Adaptive Noise Covariance Scheduling under Riemannian Metrics for Diffusion Models

Bolin Deng[1], Die Hu[2] ✉, Bin Tan[3], and Jun Wu[1] ✉

[1] College of Computer Science and Artificial Intelligence, Fudan University, Shanghai, China
bldeng23@m.fudan.edu.cn, wujun@fudan.edu.cn
[2] College of Future Information Technology, Fudan University, Shanghai, China
hudie@fudan.edu.cn
[3] College of Electronics and Information Engineering, Jinggangshan University, Ji'an, China
tanbin@jgsu.edu.cn

Abstract. Diffusion models typically reconstruct coarse, low-frequency structure before recovering high-frequency detail during the reverse denoising. However, the standard forward process injects Gaussian white noise that corrupts all frequencies uniformly, creating a mismatch between forward corruption and reverse denoising and hindering the recovery of fine detail. We address this mismatch with an adaptive noise covariance schedule that evolves along the geodesic on the Symmetric Positive Definite (SPD) manifold under the Bures-Wasserstein (BW) metric. Specifically, we evolve the covariance along the BW geodesic from a blue-noise endpoint to a white-noise endpoint, so the noise power spectrum transitions smoothly from high-frequency emphasis to uniform. This transition better aligns forward corruption with reverse denoising and enhances textural detail. We provide theoretical and quantitative analyses showing that this transition path yields smoother noise power spectrum evolution and lower noise power fluctuations. In addition, we adapt the transition speed to the characteristics of each image, yielding finer forward-reverse alignment and improved perceptual quality. Experiments on multiple datasets show consistent gains over deterministic diffusion baselines in FID and KID. Our code and pretrained models are available at https://github.com/BolinDeng/RMDM.

Keywords: Time-varying noise covariance · Riemannian manifold · Diffusion models

1 Introduction

Diffusion models have become a central paradigm for generative modeling, building on the seminal formulation of Sohl-Dickstein *et al.* [27] and subsequent advances by Ho *et al.* [12] and Song *et al.* [30]. They typically surpass Generative Adversarial Networks (GANs) in perceptual quality and training stability

✉ Corresponding authors.

across image synthesis [9], text-to-image [25], video [6, 39], and audio [16]. A diffusion framework consists of a forward process that maps data to a simple prior and a reverse process that reconstructs data via iterative denoising. Recent progress in diffusion models has explored continuous-time formulations [31], improved samplers that reduce inference steps while preserving fidelity and stability [10, 15, 21, 28, 29], adaptive timestep selection and planning based on latent features or denoising difficulty [7, 37], and transformer-based denoising networks [22].

However, a key mismatch between reverse denoising and forward corruption remains underexplored. Specifically, the reverse process in diffusion models typically reconstructs low-frequency structure before recovering high-frequency detail, whereas the classical forward process injects Gaussian white noise with a flat power spectrum that corrupts all frequency components uniformly. This mismatch can hinder the synthesis of fine-grained textures. To mitigate it, Rissanen *et al.* [24] formulate diffusion as the inversion of the heat equation, explicitly connecting diffusion dynamics with frequency control. Huang *et al.* [14] introduce a time-varying forward noise that transitions from blue noise to white noise to better align the two processes.

Despite these advances, two limitations remain. **(i) Off-geodesic covariance transition.** Noise covariance matrices reside on the SPD manifold, yet existing methods overlook this and interpolate the Cholesky factors for blue noise and white noise covariances in Euclidean space (Cholesky interpolation), yielding covariance trajectories that deviate from geodesics on the SPD manifold. Such deviation results in longer path length under Riemannian metrics, larger noise power fluctuations, and anisotropic noise power spectrum evolution, which can complicate denoising, require more sampling steps and degrade perceptual quality. **(ii) Lack of adaptivity.** Different images exhibit distinct frequency characteristics, yet existing methods rely on a single transition schedule that is sensitive to dataset shifts, requires per-dataset retuning, and can degrade the synthesis of fine-grained textures.

To address these issues, we decouple the choice of the transition path on the SPD manifold from the transition schedule along it, enabling independent optimization of each component. For the transition path, we treat the noise covariance as a point on the SPD manifold and evolve it along the BW geodesic between the blue-noise and the white-noise endpoint. This geodesic has a closed-form expression and is the shortest path under the BW metric. As a result, it yields smaller noise power fluctuations (Fig. 1) and a smoother evolution of the noise power spectrum (Fig. 2) than Cholesky interpolation and other alternative geodesic choices (e.g., geodesics under Affine-Invariant Riemannian metric and Log-Euclidean metric), thereby better aligning forward corruption with reverse denoising and improving textural detail and perceptual quality. For the transition schedule, we introduce an adaptive blue-to-white schedule that modulates the transition speed per image, thereby avoiding per-dataset schedule retuning and yielding finer forward-reverse alignment, which improves perceptual quality.

In summary, our contributions are summarized as follows:

- We introduce a noise covariance transition path based on the BW geodesic on the SPD manifold. Our theoretical analysis and quantitative evaluation show that this path yields smoother noise power spectrum evolution and lower noise power fluctuations.
- We propose an adaptive covariance transition schedule that adjusts the speed of the blue-to-white covariance evolution based on image characteristics.
- Extensive experiments demonstrate consistent improvements over deterministic diffusion models on FID and KID, and ablation studies validate the effectiveness of each component.

2 Related Work

2.1 Diffusion Models

Classical diffusion models perturb data with independent Gaussian white noise in the forward process so that the distribution converges to an isotropic Gaussian prior as the corruption level increases. In parallel, recent studies have begun to explore alternative ways of corrupting data in the forward process [3, 10, 13, 14, 24, 26, 36]. For example, Voleti *et al.* [36] introduce anisotropic Gaussian noise to model correlations in the score-based framework and provide a detailed mathematical treatment. Bansal *et al.* [3] replace stochastic noising with completely deterministic degradations such as blur and masking. Heitz *et al.* [10] propose a minimalist deterministic forward process by iteratively blending the clean image and noise. Sahoo *et al.* [26] apply spatially varying noise intensities and re-examine the common assumption about ELBO invariance to the noise process. Moreover, Rissanen *et al.* [24] formulate diffusion as the inversion of the heat equation, explicitly connecting diffusion dynamics with frequency control. Building on this, Hoogeboom *et al.* [13] combine Gaussian noise with blurring to preserve structural fidelity while reintroducing stochasticity, leading to better generation quality. To further align the forward process with the empirical coarse-to-fine behavior of the reverse denoising, Huang *et al.* [14] introduce time-varying noise that transitions from blue noise to white noise. Most prior work focuses on what noise to inject or where to apply it, rather than how the noise covariance should evolve on the SPD manifold. We decompose the covariance evolution into a BW geodesic path on the SPD manifold and an adaptive covariance transition schedule, improving frequency alignment between forward corruption and reverse denoising.

2.2 Geodesics on the SPD Manifold

A manifold is locally Euclidean but may have a non-Euclidean global structure [18]. With a Riemannian metric, it becomes a Riemannian manifold that admits inner products on tangent spaces and induces notions of length, distance, and geodesics [1]. A geodesic between two endpoints is the shortest path between them with respect to the chosen metric and reflects the intrinsic geometry of the

manifold. Consequently, linear interpolation in Euclidean coordinates is generally not geometrically consistent on a curved manifold, even when the interpolated points remain within the set of valid points, because it ignores the Riemannian metric and its associated invariances. Since geodesics are metric-dependent, different metrics yield different shortest paths between the same endpoints.

The SPD manifold is the set of all real symmetric positive definite matrices, and noise covariance matrices lie on this space [4]. Several Riemannian metrics are commonly used for the SPD manifold, each inducing a specific geometry and a corresponding family of geodesics suited to different modeling goals. Specifically, the Log-Euclidean metric [2] maps SPD matrices to a flat vector space via the matrix logarithm and returns via the matrix exponential, providing closed-form operations with low computational cost while preserving positive-definiteness. The Affine Invariant Riemannian Metric [23] is invariant under congruence transformations and captures intrinsic curvature, which is useful when invariance and smoothness are essential. The Bures-Wasserstein metric arises from optimal transport between Gaussian measures and can be interpreted as the 2-Wasserstein distance between zero-mean Gaussians, minimizing quadratic transport cost between covariances [34, 35]. Different metrics induce different geodesics between the same endpoints, and the choice is task-dependent. In our setting, we adopt the BW geodesic, which yields smoother noise power spectrum evolution and lower power fluctuations than Euclidean Cholesky interpolation.

3 Method

To investigate how the time-varying noise covariance influences generative performance, we decompose the evolution of noise covariance into a transition path on the SPD manifold and a transition schedule that controls transition speed. To this end, we first review the necessary background.

3.1 Preliminary

Geodesics on the SPD manifold. Let $\mathcal{S}_{++}^n$ denote the set of $n \times n$ SPD matrices and let $A, B \in \mathcal{S}_{++}^n$. The closed-form expression of geodesic joining A and B under the Log-Euclidean metric (LE geodesic) is:

$$\Sigma_\gamma^{LE} = \exp((1-\gamma)\log A + \gamma \log B), \tag{1}$$

with $\gamma \in [0, 1]$. The closed-form expression of the geodesic joining A and B under the Affine-Invariant Riemannian metric (AIRM geodesic) is:

$$\Sigma_\gamma^{AIRM} = A^{1/2}(A^{-1/2}BA^{-1/2})^\gamma A^{1/2}. \tag{2}$$

The closed-form expression of BW geodesic joining A and B is defined as:

$$\Sigma_\gamma^{BW} = (1-\gamma)^2 A + \gamma^2 B + 2\gamma(1-\gamma)\, M. \tag{3}$$

where M is the geometric mean:

$$M = A^{1/2}(A^{-1/2}BA^{-1/2})^{1/2}A^{1/2}. \tag{4}$$

Covariance transition via Cholesky interpolation. Let white noise covariance be $\Sigma_w = I$, where I denotes the identity matrix, and blue noise covariance $\Sigma_b \in \mathcal{S}^n_{++}$ with Cholesky factorization $\Sigma_b = L_b L_b^\top$. Rather than directly modifying the noise covariance, Huang *et al.* [14] (referred to as BNDM) achieve noise covariance transition by performing linear interpolation of Cholesky factors in the Euclidean space (Cholesky interpolation) with the transition schedule $\gamma_t \in [0, 1]$:

$$\mathbf{C}^{CH}_{\gamma_t} = (1 - \gamma_t)\, L_b + \gamma_t\, I, \tag{5}$$

where $\mathbf{C}^{CH}_{\gamma_t}$ denotes the coloring operator via Cholesky interpolation and colors white noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$ to colored noise $\boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \Sigma^{CH}_{\gamma_t})$ by $\boldsymbol{\epsilon}_t = \mathbf{C}^{CH}_{\gamma_t}\boldsymbol{\epsilon}$, where $\Sigma^{CH}_{\gamma_t} = \mathbf{C}^{CH}_{\gamma_t}(\mathbf{C}^{CH}_{\gamma_t})^\top$. The transition schedule γ_t is defined by a normalized sigmoid function over $t \in [0, T]$ and is monotonically increasing:

$$\gamma_t = \frac{\sigma((s + (e - s)(t/T))/\tau) - \sigma(s/\tau)}{\sigma(e/\tau) - \sigma(s/\tau)}, \tag{6}$$

$$\sigma(x) = \frac{1}{1 + e^{-x}}, \tag{7}$$

and is fixed with global constants $s < e$ and $\tau > 0$. Here, t denotes the discrete timestep and T is the total number of diffusion timesteps.

For a fair comparison, we use BNDM as the host framework and replace only the transition path and schedule. The architecture, objective and hyperparameters remain unchanged. But our method can also be extended to DDIM [28].

3.2 Covariance Transition via the BW Geodesic

We replace the transition path induced by Cholesky interpolation with the BW geodesic on the SPD manifold. We first give the blue-to-white closed forms for time-varying covariance under different Riemannian metrics and the associated coloring operators, and then compare their properties to show why we choose the BW geodesic.

Geodesics from blue to white noise. For the blue-to-white noise transition controlled by the schedule γ_t, we set $A = \Sigma_b$, $B = I$. For the covariance matrix Σ_b, there exist an orthogonal matrix Q and a diagonal matrix Λ such that $\Sigma_b = Q\Lambda Q^\top$ and $\Sigma_b^{1/2} = Q\Lambda^{1/2}Q^\top$. The BW geodesic from Σ_b to I is:

$$\Sigma^{BW}_{\gamma_t} = \Big((1 - \gamma_t)\, Q\Lambda^{1/2}Q^\top + \gamma_t I\Big)^2, \tag{8}$$

and the associated coloring operator $\mathbf{C}^{BW}_{\gamma_t}$ is:

$$\mathbf{C}^{BW}_{\gamma_t} = (1 - \gamma_t)\, Q\Lambda^{1/2}Q^\top + \gamma_t\, I. \tag{9}$$

The LE and AIRM (LE/AIRM) geodesics share the same closed form, as do their associated coloring operators:

$$\Sigma^{LA}_{\gamma_t} = Q\Lambda^{1-\gamma_t}Q^\top, \tag{10}$$

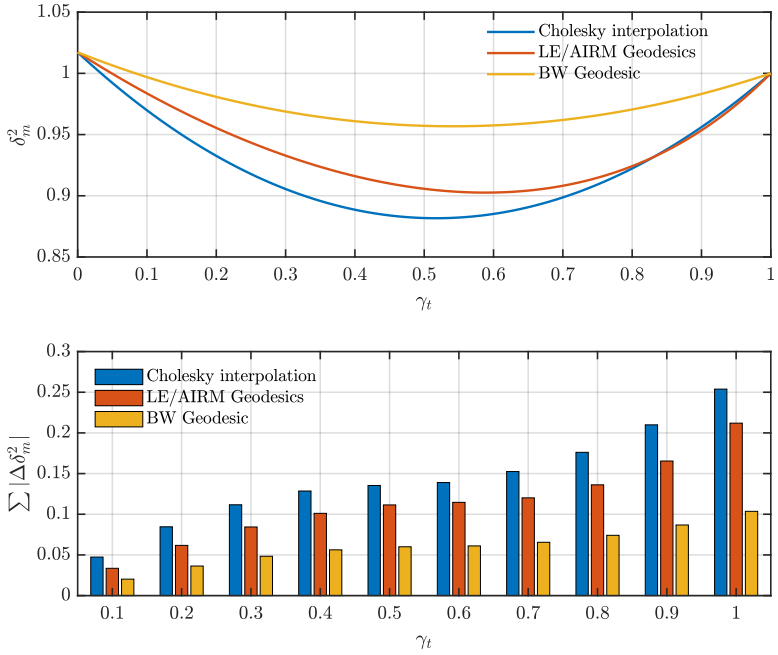

Fig. 1: Average noise power along the same transition schedule under different transition paths (*top*), and the cumulative absolute change in the average noise power (*bottom*).

$$\mathbf{C}_{\gamma_t}^{LA} = Q\Lambda^{(1-\gamma_t)/2}Q^\top. \tag{11}$$

Why the BW geodesic. We explain why we chose the BW geodesic as the transition path through three aspects.

Average noise power fluctuations. Along the BW geodesic, the noise power exhibits the smallest fluctuations, which benefits training stability and sample quality. Concretely, we quantify the average noise power δ_m^2 by the mean marginal variance of time-varying noise:

$$\delta_m^2 = \frac{1}{n}\,\mathrm{tr}\left(\Sigma_{\gamma_t}\right), \tag{12}$$

where tr denotes the matrix trace and Σ_{γ_t} denotes the noise covariance. As shown in Fig. 1, the BW geodesic yields the smallest fluctuations in noise power, LE/AIRM geodesics are intermediate, and the Cholesky interpolation induces the largest fluctuations. Smaller noise power fluctuations make adjacent time steps statistically closer, which reduces the variance of denoising targets and gradients and leads to more stable training.

Sensitivity to the schedule γ_t. Along the BW geodesic, $\mathbf{C}_{\gamma_t}^{BW}$ exhibits position-independent sensitivity to the schedule γ_t, which simplifies schedule tuning and

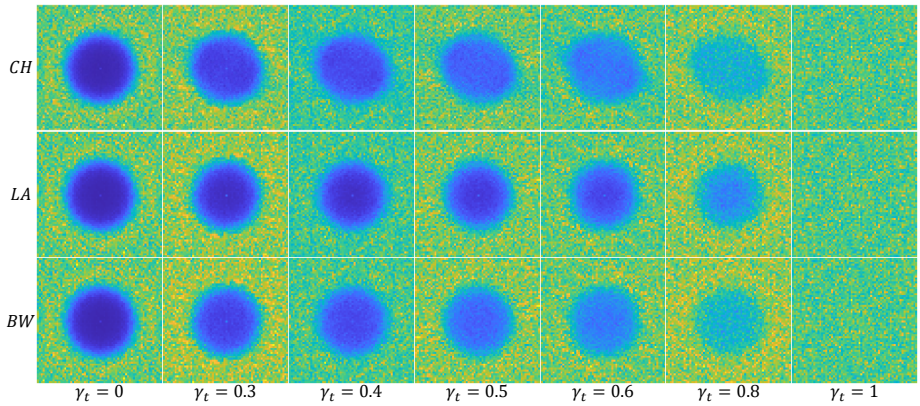

Fig. 2: Noise power spectrum evolution from blue noise to white noise under different transition paths on the SPD manifold. CH denotes Cholesky interpolation, which tends to introduce anisotropic, oblique spectral artifacts during the transition. LA denotes evolution along the LE/AIRM geodesic. BW denotes evolution along the BW geodesic, which produces a smoother, more isotropic power spectrum evolution.

yields uniform coloring operator updates across the trajectory. Formally, let $\mathbf{C}_{\gamma_t}$ map Gaussian white noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$ to noise with covariance Σ_{γ_t}. Its derivative with respect to the schedule γ_t is:

$$\mathbf{C}'_{\gamma_t} \triangleq \left.\frac{\partial \mathbf{C}_\gamma}{\partial \gamma}\right|_{\gamma=\gamma_t}, \tag{13}$$

which quantifies how fast the coloring operator changes per unit increment of γ_t. By the chain rule:

$$\frac{\partial \mathbf{C}_{\gamma_t}}{\partial t} = \mathbf{C}'_{\gamma_t} \frac{\partial \gamma_t}{\partial t}. \tag{14}$$

Hence, for the Cholesky interpolation:

$$(\mathbf{C}^{CH}_{\gamma_t})' = \frac{\partial \mathbf{C}^{CH}_{\gamma_t}}{\partial \gamma_t} = I - L_b. \tag{15}$$

For the LE/AIRM geodesics:

$$(\mathbf{C}^{LA}_{\gamma_t})' = \frac{\partial \mathbf{C}^{LA}_{\gamma_t}}{\partial \gamma_t} = -\frac{1}{2} Q \Lambda^{(1-\gamma_t)/2} \ln(\Lambda) Q^\top. \tag{16}$$

For the BW geodesic:

$$(\mathbf{C}^{BW}_{\gamma_t})' = \frac{\partial \mathbf{C}^{BW}_{\gamma_t}}{\partial \gamma_t} = I - Q \Lambda^{1/2} Q^\top. \tag{17}$$

$(\mathbf{C}^{CH}_{\gamma_t})'$ and $(\mathbf{C}^{BW}_{\gamma_t})'$ are constant. Therefore, $\frac{\partial \mathbf{C}^{CH}_{\gamma_t}}{\partial t}$ and $\frac{\partial \mathbf{C}^{BW}_{\gamma_t}}{\partial t}$ are controlled solely by $\frac{\partial \gamma_t}{\partial t}$, and are independent of the instantaneous value of γ_t, so the same

increment $\Delta\gamma_t$ induces the same update everywhere along the path. In contrast, for LE/AIRM geodesics, $\frac{\partial \mathbf{C}_{\gamma_t}^{LA}}{\partial t}$ depends explicitly on the instantaneous value of γ_t, making $\frac{\partial \mathbf{C}_{\gamma_t}^{LA}}{\partial t}$ more sensitive to the transition schedule γ_t. This position-dependence introduces extra complexity as the effect of the same $\Delta\gamma_t$ varies with the instantaneous value of γ_t, which makes the response more nonlinear and harder to control.

Evolution of the noise power spectrum. Along the BW geodesic, the noise power spectrum evolves isotropically, thereby reducing distribution drift across timesteps for the denoising network. Specifically, since $QQ^\top = I$, Eq. (8) can be rewritten as:

$$\Sigma_{\gamma_t}^{BW} = Q\Big((1-\gamma_t)\,\Lambda^{1/2} + \gamma_t I\Big)^2 Q^\top. \tag{18}$$

Eq. (10) and Eq. (18) show that along the BW and LE/AIRM geodesics the eigenvectors of the noise covariance remain fixed, while only the eigenvalues vary smoothly. Consequently, the noise power spectrum evolves isotropically from blue to white and is free of directional artifacts. In contrast, for the Cholesky interpolation, the eigenvectors of $\Sigma_{\gamma_t}^{CH}$ vary with γ_t and generally have no closed form (we prove this in the Supplementary material by projecting $\Sigma_{\gamma_t}^{CH}$ onto the eigenspace of Σ_b), yielding anisotropic power spectrum evolution. As shown in Fig. 2, for the power spectrum of $\Sigma_{\gamma_t}^{BW}$ and $\Sigma_{\gamma_t}^{LA}$, the central low-frequency hole shrinks steadily and the rings fade uniformly until the spectrum becomes flat (white noise), whereas the power spectrum of $\Sigma_{\gamma_t}^{CH}$ exhibits anisotropic and oblique artifacts. Such power spectrum drift increases the difficulty of denoising, undermines robustness, and ultimately degrades sample quality.

Overall, the BW geodesic provides smoother power fluctuations, position-independent sensitivity to the transition schedule, and isotropic power spectrum evolution. We therefore choose the BW geodesic as the covariance transition path.

3.3 Adaptive Transition Schedule

Since the $\mathbf{C}_{\gamma_t}^{BW}$ responds uniformly to the transition schedule, a single global parameter provides a parsimonious yet effective control of the schedule. We thus retain the normalized sigmoid family in Eq. (6) and make only its temperature τ adaptive. The adaptive transition schedule γ_t^a is defined as:

$$\gamma_t^a = \frac{\sigma((s+(e-s)(t/T))/\tau_{\mathbf{z}}) - \sigma(s/\tau_{\mathbf{z}})}{\sigma(e/\tau_{\mathbf{z}}) - \sigma(s/\tau_{\mathbf{z}})}, \tag{19}$$

where $\tau_{\mathbf{z}} > 0$ denotes the adaptive temperature parameter that controls the transition speed and is predicted from a latent $\mathbf{z}$ by a small MLP network p with parameter θ:

$$\tau_{\mathbf{z}} = p_\theta(\mathbf{z}). \tag{20}$$

The latent $\mathbf{z}$ is associated with the image $\mathbf{x_0}$. Specifically, we learn a Variational Autoencoder (VAE)-style posterior:

$$q_\phi(\mathcal{Z} \mid \mathbf{x_0}) = \mathcal{N}\left(\boldsymbol{\mu}_\phi(\mathbf{x_0}), \operatorname{diag}\boldsymbol{\sigma}_\phi^2(\mathbf{x_0})\right), \tag{21}$$

Algorithm 1 Training

Repeat
  $\mathbf{x_0} \sim p(\mathcal{X}_0),\ t \sim \mathcal{U}(\{1, ..., T\}),\ \boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$
  $\mathbf{z} \sim q_\phi(\mathcal{Z} \mid \mathbf{x_0})$
  $\tau_{\mathbf{z}} \leftarrow p_\theta(\mathbf{z}),\ \alpha_t \leftarrow t/T$
  $\gamma_t^a \leftarrow Eq.\ (19)$
  $\mathbf{C}_{\gamma_t^a}^{BW} \leftarrow (1-\gamma_t^a)\, Q\Lambda^{1/2}Q^\top + \gamma_t^a I$
  $\mathbf{x}_t \leftarrow \alpha_t \mathbf{C}_{\gamma_t^a}^{BW}\boldsymbol{\epsilon} + (1-\alpha_t)\mathbf{x_0}$
  Take gradient descent step on $\nabla_{\theta,\phi,\varphi}\,\mathcal{L}$
Until converged

Algorithm 2 Sampling

$\mathbf{x}, \mathbf{z} \sim \mathcal{N}(0, I)$
$\tau_{\mathbf{z}} \leftarrow p_\theta(\mathbf{z})$
for $t \leftarrow T$ **to** 1 **do**
    $\gamma_t^a,\ \gamma_{t-1}^a \leftarrow Eq.\ (19)$
    $\alpha_t \leftarrow t/T,\ \alpha_{t-1} \leftarrow (t-1)/T$
    $\mathbf{x} \leftarrow \mathbf{x} + \Delta\alpha_t f_\varphi^{(1)}(\mathbf{x}, t) + \Delta\gamma_t^a f_\varphi^{(2)}(\mathbf{x}, t)$
end for
return x

where $\mathbf{x_0}$ denotes the clean image, $\mathrm{diag}(\cdot)$ forms a diagonal covariance matrix from a vector, $\mathcal{Z}$ denotes the latent random variable, $\boldsymbol{\mu}_\phi$ and $\boldsymbol{\sigma}_\phi^2$ denote predicted mean and variance, respectively. During training, we sample $\mathbf{z} \sim \mathcal{Z}$ via the reparameterization trick:

$$\mathbf{z} = \boldsymbol{\mu}_\phi(\mathbf{x_0}) + \boldsymbol{\sigma}_\phi(\mathbf{x_0}) \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, I), \tag{22}$$

and regularize $\mathcal{Z}$ toward the prior $p(\mathcal{Z}) = \mathcal{N}(0, I)$ via a Kullback-Leibler (KL) regularizer term $\mathrm{KL}\left(q_\phi(\mathcal{Z} \mid \mathbf{x_0}) \,\|\, \mathcal{N}(0, I)\right)$. During the sampling process, we discard the VAE-style encoder q_ϕ and sample $\mathbf{z} \sim \mathcal{N}(0, I)$. Here, $\odot$ denotes the element-wise product.

Since we only modify the transition path and schedule of blue-to-white noise, the forward process and reverse process update rules remain identical to those in [14]. Under our adaptive transition schedule along the BW geodesic, the forward process is defined as:

$$\mathbf{x}_t = \alpha_t \mathbf{C}_{\gamma_t^a}^{BW}\boldsymbol{\epsilon} + (1-\alpha_t)\mathbf{x_0}, \tag{23}$$

where $\alpha_t \in [0, 1]$. The reverse process can be written as:

$$\mathbf{x}_{t-1} = \mathbf{x}_t + \Delta\alpha_t(\mathbf{x_0} - \mathbf{C}_{\gamma_t^a}^{BW}\boldsymbol{\epsilon}) + \Delta\gamma_t^a\alpha_{t-1}(Q\Lambda^{1/2}Q^\top - I)\boldsymbol{\epsilon}, \tag{24}$$

with $\Delta\alpha_t = \alpha_t - \alpha_{t-1}$ and $\Delta\gamma_t^a = \gamma_t^a - \gamma_{t-1}^a$. Let the two terms in Eq. (24) be the learning objectives of the denoising network:

$$\mathbf{h_1} = \mathbf{x_0} - \mathbf{C}_{\gamma_t^a}^{BW}\boldsymbol{\epsilon}, \tag{25}$$

and:

$$\mathbf{h_2} = \alpha_{t-1}(Q\Lambda^{1/2}Q^\top - I)\boldsymbol{\epsilon}. \tag{26}$$

We split the output of the denoising network f_φ into two parts, denoted as $f_\varphi^{(1)}(\mathbf{x}_t, t)$ and $f_\varphi^{(2)}(\mathbf{x}_t, t)$ to predict $\mathbf{h_1}$ and $\mathbf{h_2}$, respectively. Therefore, our loss function is defined as:

$$\begin{aligned}\mathcal{L} = \mathbb{E}_{\mathbf{x_0}\sim p(\mathcal{X}_0),\, t}\Big[f_\varphi^{(1)}(\mathbf{x}_t, t) - \mathbf{h_1})^2 + \frac{\Delta\gamma_t^a}{\Delta\alpha_t}(f_\varphi^{(2)}(\mathbf{x}_t, t) - \mathbf{h_2})^2 \\ + \beta\,\mathrm{KL}\left(q_\phi(\mathcal{Z} \mid \mathbf{x_0}) \,\|\, \mathcal{N}(0, I)\right)\Big],\end{aligned} \tag{27}$$

which optimizes θ, ϕ, φ jointly, where $p(\mathcal{X}_0)$ denotes the true data distribution, β is the weight for the KL regularizer and the KL term is applied per training example and encourages $q_\phi(\mathcal{Z} \mid \mathbf{x_0})$ to match the prior $p(\mathcal{Z}) = \mathcal{N}(0, I)$, ensuring training and sampling consistency. Training and sampling algorithms are summarized in Algorithm 1 and Algorithm 2, respectively, where $\mathcal{U}$ denotes the uniform distribution and the encoder q_ϕ is not used during sampling.

4 Experiments

4.1 Implementation Details

We conduct experiments on the CelebA [19], AFHQ [8], AFHQ-Cat [8], and LSUN-Church [38] datasets. Our framework is implemented based on the official codebase of Huang *et al.* [14]. The network architecture of the denoising network, the blue noise covariance Σ_b, and the hyperparameters remain consistent with Huang *et al.* [14]. Specifically, the architecture of the denoising network remains a U-Net. For the number of steps, we use $T = 1000$ during training and $T = 250$ for sampling. We set the s and e in Eq. (19) as 0 and 3, respectively. The encoder q_ϕ is a lightweight CNN followed by three residual blocks, and the temperature predictor p_θ is a two-layer MLP. For the learning rate of the denoising network with parameter φ is set to 1×10^{-4}, while the learning rate for the encoder and MLP are set to 1×10^{-5}, with the weight $\beta = 1$. Σ_b is obtained from offline statistics and $\Sigma_b^{1/2} = Q\Lambda^{1/2}Q^\top$ is computed via offline eigenvalue decomposition. Network parameters are optimized using the AdamW optimizer [20]. Evaluation metrics include FID [11], KID [5], Precision, and Recall [17] to measure the generative quality of all models. These metrics are computed based on the implementation in [32], using the Inception-v3 network [33] as the backbone. More details can be found in the Supplementary material.

4.2 Unconditional Image Generation

We compare our method with the stochastic baseline DDPM [12], as well as deterministic baselines DDIM [28], IHDM [24], IADB [10], and BNDM [14] for unconditional image generation. To ensure fair comparisons, we generate samples per dataset using the same initial Gaussian white noise for all methods.

Table 1: (a) Comparison of FID across multiple datasets in the pixel space. (b) Comparisons including FID, KID, Precision, and Recall in the pixel space.

FID (↓)	IHDM	DDPM	DDIM	IADB	BNDM	Ours
AFHQ-Cat (64^2)	11.02	9.75	9.82	9.18	7.95	**6.17**
AFHQ-Cat (128^2)	15.40	12.41	10.73	10.81	9.47	**8.47**
CelebA (64^2)	14.30	8.56	9.26	7.53	7.17	**6.19**
LSUN-Church (64^2)	17.76	13.07	16.46	9.95	11.35	**8.26**

(a)

Dataset	Method	FID↓	KID↓	Precision↑	Recall↑
AFHQ-Cat (64^2)	IADB	9.18	0.0062	0.72	0.37
	BNDM	7.95	0.0052	0.73	**0.49**
	Ours	**6.17**	**0.0034**	**0.79**	0.45
AFHQ-Cat (128^2)	IADB	10.81	0.0068	0.78	0.31
	BNDM	9.47	0.0063	0.78	**0.34**
	Ours	**8.47**	**0.0050**	**0.80**	0.32
LSUN-Church (64^2)	IADB	9.95	0.0073	**0.84**	0.44
	BNDM	11.35	0.0091	0.79	**0.48**
	Ours	**8.26**	**0.0064**	0.82	**0.48**
AFHQ (64^2)	IADB	11.06	0.0053	0.68	0.69
	BNDM	12.79	0.0069	0.61	**0.74**
	Ours	**9.51**	**0.0043**	**0.69**	0.69
AFHQ-Cat (256^2)	IADB	12.56	0.008	0.80	0.37
	BNDM	13.50	0.011	0.79	0.39
	Ours	**8.90**	**0.004**	**0.82**	**0.41**
CelebA (256^2)	IADB	19.25	0.012	0.80	**0.40**
	BNDM	16.11	0.010	0.83	0.38
	Ours	**12.14**	**0.006**	**0.86**	**0.40**

(b)

For BNDM, we adopt their transition schedules without additional tuning; a schedule-sensitivity check is provided in the Supplementary material.

As shown in Tab. 1a, our method achieves the best FID on all datasets, indicating consistent gains in perceptual quality. Tab. 1b further reports KID, Precision, and Recall for IADB, BNDM, and our method. Our approach attains the lowest KID on all datasets and achieves the highest Precision in most cases, while maintaining Recall comparable to baselines. Additionally, on the AFHQ (64^2) dataset, BNDM yields a higher FID score than IADB. A plausible explanation is that AFHQ has greater diversity (three domains of cat, dog, and wildlife), and a single transition schedule may not accommodate the frequency characteristics of all images, whereas our method remains effective. For qualitative evaluation, Fig. 3 visualizes image generation results on the AFHQ-Cat

Fig. 3: Image generation comparisons among IADB, BNDM and Ours on AFHQ-Cat dataset. Our method achieves the best visual quality, producing more realistic images with sharper textures and fewer artifacts.

Table 2: Comparison of the FID score between BNDM and Ours across different numbers of sampling steps T.

		$T=5$	$T=15$	$T=25$	$T=50$	$T=75$	$T=100$	$T=125$	$T=250$
AFHQ-Cat	BNDM	82.93	15.20	10.58	8.48	8.11	8.00	7.96	7.95
	Ours	**57.35**	**11.85**	**8.38**	**6.74**	**6.38**	**6.26**	**6.21**	**6.17**
AFHQ	BNDM	130.27	18.98	13.33	11.96	12.12	12.30	12.43	12.79
	Ours	**83.28**	**17.04**	**11.57**	**9.70**	**9.47**	**9.37**	**9.38**	**9.51**

dataset. Compared with IADB and BNDM, our method produces sharper eye contours, cleaner mouth and whiskers, richer fur textures, and fewer artifacts, with more natural color transitions. Overall, our method delivers the best FID and KID across all datasets while maintaining diversity comparable to baselines, and yields clear improvements in fine textures and reduced artifacts.

4.3 Extensions and Efficiency

We compare the FID score of our method with BNDM under different sampling steps T. As shown in Tab. 2, compared with BNDM, our method demonstrates superior performance across all step counts T. Moreover, we extend our approach to DDIM. As shown in Tab. 3, on the AFHQ-Cat dataset, the DDIM variant of our method achieves lower FID than the original DDIM and the BNDM-style DDIM extension. We also support conditional generation, such as image super-resolution. For image super-resolution, the learnable $\tau_{\mathbf{z}}$ that controls the blue-to-white noise transition schedule is predicted from the low-resolution input. The denoising network receives the noisy image and the low-resolution image

Table 3: Preliminary results on comparing DDIM, BNDM-(DDIM) and Ours-(DDIM) on the AFHQ-Cat dataset. BNDM-(DDIM) means extending BNDM to DDIM, Ours-(DDIM) means extending ours to DDIM, respectively.

Method	DDIM	BNDM-(DDIM)	Ours-(DDIM)
FID (↓)	9.82	7.11	**6.33**

Table 4: Image super-resolution performance of IADB, BNDM and our method. Our method outperforms IADB and BNDM in terms of PSNR.

PSNR (↑)	IADB	BNDM	Ours
AFHQ-Cat (64^2 to 128^2)	29.15	29.30	**30.01**
AFHQ-Cat (32^2 to 128^2)	23.91	23.94	**24.29**
AFHQ-Cat (128^2 to 256^2)	28.93	29.28	**29.65**
AFHQ (64^2 to 128^2)	29.05	28.98	**29.48**

concatenated along the channel dimension as its input. As shown in Tab. 4, our method shows consistent improvements over IADB and BNDM according to the PSNR metric.

For efficiency, we report the total parameter count, per-step training GFLOPs, and wall-clock time for training 100 epochs on CelebA-256 in Tab. 5. Our method introduces 0.98M additional parameters, with limited increases in computational cost and training time.

4.4 Ablation Study and Analysis

We conduct ablation experiments on the AFHQ-Cat dataset.

Different transition paths with fixed schedule. To confirm the role of the BW geodesic as a transition path, we compare it with results from Cholesky interpolation and the LE/AIRM geodesics under a shared fixed schedule. As shown in Tab. 6, the BW geodesic attains the best FID, KID, Precision and Recall, outperforming both Cholesky interpolation and the LE/AIRM geodesics, which matches our analysis in Sec. 3. We also observe that the LE/AIRM geodesics exhibit performance comparable to that of Cholesky interpolation. As discussed in Sec. 3, although Cholesky interpolation induces larger power fluctuations and anisotropic power spectrum evolution, LE/AIRM geodesics are more sensitive to the transition schedule. Consequently, with a fixed transition schedule γ_t, these effects partly offset, leading to similar performance.

Adaptive transition schedule for different paths. We enabled the adaptive transition schedule on different paths to investigate its impact. As shown in Tab. 6, the performance of LA+A shows a marked decline, which agrees with our analysis in Sec. 3 that the evolution of the coloring operator for LE/AIRM geodesics is more sensitive to the transition schedule, making it hard to control and train. This sensitivity is further amplified under the adaptive transition schedule and

Table 5: Computational efficiency on CelebA-256. We report the additional per-step training costs of the encoder (Enc.), predictor (Pred.), and coloring operation (CO), together with the total training GFLOPs.

Method	Parameters (M)	Training GFLOPs				Wall-clock (h)
		Enc.	Pred.	CO	Total	
IADB	118.9	–	–	–	196.91	5.7
BNDM	118.9	–	–	1.61	198.75	6.0
Ours	119.9	3.81	1.65e-5	1.61	202.56	6.2

Table 6: Ablation study on different combinations of blue-to-white noise transition path and schedule on AFHQ-Cat. F means a shared fixed transition schedule, and A means the adaptive transition schedule, respectively.

Transition methods	FID (↓)	KID (↓)	Precision (↑)	Recall (↑)
CH+F	7.95	0.0052	0.73	**0.49**
LA+F	7.95	0.0047	0.72	0.43
BW+F	**6.40**	**0.0034**	**0.76**	**0.49**
CH+A	6.43	0.0037	0.78	0.44
LA+A	8.36	0.0053	0.69	0.44
BW+A	**6.17**	**0.0034**	**0.79**	**0.45**

produces a larger degradation than with a fixed schedule. In contrast, both Cholesky interpolation and BW geodesic benefit from the adaptive schedule and achieve lower FID than their fixed-schedule versions. Overall, the combination of the BW geodesic and adaptive schedule achieves the best results.

5 Conclusion

We revisit the mismatch between forward corruption and reverse denoising in diffusion models and address it by modeling the forward noise with time-varying covariance on the SPD manifold. Specifically, we decompose the evolution of noise covariance into a transition path on the SPD manifold and a transition schedule. We adopt the BW geodesic as the path between the blue-noise and white-noise covariances and learn an adaptive schedule. Compared with off-geodesic Cholesky interpolation and geodesics under other metrics, the BW geodesic-based transition path yields lower power fluctuations and smoother noise power spectrum evolution across timesteps, which reduces distribution drift between consecutive timesteps and eases reverse denoising. Extensive experiments across multiple datasets and deterministic baselines show consistent improvements in FID and KID, and ablations further corroborate both the correctness of our theoretical analysis and the effectiveness of our method. Overall, our method

provides finer forward-reverse alignment and improved perceptual quality without modifying the denoising network.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 62271155, the National Key R&D Program of China under Grant 2020YFA0711400, the Key-Area Research and Development Program of Guangdong Province under Grant 2020B010166003.

References

1. Absil, P.A., Mahony, R., Sepulchre, R.: Optimization algorithms on matrix manifolds. Princeton University Press (2008)
2. Arsigny, V., Fillard, P., Pennec, X., Ayache, N.: Log-euclidean metrics for fast and simple calculus on diffusion tensors. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine **56**(2), 411–421 (2006)
3. Bansal, A., Borgnia, E., Chu, H.M., Li, J., Kazemi, H., Huang, F., Goldblum, M., Geiping, J., Goldstein, T.: Cold diffusion: Inverting arbitrary image transforms without noise. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 41259–41282. Curran Associates, Inc. (2023), https://proceedings.neurips.cc/paper_files/paper/2023/file/80fe51a7d8d0c73ff7439c2a2554ed53-Paper-Conference.pdf
4. Bhatia, R.: Positive definite matrices. In: Positive Definite Matrices. Princeton university press (2009)
5. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans. In: International Conference on Learning Representations (2018)
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. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 22563–22575 (2023), https://api.semanticscholar.org/CorpusID:258187553
7. Chen, P., Liu, X., Zhang, X., Shen, F., Gong, X., Liu, Z., Chen, Z., Hu, H., Wang, K., Lian, S.: Chain-of-trajectories: Unlocking the intrinsic generative optimality of diffusion models via graph-theoretic planning. arXiv preprint arXiv:2603.14704 (2026)
8. Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: Stargan v2: Diverse image synthesis for multiple domains. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8188–8197 (2020)
9. Dhariwal, P., Nichol, A.Q.: Diffusion models beat gans on image synthesis. Neural Information Processing Systems **34** (Dec 2021), https://dblp.uni-trier.de/db/journals/corr/corr2105.html#abs-2105-05233
10. Heitz, E., Belcour, L., Chambon, T.: Iterative α -(de)blending: a minimalist deterministic diffusion model. ACM SIGGRAPH 2023 Conference Proceedings (2023), https://api.semanticscholar.org/CorpusID:258547068

11. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
12. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
13. Hoogeboom, E., Salimans, T.: Blurring diffusion models. ArXiv **abs/2209.05557** (2022), https://api.semanticscholar.org/CorpusID:252211878
14. Huang, X., Salaun, C., Vasconcelos, C., Theobalt, C., Oztireli, C., Singh, G.: Blue noise for diffusion models. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–11 (2024)
15. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. arXiv:2206.00364 (2022)
16. Kong, Z., Ping, W., Huang, J., Zhao, K., Catanzaro, B.: Diffwave: A versatile diffusion model for audio synthesis. ArXiv **abs/2009.09761** (2020), https://api.semanticscholar.org/CorpusID:221818900
17. Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., Aila, T.: Improved precision and recall metric for assessing generative models. Advances in neural information processing systems **32** (2019)
18. Lee, J.M.: Smooth manifolds. In: Introduction to smooth manifolds, pp. 1–29. Springer (2003)
19. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (December 2015)
20. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
21. Lu, C., Zhou, T., Bao, J., Chen, C.: Dpm-solver: A fast ode solver for diffusion probabilistic models. In: NeurIPS (2022)
22. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV (2023)
23. Pennec, X., Fillard, P., Ayache, N.: A riemannian framework for tensor computing. International Journal of computer vision **66**(1), 41–66 (2006)
24. Rissanen, S., Heinonen, M., Solin, A.: Generative modelling with inverse heat dissipation. In: International Conference on Learning Representations (2023)
25. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
26. Sahoo, S., Gokaslan, A., De Sa, C.M., Kuleshov, V.: Diffusion models with learned adaptive noise. Advances in Neural Information Processing Systems **37**, 105730–105779 (2024)
27. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. pmlr (2015)
28. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv:2010.02502 (October 2020), https://arxiv.org/abs/2010.02502
29. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models. In: ICML (2023)
30. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. Advances in neural information processing systems **32** (2019)
31. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. International Conference on Learning Representations (May 2021), https://openreview.net/pdf?id=PxTIG12RRHS

32. Stein, G., Cresswell, J., Hosseinzadeh, R., Sui, Y., Ross, B., Villecroze, V., Liu, Z., Caterini, A.L., Taylor, E., Loaiza-Ganem, G.: Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. Advances in Neural Information Processing Systems **36**, 3732–3784 (2023)
33. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
34. Takatsu, A.: Wasserstein geometry of gaussian measures (2011)
35. Thanwerdas, Y., Pennec, X.: Bures–wasserstein minimizing geodesics between covariance matrices of different ranks. SIAM Journal on Matrix Analysis and Applications **44**(3), 1447–1476 (2023)
36. Voleti, V.S., Pal, C.J., Oberman, A.M.: Score-based denoising diffusion with non-isotropic gaussian noise models. ArXiv **abs/2210.12254** (2022), `https://api.semanticscholar.org/CorpusID:253098535`
37. Ye, Z., Chen, Z., Li, T., Huang, Z., Luo, W., Qi, G.J.: Schedule on the fly: Diffusion time prediction for faster and better image generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23412–23422 (2025)
38. Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., Xiao, J.: Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365 (2015)
39. Yu, S., Sohn, K., Kim, S., Shin, J.: Video probabilistic diffusion models in projected latent space. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 18456–18466 (2023), `https://api.semanticscholar.org/CorpusID:256868405`