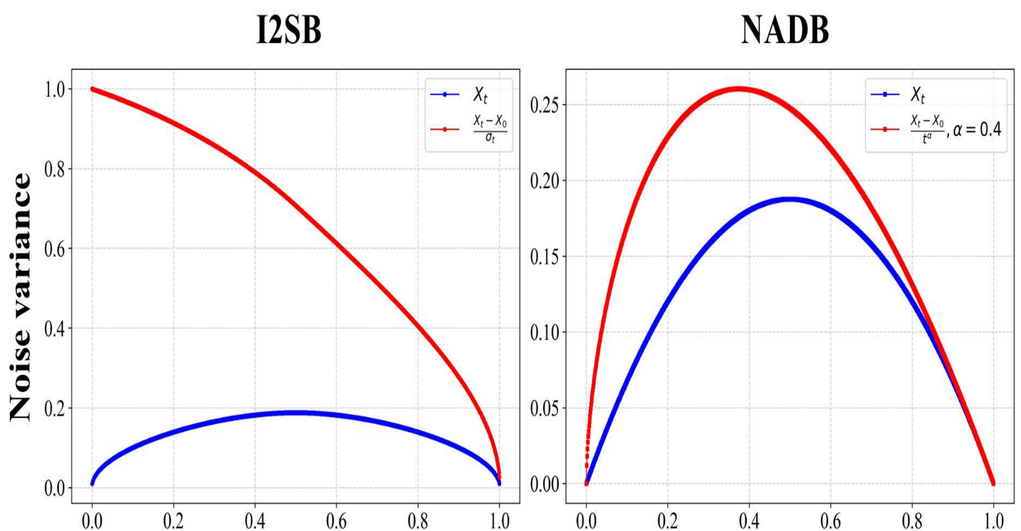
$$\mathcal{L} = \mathbb{E}_{t,\mathbf{X}_0,\mathbf{X}_1,Z}\left[\left\|\epsilon(X_t,t;\theta) - \frac{X_t - X_0}{\sigma_t}\right\|^2\right].$$

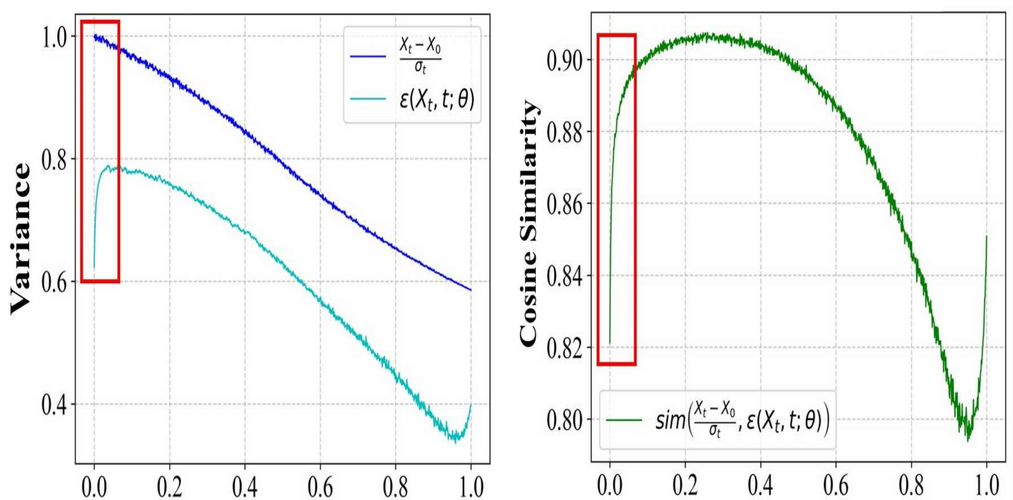
Noise Level Mismatch

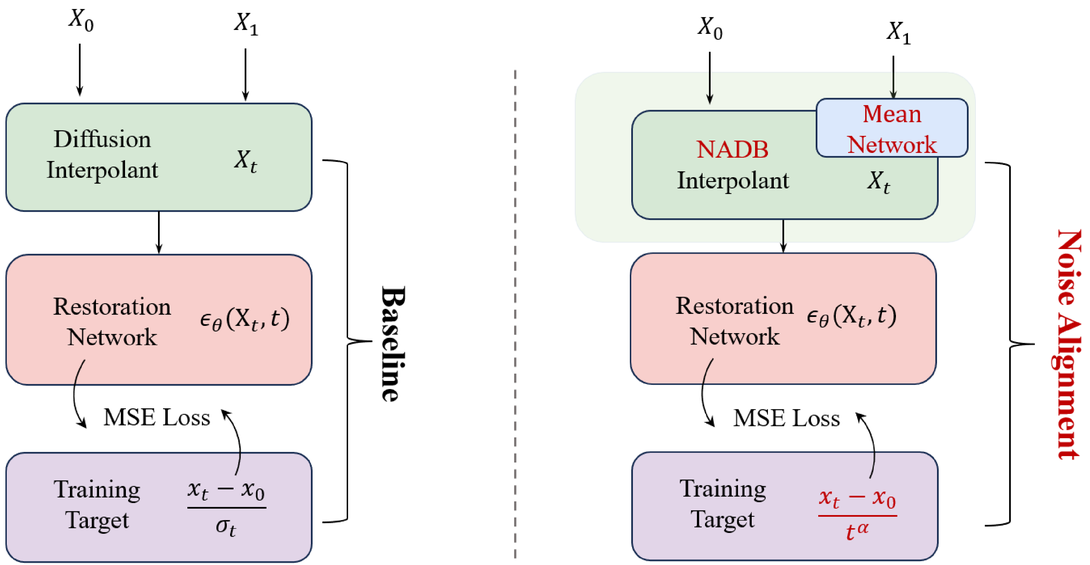
1.To quantify the network's failure on this ill-conditioned task, we assess the difference between the network's prediction and the ground truth target along two dimensions:

- **Magnitude:** We compare the Variance between the two. This measures whether the network predicts the correct amplitude of stochasticity.
- **Direction:** We compute the Cosine Similarity between the two. This measures whether the direction of the prediction is pointing in the correct direction.

2.A significant discrepancy exists at the target endpoint $(t \to 0)$: For the network input (X_t), its noise coefficient approaches 0 as $t \to 0$, making the input become nearly deterministic X_0. For the training target, its noise coefficient approaches 1 as $t \to 0$. The target becomes highly stochastic Z. This mismatch imposes an ill-conditioned task on the network: it predicts random noise from a clean and deterministic input.







Definition 3 (Magnitude-Aligned Stochastic Interpolant)**.** *Let $\alpha \in (0,1)$ and k be a finite constant. We define the magnitude-aligned stochastic interpolant X_t connecting the target $X_0 \sim \rho_0$ and the degraded $X_1 \sim \rho_1$ as:*

$$X_t := (1 - t^\alpha)X_0 + t^\alpha X_1 + kt(1-t)Z. \tag{5}$$

We define the corresponding magnitude-aligned training objective Y_t as the scaled displacement, which decomposes as:

$$Y_t := \frac{X_t - X_0}{t^\alpha} = (X_1 - X_0) + kt^{1-\alpha}(1-t)Z. \tag{6}$$

The training objective is to predict this target:

$$\mathcal{L}_{base} = \mathbb{E}_{t,X_0,X_1,Z}\left[\|\epsilon(X_t, t; \theta) - Y_t\|^2\right]. \tag{7}$$

Definition 4 (Mean Network)**.** *Given the joint distribution of the paired data $(X_0, X_1) \sim (\rho_0, \rho_1)$, we define a mean network $M(\cdot; \phi)$ trained to approximate the posterior mean $\mathbb{E}[X_0|X_1]$. The network's output is defined as:*

$$\hat{X}_0 = M(X_1; \phi). \tag{10}$$

The parameters ϕ are optimized by minimizing the mean squared error (MSE) objective:

$$\mathcal{L}_{MSE}(\phi) = \mathbb{E}_{(X_0,X_1)}\left[\|M(X_1; \phi) - X_0\|^2\right]. \tag{11}$$

Results

Table 1. Comparison with I2SB on numerous image restoration tasks.

Method	JPEG Restoration						4× Super-Resolution						Deblurring					
	QF	NFE	FID↓	PSNR↑	SSIM↑	LPIPS↓	Filter	NFE	FID↓	PSNR↑	SSIM↑	LPIPS↓	Kernel	NFE	FID↓	PSNR↑	SSIM↑	LPIPS↓
I2SB	5	10	8.0	24.50	0.69	0.30	Pool	10	7.3	24.86	0.70	0.27	Uniform	10	10.3	24.19	0.65	0.32
NADB			6.9	24.45	0.69	0.30			5.3	24.75	0.71	0.23			4.8	27.70	0.81	0.18
I2SB		100	4.7	23.60	0.65	0.31		100	4.1	23.57	0.65	0.26		100	5.0	22.47	0.58	0.32
NADB			4.3	23.63	0.65	0.30			3.7	23.63	0.67	0.22			3.4	27.39	0.80	0.17
I2SB	10	10	6.1	26.27	0.76	0.24	Bicubic	10	8.1	25.07	0.70	0.27	Gaussian	10	7.4	25.42	0.71	0.27
NADB			5.7	26.55	0.77	0.23			6.8	25.39	0.73	0.24			4.2	30.03	0.87	0.15
I2SB		100	3.6	25.34	0.72	0.24		100	4.1	23.74	0.65	0.27		100	3.9	23.98	0.66	0.27
NADB			3.5	25.81	0.74	0.23			4.1	24.46	0.70	0.22			3.1	30.34	0.88	0.13

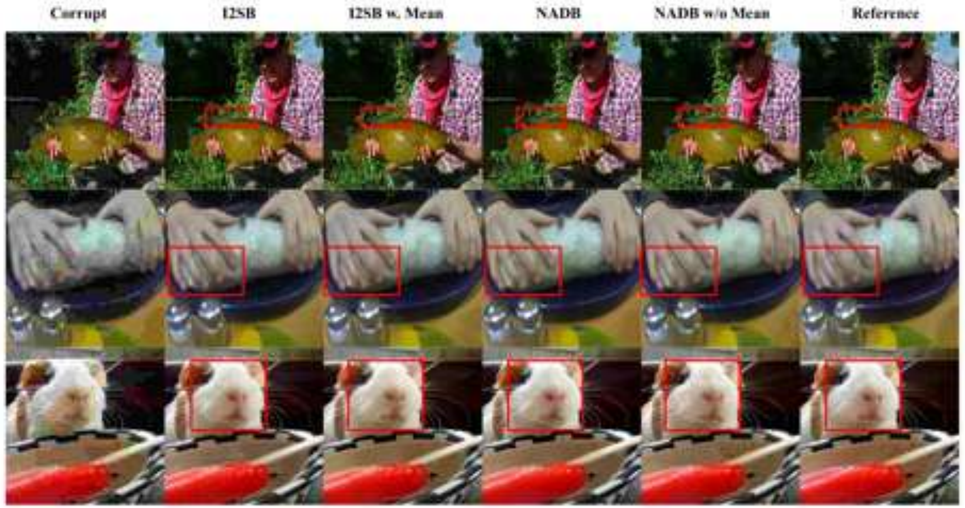


Table 2. Comparison of NADB with diffusion models (NFE=100). Reported results are taken from [21], and 4× Super-Resolution is evaluated on the full validation set, differing from Tab. 1.

JPEG Restoration			4× Super-Resolution			Deblurring		
QF	Method	FID↓	Filter	Method	FID↓	Kernel	Method	FID↓
5	DDRM	28.2	Pool	DDRM	14.8	Uniform	DDRM	9.9
	ΠGDM	8.6		DDNM	9.9		DDNM	3.0
	Palette	8.3		ΠGDM	3.8		Palette	4.1
	NADB	4.3		NADB	1.1		NADB	3.4
10	DDRM	16.7	Bicubic	DDRM	21.3	Gaussian	DDRM	6.1
	ΠGDM	6.0		DDNM	13.6		DDNM	2.9
	Palette	5.4		ΠGDM	3.6		Palette	3.1
	NADB	3.5		NADB	1.0		NADB	3.1

Table 4. Qualitative comparisons with four models on JPEG-5 restoration tasks.

QF	NFE	Method	FID ↓	PSNR ↑	SSIM↑	LPIPS ↓
5	10	I2SB	8.0	24.50	0.69	0.30
		NADB	6.9	24.45	0.69	0.30
		I2SB w. Mean	8.8	24.51	0.69	0.31
		NADB w/o Mean	7.0	24.36	0.69	0.30
	100	I2SB	4.7	23.60	0.65	0.31
		NADB	4.3	23.63	0.65	0.30
		I2SB w. Mean	4.4	23.64	0.65	0.31
		NADB w/o Mean	5.2	23.54	0.65	0.31