Isometric Representation Learning for Disentangled Latent Space of Diffusion Models

Jaehoon Hahm [*] [1] **Junho Lee** [*] [1] **Sunghyun Kim** [1] **Joonseok Lee** [1] [2]

Abstract

The latent space of diffusion model mostly still remains unexplored, despite its great success and potential in the field of generative modeling. In fact, the latent space of existing diffusion models are entangled, with a distorted mapping from its latent space to image space. To tackle this problem, we present *Isometric Diffusion*, equipping a diffusion model with a geometric regularizer to guide the model to learn a geometrically sound latent space. Our approach allows diffusion models to learn a more disentangled latent space, which enables smoother interpolation, more accurate inversion, and more precise control over attributes directly in the latent space. Extensive experiments illustrate advantages of the proposed method in image interpolation, image inversion, and linear editing.

1. Introduction

Recently, diffusion models (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Ho et al., 2020; Song et al., 2021b) have achieved unprecedented success across multiple fields, including image generation (Dhariwal & Nichol, 2021; Nichol et al., 2022; Ramesh et al., 2022; Saharia et al., 2022; Rombach et al., 2022; Lee & Lee, 2024), image editing (Kawar et al., 2023; Ruiz et al., 2023; Hertz et al., 2022), video generation (Ho et al., 2022; Blattmann et al., 2023), and scientific applications (Cho et al., 2023). However, compared to other generative models like GANs (Goodfellow et al., 2014) or VAEs (Kingma & Welling, 2013), there are few studies exploring the latent space of diffusion models.

Learning a better latent space, particularly learning a *disentangled* latent space has been historically an important problem in generative modeling. The definition of a disentangled latent space varies depending on the field, but in generative modeling, it is defined as a latent space composed of linear subspaces, where each solely controls one factor of the variations (Bengio et al., 2013; Higgins et al., 2017). Through various literatures on GANs (Karras et al., 2020; Chen et al., 2016; Shen et al., 2020b; Kim & Mnih, 2018) and VAEs (Burgess et al., 2017; Chen et al., 2018), disentanglement is known to be beneficial for downstream tasks such as image interpolation, inversion, and editing. However, despite these benefits, only a few studies have addressed disentanglement of the latent space of diffusion models, possibly due to the relatively challenging analysis caused by their iterative sampling process.

Empirically exploring the latent space of diffusion models, we observe they are often entangled, aligned with recent discoveries and demonstration (Park et al., 2023; Peebles & Xie, 2023). For example, a naive latent walking by linear interpolation between two latent vectors produces unwanted intermediate images, as illustrated in Fig. 1 (top). Latent walking on a spherically interpolated trajectory between two latent vectors leads to a smoother intermediate images, as illustrated in Fig. 1 (mid), but it is still not a geodesic on the data manifold; on the trajectory between two men, it unnecessarily goes through an unrelated woman.

This can be interpreted that there exist some distortions in the latent space of diffusion models, implying that they fail to accurately reflect the geometry of the data manifold; geodesic of latent space is not necessarily mapped to geodesic on the data manifold. Such a misalignment often leads to entanglement of multiple semantic concepts, which induces a sub-optimal image interpolation, image inversion, or fine-grained image editing.

Motivated from the desire to guide diffusion models to learn a better disentangled latent space, we present *Isometric Diffusion*, a diffusion model equipped with isometric representation learning. Isometry is a map that preserves distance and angle between two metric spaces, and employing its geodesic preserving property, *Isometric Diffusion* guides to obtain a geometrically sound latent space that better reflects the data manifold. Specifically, we introduce a novel loss to encourage isometry between the latent space and the image

[*]Equal contribution [1]Seoul National University, Seoul, Korea [2]Google Research, Mountain View, California, United States. Correspondence to: Joonseok Lee <joonseok@snu.ac.kr>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

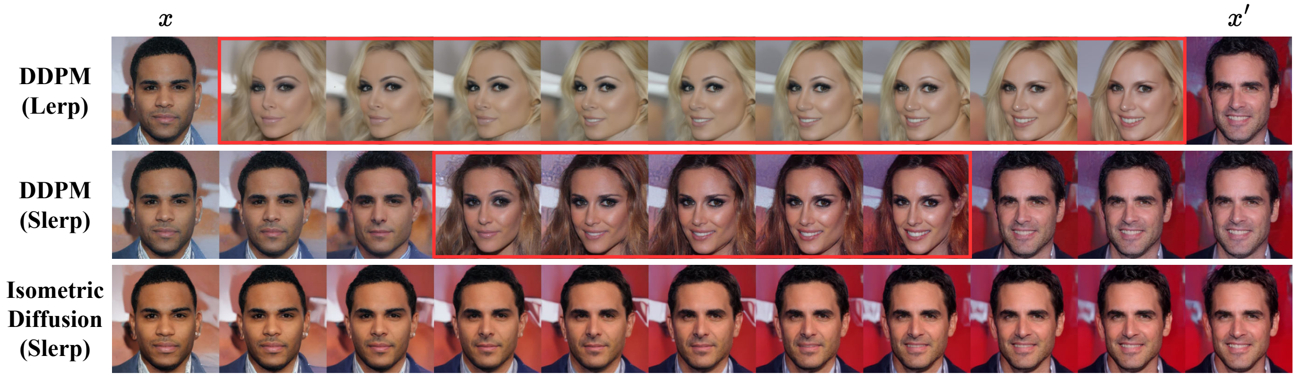


Figure 1. **An illustration of latent traversal between two latents $\boldsymbol{x}$ and $\boldsymbol{x}'$.** *Top*: naive linear interpolation (Lerp) assuming Euclidean space, *Mid*: spherical interpolation (Slerp) between $\boldsymbol{x}$ and $\boldsymbol{x}'$ (direction $\boldsymbol{x} \to \boldsymbol{x}'$ is entangled with unwanted gender axis inducing abrupt changes), *Bottom*: Slerp with the same latents with our Isometric Diffusion resolving unwanted entanglement.

space. With this additional guidance, latent walking induces a path closer to geodesic on the data manifold, and hence enables a smoother interpolation with less abrupt changes as in Fig. 1 (bottom).

To sum up, for the first time to the best of our knowledge, this paper proposes *Isometric Diffusion*, a diffusion model equipped with geometric considerations that lead to a better disentangled latent space. In order to obtain such geometrically sound latent space, we regularize the mapping from latent space to data manifold to be isometric. Our proposed method achieves superior disentanglement, without substantial degradation in quality of the generated images. We verify the effectiveness of our proposed method through quantitative and qualitative evaluations on various applications, including image interpolations, image inversions, and linear editing.

2. Background

We briefly review the sampling and inversion techniques using DDIM (Song et al., 2021a), latent spaces of diffusion models, and illustrate the objective for a better disentangled latent space.

2.1. Diffusion Model

Training. Given an observed image space, denoted by $\mathcal{X}_0$, the forward process of diffusion models repeatedly perturbs an image $\boldsymbol{x}_0 \in \mathcal{X}_0$ by $\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_0$, with noise $\epsilon_0 \sim \mathcal{N}(0, \boldsymbol{I})$ for $t = 1, ..., T$ where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$. These perturbed images $\boldsymbol{x}_t$ construct a chain of latent spaces $\{\mathcal{X}_t\}$ for $t \in \{1, ..., T\}$, where the intermediate latent space at each time step t is denoted by $\mathcal{X}_t$. For simplicity, we denote $\mathcal{X} \equiv \mathcal{X}_T$. To recover the original image $\boldsymbol{x}_0$ from $\boldsymbol{x}_T$, diffusion models train a score model $\mathbf{s}_\theta$ by minimizing the following denoising score matching loss (Vincent, 2011; Song et al., 2021b):

$$\mathcal{L}_{\text{dsm}}(t) = \lambda(t)\mathbb{E}_{\boldsymbol{x}_0}\mathbb{E}_{\boldsymbol{x}_t|\boldsymbol{x}_0}\left[\|\mathbf{s}_\theta(\boldsymbol{x}_t, t) - \nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t|\boldsymbol{x}_0)\|_2^2\right],$$

where θ is a set of learnable parameters of the score model and $\lambda(t)$ is a positive weighting function.

DDIM Sampling and Inversion. With the trained $\mathbf{s}_\theta$, we may generate an image $\boldsymbol{x}_0$ from a sample $\boldsymbol{x}_T \sim \mathcal{N}(0, \boldsymbol{I})$ through the reverse diffusion process. DDIM sampling accelerates the denoising process by skipping sampling steps (Song et al., 2021a; Dhariwal & Nichol, 2021):

$$\boldsymbol{x}_{t-1} = \sqrt{\tfrac{\bar{\alpha}_{t-1}}{\bar{\alpha}_t}}\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}\left(\sqrt{\tfrac{1}{\bar{\alpha}_{t-1}} - 1} - \sqrt{\tfrac{1}{\bar{\alpha}_t} - 1}\right)\epsilon_\theta(\boldsymbol{x}_t, t).$$

DDIM inversion finds the corresponding latent of a given image $\boldsymbol{x}_0$ by reversing the sampling process in the forward direction (Song et al., 2021a; Dhariwal & Nichol, 2021):

$$\boldsymbol{x}_{t+1} = \sqrt{\tfrac{\bar{\alpha}_{t+1}}{\bar{\alpha}_t}}\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t+1}}\left(\sqrt{\tfrac{1}{\bar{\alpha}_{t+1}} - 1} - \sqrt{\tfrac{1}{\bar{\alpha}_t} - 1}\right)\epsilon_\theta(\boldsymbol{x}_t, t).$$

2.2. Analysis on Latent Space $\mathcal{X}$ of Diffusion Models

The distribution of the norm of completely noised images $\|\boldsymbol{x}_T\|_2$ follows a χ-distribution, and they are distributed on the shell of a sphere, not uniformly within the sphere (see Sec. 3.1 for more details). For this reason, linearly interpolating two images within $\mathcal{X}$, as shown in Fig. 1 (top), results in a path far from geodesic on the data manifold, while spherical linear interpolation follows a shorter path. However, as seen in Fig. 1 (mid), the spherical linear interpolation is still semantically not disentangled, indicating that entangled regions exist in $\mathcal{X}_T$.

2.3. Intermediate Latent Space $\mathcal{H}$ as a Semantic Space

Kwon et al. (2023) discovers that diffusion models have a semantic latent space $\mathcal{H}$ in the intermediate feature space of its score model. They suggest that the learned intermediate feature space $\mathcal{H}$ of the score model $\mathbf{s}_\theta$ sufficiently represents the semantics of the observed images. Also, it is reported that a linear scaling by $\Delta\mathbf{h}$ on $\mathcal{H}$ controls the magnitude of semantic changes.

2.4. Path Length Regularizer

Motivated to obtain a disentangled and smoother latent space of GANs, path length regularizer (Karras et al., 2020) guides the generator $f : X \to Y$ to obtain a scaled-isometry, using an exponential moving average (EMA):

$$\mathcal{L}_{\text{pl}}(f) = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim\mathcal{N}(0,\boldsymbol{I})}\left[(\|\boldsymbol{J_x}^\top \boldsymbol{y}\|_2 - a)^2\right], \qquad (1)$$

where $\boldsymbol{x} \in X$ and $\boldsymbol{y} \in Y$ are random samples from a normal distribution, $\boldsymbol{J_x} = \frac{\partial f}{\partial \boldsymbol{x}}$ is the Jacobian of f, and a is the exponential moving average of $\|\boldsymbol{J_x}^\top \boldsymbol{y}\|_2$. The objective is minimized when $\boldsymbol{J}_x$ is orthogonal up to a global scale.

3. Isometric Representation Learning for Diffusion Models

The goal of our work is to learn a latent space $\mathcal{X}$ which better reflects the geometry of the training data manifold by encouraging the mapping between them to be closer to geodesic-preserving. We first explain the spherical approximation of latent space (Sec. 3.1), definition and geodesic preserving property of scaled isometry (Sec. 3.2), and how to guide the score model to learn an isometric mapping from $\mathcal{X}$ to $\mathcal{X}_0$ using a property of semantic latent space $\mathcal{H}$ (Sec. 3.3). Fig. 2 illustrates the overall flow of our approach. Lastly, we discuss computational considerations (Sec. 3.4).

3.1. Spherical Approximation of the Latent Space

Recall that the sampling process of diffusion models starts from a Gaussian noise, $\boldsymbol{x}_T \sim \mathcal{N}(0, \boldsymbol{I}_n) \in \mathbb{R}^n$, where T is the number of reverse time steps. Then, the radii of Gaussian noise vectors $\boldsymbol{x}_T$ follow χ-distribution: $r = \sqrt{\sum_{i=1}^n \boldsymbol{x}_{T,i}^2} \sim \chi(n)$, whose mean and variance are approximately $\sqrt{n - \frac{1}{2}}$ and 1, respectively. For a sufficiently large n (*e.g.*, $n = 256 \times 256 \times 3$), the noise vectors reside within close proximity of a hypersphere with $r = \sqrt{n - \frac{1}{2}}$.

From this observation, we approximate the noise vectors $\boldsymbol{x} \in \mathcal{X}$ (we omit subscripts to be uncluttered) reside on the hypersphere manifold $S^{n-1}(r) = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\| = r\}$. To define a Riemannian metric on $S^{n-1}(r)$, we need to choose charts and local coordinates to represent the Riemannian manifolds (Miranda, 1995). We choose the stereographic coordinates (Apostol, 1974) as the local coordinates to represent $\mathcal{X}$, and we set $\Phi = \text{id}$, the identity mapping defined at $\mathcal{H}$, which is the range of function we are interested in. Stereographic projection $\Pi_{n-1} : S^{n-1}(r) \setminus \{N\} \to \mathbb{R}^{n-1}$ is a bijective transformation from every point except for the north pole (N) on the hypersphere to a plane with the north pole as the reference point. Π_{n-1} and its inverse projection Π_{n-1}^{-1} are given by

$$\Pi_{n-1}(\boldsymbol{x}) = \frac{1}{r - \boldsymbol{x}_n}(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{n-1}), \qquad (2)$$
$$\Pi_{n-1}^{-1}(\boldsymbol{z}) = \frac{r}{|\boldsymbol{z}|^2 + 1}(2\boldsymbol{z}_1, 2\boldsymbol{z}_2, \cdots, 2\boldsymbol{z}_{n-1}, |\boldsymbol{z}|^2 - 1).$$

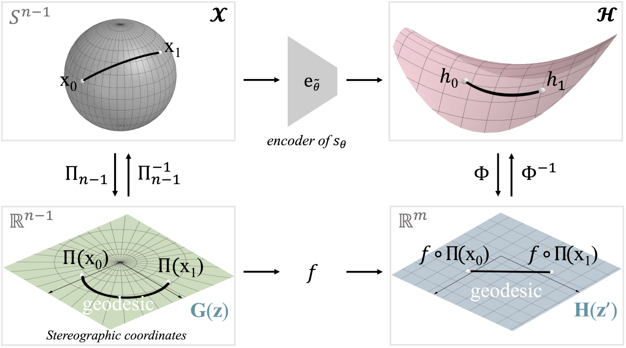


Figure 2. Illustration of $\mathcal{X}, \mathcal{H}$, and local coordinates of those two manifolds. Our isometric loss regularizes the encoder of the score model to map a spherical trajectory in $\mathcal{X}$ to a linear trajectory in $\mathcal{H}$, preserving a geodesic in $\mathcal{X}$ to a geodesic in $\mathcal{H}$. $e_{\tilde{\theta}}$ denotes the encoder of score model s_θ. Π_{n-1} and Φ are charts mapping from Riemmanian manifolds to local coordinate spaces. $\boldsymbol{z}, \boldsymbol{z}'$ denote the local coordinates of $\mathcal{X}, \mathcal{H}$, respectively.

In stereographic coordinates, the Riemannian metric of the $S^{n-1}(r)$ (do Carmo, 1992) is given by

$$\mathbf{G_s}(\boldsymbol{z}) = \frac{4r^4}{(|\boldsymbol{z}|^2 + r^2)^2}\boldsymbol{I}_{n-1}, \quad \forall \boldsymbol{z} \in \mathbb{R}^{n-1}. \qquad (3)$$

Recall that a diffusion model consists of a chain of latent spaces. Hence, it is needed to verify at every time step the validity of spherical approximation. From $\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon_0$, the variance of perturbation kernels is $\text{Var}[p(\boldsymbol{x}_t|\boldsymbol{x}_0)] = 1 - \bar{\alpha}_t = 1 - e^{\int -\beta(t)dt}$ (Song et al., 2021b). We use a linear noise schedule $\beta_t = \beta_0(1 - \frac{t}{T}) + \beta_T \frac{t}{T}$ with $\beta_t = 1 - \alpha_t$. We claim that for a sufficiently large t, $\sqrt{1 - \bar{\alpha}_t} \approx 1$ and thus the latent space can be approximated to a sphere. That is, we approximate $\mathcal{X}_t \approx S^{n-1}(r)$ with $r = \sqrt{1 - \bar{\alpha}_t}\mathbb{E}[\chi(n)] = \sqrt{(1 - \bar{\alpha}_t)n}$ for $t > pT$, where we set $p \in [0, 1]$ as a hyperparamter.

3.2. Isometric Mappings

Definition. A mapping between two Riemannian manifolds $\mathbf{f} : \mathcal{M}_1 \to \mathcal{M}_2$ (f in local coordinates; $f = \Phi \circ e_\theta \circ \Pi_{n-1}^{-1}$) is a *scaled isometry* (Lee et al., 2021) if and only if

$$\mathbf{G}(\boldsymbol{z}) = c\mathbf{J}_f(\boldsymbol{z})^\top \mathbf{H}(f(\boldsymbol{z}))\mathbf{J}_f(\boldsymbol{z}), \quad \forall \boldsymbol{z} \in \mathbb{R}^{n-1}, \qquad (4)$$

where $c \in \mathbb{R}$ is a constant, $\mathbf{J}_f(\boldsymbol{z}) = \frac{\partial f}{\partial \boldsymbol{z}} \in \mathbb{R}^{(n-1)\times m}$ is the Jacobian of f, $\mathbf{G}(\boldsymbol{z}) \in \mathbb{R}^{(n-1)\times(n-1)}$ and $\mathbf{H}(\boldsymbol{z}') \in \mathbb{R}^{m\times m}$ are the Riemannian metrics defined at the local coordinates $\boldsymbol{z}, \boldsymbol{z}'$ of $\mathcal{M}_1 = \mathbb{R}^{n-1}$ and $\mathcal{M}_2 = \mathbb{R}^m$, respectively.

Equivalently, f is a scaled isometry if and only if $\mathbf{J}_f^\top \mathbf{H} \mathbf{J}_f \mathbf{G}^{-1} = c\boldsymbol{I}$ where $c \in \mathbb{R}$ is a global constant. As its special case, f is called a strict isometry when $c = 1$, where a transformation between two metric spaces globally preserves distances and angles. Scaled isometry allows the constant c to vary, preserving only the *scaled* distances and angles. This relaxation makes it easier to optimize a function to preserve geodesic with less restrictions, hence leading to easier and more stable training than strict isometry.

In our problem formulation, $\mathcal{M}_1 = S^{n-1}$ ($\mathcal{X}$), $\mathcal{M}_2 = \mathbb{R}^m$ ($\mathcal{H}$), and $\mathbf{H}(\boldsymbol{z}') = \boldsymbol{I}_m$, as introduced in Sec. 3.1. Although evaluation of $\mathbf{J}_f^\top \mathbf{H} \mathbf{J}_f \mathbf{G}^{-1}$ is coordinate-invariant, our choice of stereographic coordinates is computationally advantageous, as its Riemannian metric in Eq. (3) is proportional to the identity matrix (see Sec. 3.4 for details).

Properties. To motivate the use of isometric mapping to learn disentangled representation, we introduce two important properties that isometry satisfies: geodesic-preserving and angle-preserving. We follow the definition of disentanglement from Bengio et al. (2013) and Higgins et al. (2017), which argue that a disentangled representation can be defined as one where a single latent unit is sensitive solely to changes in a single generative factor, while being invariant to changes in other factors.

1) Geodesic-preserving Property. Distance-preserving property of isometry naturally guarantees geodesic-preserving:

$$\begin{aligned}&\arg\min_{\gamma(t)} \int_0^1 \sqrt{\dot{\gamma}(t)^\top \mathbf{G}(\gamma(t))\dot{\gamma}(t)}\mathrm{d}t \qquad (5)\\ &= \arg\min_{\gamma(t)} \int_0^1 \sqrt{\dot{\gamma}(t)^\top \mathbf{J}(\gamma(t))^\top \mathbf{H}(f(\gamma(t)))\mathbf{J}(\gamma(t))\dot{\gamma}(t)}\mathrm{d}t,\end{aligned}$$

for an arbitrary trajectory $\gamma : [0, 1] \to \mathbb{R}^n$ in local coordinates of $\mathcal{M}_1$ with fixed endpoints ($\gamma(0) = \boldsymbol{x}_0, \gamma(1) = \boldsymbol{x}_1$), where $\boldsymbol{x}_0, \boldsymbol{x}_1 \in \mathbb{R}^n$ are constant vectors and $\dot{\gamma}(t) = \frac{d\gamma}{dt}(t)$.

This property induces equal sensitivity of each latent basis vector; a fixed-size step in the latent space results in equal amount of change in the semantic space, which is related to obtaining a smooth latent space.

2) Angle-preserving Property. This follows from the fact that if $G(x) = cJ^\top(x)H(f(x))J(x)$, then

$$\begin{aligned}\cos(\theta_1) &= \frac{\langle v_1, v_2\rangle_{\mathcal{M}_1}}{\|v_1\|_{\mathcal{M}_1}\|v_2\|_{\mathcal{M}_1}}\\ &= \frac{\langle df_p(v_1), df_p(v_2)\rangle_{\mathcal{M}_2}}{\|df_p(v_1)\|_{\mathcal{M}_2}\|df_p(v_2)\|_{\mathcal{M}_2}} = \cos(\theta_2),\end{aligned} \tag{6}$$

where $\langle v_1, v_2\rangle_{\mathcal{M}_1} = \dot{x_1}(0)^\top G\dot{x_2}(0)$, $\langle df_p(v_1), df_p(v_2)\rangle_{\mathcal{M}_2} = \dot{y_1}(0)^\top H\dot{y_2}(0) = \dot{x_1}(0)^\top J^\top HJ\dot{x_2}(0)$, and df_p is the pushforward at p. $x_1(t), x_2(t), y_1(t), y_2(t)$ are the trajectories on manifolds $\mathcal{M}_1, \mathcal{M}_2$ such that $x_1(0) = p$, $x_2(0) = p$, $y_1(0) = f(p)$, $y_2(0) = f(p)$, and $\dot{x} = \frac{dx}{dt}(t)$.

Recalling the semantic space $\mathcal{H}$ discovered by Kwon et al. (2023), we pose that an orthogonal basis corresponding to meaningful visual attributes exists in the semantic space. Due to the angle-preserving property, if the latent space $\mathcal{X}$ is mapped to $\mathcal{H}$ with an isometry, there exists orthogonal basis of $\mathcal{X}$ which is mapped to an orthogonal basis of $\mathcal{H}$ (assuming existence of the inverse). This implies that a vector corresponding to a specific attribute is mapped to a single latent vector, orthogonal to other latent vectors corresponding to other factors. This is related to the desired property of a disentangled latent space.

3.3. Isometry Loss for Diffusion Models

Isometry Loss. To sum up, we can encourage the mapping $\mathbf{f} : \mathcal{X} \to \mathcal{H}$ to preserve geodesics and angles by regularizing $\mathbf{R}(\boldsymbol{z}) \equiv \mathbf{J}_f(\boldsymbol{z})^\top \mathbf{H}(f(\boldsymbol{z}))\mathbf{J}_f(\boldsymbol{z})\mathbf{G}^{-1}(\boldsymbol{z}) = c\boldsymbol{I}$, for some $c \in \mathbb{R}$. It can be achieved by minimizing the following isometry loss (Lee et al., 2021):

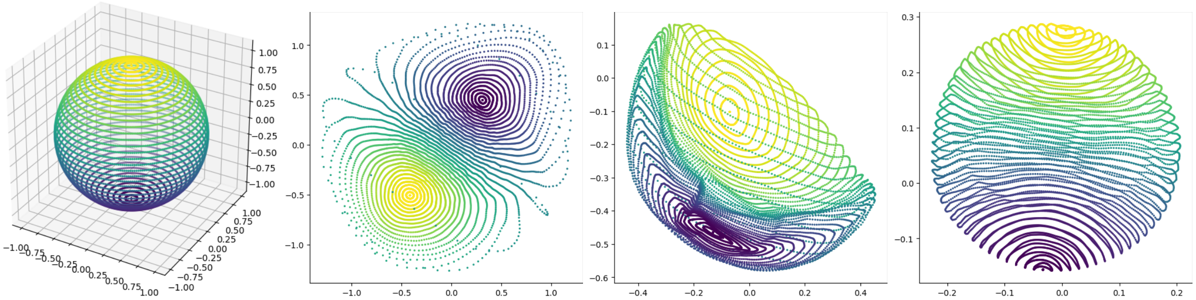
$$\begin{aligned}\mathcal{L}_{\text{iso}}(f, t) &= \frac{\mathbb{E}_{\boldsymbol{x}_t\sim P(\boldsymbol{x}_t)}[\mathrm{Tr}(\mathbf{R}^2(\boldsymbol{x}_t))]}{\mathbb{E}_{\boldsymbol{x}_t\sim P(\boldsymbol{z}_t)}[\mathrm{Tr}(\mathbf{R}(\boldsymbol{z}_t))]^2} \qquad (7)\\ &= \frac{\mathbb{E}_{\boldsymbol{x}_t\sim P(\boldsymbol{x}_t)}\mathbb{E}_{\boldsymbol{v}\sim\mathcal{N}(0,\boldsymbol{I})}[\boldsymbol{v}^\top\mathbf{R}(\boldsymbol{z}_t)^\top\mathbf{R}(\boldsymbol{z}_t)\boldsymbol{v}]}{\mathbb{E}_{\boldsymbol{x}_t\sim P(\boldsymbol{x}_t)}\mathbb{E}_{\boldsymbol{v}\sim\mathcal{N}(0,\boldsymbol{I})}[\boldsymbol{v}^\top\mathbf{R}(\boldsymbol{z}_t)\boldsymbol{v}]^2},\end{aligned}$$

where $P(\boldsymbol{x}_t)$ is the noise probability distribution at timestep t, and $\boldsymbol{z}_t = \Pi_{n-1}(\boldsymbol{x}_t)$. The second equality holds due to the stochastic trace estimator (Hutchinson, 1989), where $\boldsymbol{v} \in \mathbb{R}^{n-1}$ is a random vector such that $\mathbb{E}[\boldsymbol{v}\boldsymbol{v}^\top] = \boldsymbol{I}$.

Applying to Diffusion Models. Applying the isometry regularizer directly to the generating path of diffusion models is intractable, due to its iterative nature of sample generation. Specifically, calculating $\mathbf{R}(\boldsymbol{z})$ in Eq. (7) requires the Jacobian of $f = f_0 \circ \cdots \circ f_{N-1}$, where f_i is the i-th reverse step and N is the number of reverse steps, resulting in a long chain of function compositions.

Motivated from the training method of diffusion models, we apply isometric regularizer at each time step. To guide a mapping from $\mathcal{X}_T$ to $\mathcal{X}_0$ to be geodesic-preserving, we regularize each timestep of the iterative sequence; that is, the mapping between $\mathcal{X}_t$ and $\mathcal{X}_{t-1}$ for all $t \in \{T, ..., 1\}$. Instead of regularizing all steps, we may selectively apply it. For time steps closer to T, samples are closer to a Gaussian, so our assumption may reasonably hold. For time steps closer to 0, samples are not sufficiently perturbed yet and thus they would follow some intermediate distribution between the Gaussian and the original data distribution. Therefore, applying isometry loss to all timesteps can be sub-optimal and we let the portion of timesteps to apply it as a hyperparameter.

Also, to address the entanglement problem, we need to consider the semantic space of images rather than the pixel space. Hence, we assume the semantic gap between images



(a) S^2 manifold (b) Reconstruction only (c) $\mathcal{L}_{\text{iso}}$ with $\mathbf{G} = \boldsymbol{I}$ (d) $\mathcal{L}_{\text{iso}}$ with $\mathbf{G} = \mathbf{G_s}$ (Eq. (7))

Figure 3. (a) The input S^2 manifold. (b–d) Mapped contours in latent coordinates learned by an autoencoder; (b) with reconstruction loss only, (c) with isometric loss assuming navie Euclidean geometry, and (d) with our isometric loss considering S^2 geometry.

as a distance metric on $\mathcal{X}_T$. The desired objective can be achieved by guiding the encoder of the score model, or equivalently a mapping from $\mathcal{X}_t$ to $\mathcal{H}_t$, to be more isometric. Thus, we let $\mathbf{f} = e_{\tilde{\theta}}$, where $e_{\tilde{\theta}}$ denotes the encoder of score model s_θ, and apply the isometry loss in Eq. equation 7.

Our overall loss to train the score model is given by

$$\mathcal{L}(t) = \mathcal{L}_{\text{dsm}}(t) + \lambda_{\text{iso}}(\gamma, t)\mathcal{L}_{\text{iso}}(e_{\tilde{\theta}}, t), \tag{8}$$

where $\lambda_{\text{iso}}(p, t)$ is a non-negative weighting function and $\gamma \in [0, 1]$ is the ratio of timesteps to skip $\mathcal{L}_{\text{iso}}$. That is, $\lambda_{\text{iso}}(\gamma, t) = \lambda_{\text{iso}} \mathbf{1}_{t' > \gamma T}(t' = t)$ where $\mathbf{1}(\cdot)$ is the indicator function, and the denoising process starts from $t = T$.

Comparison with Path Length Regularizer. Calculated with exponential moving average (EMA), the path length regularizer (Karras et al., 2020) may not equally penalize two mappings equivalent up to a global scale; that is, it may not hold $\mathcal{L}_{\text{pl}}(f) = \mathcal{L}_{\text{pl}}(f')$ even if ${\mathbf{J}_f}^\top \mathbf{J}_f = c{\mathbf{J}'_f}^\top \mathbf{J}'_f$ holds for some $c \in \mathbb{R}^+$, potentially leading to sub-optimal training. In contrast, our isometric regularizer is scale-free, and does not require EMA-based optimization. Thus, isometric regularizer can be seen as a generalization of the path length regularizer, and helps to find the optimal point achieving disentanglement without significant degradation in generation quality. We empirically demonstrate this in Sec. 4.2.

Illustration. We illustrate the purpose of isometric representation learning with a toy autoencoder example, learning an encoding map from S^2 to $\mathbb{R}^2$. The autoencoder is trained with the reconstruction loss, regularized by our isometric loss in Eq. (7). Fig. 3 illustrates an autoencoder flattening the given S^2 manifold in (a) with three different losses. Only with the reconstruction loss, we see that the manifold in (b) is significantly distorted, often locating two far-away points in the input closely in the latent space. We observe clearly less distortion with the isometric loss in (c), under the assumption of the Euclidean metric in local coordinates of S^2 ($\mathbf{G} = \mathbf{I}$), but it still does not perfectly preserve geodesic. With our full loss in (d), we see that the geometry of the input space is better preserved with $\mathbf{G} = \mathbf{G}_s$ from Eq. (3). We provide more illustrations in Appendix G.

3.4. Computational Considerations

To sidestep the heavy computation of full Jacobian matrices, we use stochastic trace estimator to substitute the trace of Jacobian to Jacobian-vector product (JVP). Exploiting the commutativity and symmetry of the Riemmanian metric in stereographic coordinates, we utilize $\mathbb{E}_{\boldsymbol{v} \sim \mathcal{N}(0,I)}[\boldsymbol{v}^\top \mathbf{J}^\top \mathbf{J} \mathbf{G}^{-1} \boldsymbol{v}] = \mathbb{E}_{\boldsymbol{v} \sim \mathcal{N}(0,\boldsymbol{I})}[\boldsymbol{v}^\top \sqrt{\mathbf{G}^{-\top}} \mathbf{J}^\top \mathbf{J} \sqrt{\mathbf{G}^{-1}} \boldsymbol{v}] = \mathbb{E}_{\boldsymbol{v} \sim \mathcal{N}(0,\boldsymbol{I})}[(\mathbf{J}\sqrt{\mathbf{G}^{-1}}\boldsymbol{v})^\top \mathbf{J}\sqrt{\mathbf{G}^{-1}}\boldsymbol{v}]$ to reduce the number of JVP evaluations. We provide more details about the computation of stochastic trace estimator in Appendix F.2.

4. Experiments

We conduct extensive experiments to verify the effectiveness of our isometric loss $\mathcal{L}_{\text{iso}}$ on disentangling the latent space of diffusion models. We obtain experimental results by fine-tuning a pre-trained model with our $\mathcal{L}_{\text{iso}}$, unless noted otherwise. Refer to Appendix. A for further details.

4.1. Experimental Settings

Dataset. We evaluate our approach on CIFAR-10, CelebA-HQ (Huang et al., 2018), LSUN-Church and LSUN-Bedrooms (Wang et al., 2017). The training partition of each dataset consists of 50K, 14K, 126K, and 3M samples, respectively. We resize each image to 256 × 256 except for CIFAR-10 and horizontally flip it with probability 0.5.

Evaluation Metrics. *Fréchet inception distance (FID)* (Heusel et al., 2017) is a widely-used metric to assess the quality of images created by a generative model by comparing the distribution of generated images with that of ground truth images. *Perceptual Path Length (PPL)* (Karras et al., 2019) evaluates how well the generator interpolates between points in the latent space, defined as PPL $= \mathbb{E}[\frac{1}{\epsilon^2} d(\boldsymbol{x}_t, \boldsymbol{x}_{t+\epsilon})]$, where $d(\cdot, \cdot)$ is a distance function. We use LPIPS (Zhang et al., 2018) distance using AlexNet (Krizhevsky et al., 2012) for d. A lower PPL in-

Table 1. **Quantitative comparison.** Diffusion models trained with our isometry loss achieve consistent improvement over the baselines.

		FID-10k↓		PPL-50k↓		mRTL↓		MCN ↓		VoR ↓		LS ↓	
Dataset	Model	Base	Ours	Base	Ours	Base	Ours	Base	Ours	Base	Ours	Base	Ours
CIFAR-10	DDPM	**10.19**	10.50	126	**101**	2.03	**1.92**	155	**107**	**0.50**	0.57	-	-
LSUN-Church	DDPM	**10.56**	12.10	2028	**1559**	3.71	**3.21**	375	**217**	1.92	**1.37**	-	-
LSUN-Bedrooms	DDPM	**11.95**	12.02	4515	**3809**	3.38	**3.21**	320	**186**	1.69	**1.12**	-	-
CelebA-HQ	DDPM	**15.89**	16.18	648	**455**	2.67	**2.50**	497	**180**	1.42	**0.85**	1.91	**1.51**
CelebA-HQ	LDM	**10.79**	11.46	439	**397**	2.89	**2.73**	322	**198**	1.04	**0.54**	2.38	**2.15**

dicates a better disentangled latent space, since when two or more axes are entangled and geodesic interpolation in $\mathcal{X}$ induces a sub-optimal trajectory in the semantic space, the LPIPS distance gets larger and thereby so does the PPL. We perform 20 and 100 steps of DDIM sampling for FID and PPL, computed with 10,000 and 50,000 images, respectively. *Linear separability (LS)* (Karras et al., 2019) measures the degree of entanglement of a latent space, by measuring how much the latent space is far from being separable by a hyperplane. Since LS requires attributes, we measure it only on CelebA-HQ. *Mean condition number* (MCN) and *variance of Riemannian metric* (VoR) measure how close a mapping is to a scaled-isometry, proposed by Lee et al. (2021). We provide further details on these metrics in Appendix B.

We additionally design a new metric called *mean Relative Trajectory Length (mRTL)*, measuring the extent to which a trajectory in $\mathcal{X}$ is mapped to geodesic in $\mathcal{H}$. Specifically, mRTL is defined as the mean ratio between the L_2 distance $d_2(t)$ of $\boldsymbol{h}, \boldsymbol{h}' \in \mathcal{H}$, corresponding to two latents $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, and another distance measured on the manifold $d_{\mathcal{M}}(t)$, following the mapped path on $\{\mathcal{H}_t\}$. That is, $\text{RTL}(t) = \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}' \in \mathcal{X}}\left[d_{\mathcal{M}}(t)/d_2(t)\right]$ and $\text{mRTL} = \mathbb{E}_t[\text{RTL}(t)]$, where t denotes the timesteps of the sampling schedule. Intuitively, it represents the degree of isometry of the encoder $\mathbf{f}$.

4.2. Quantitative Comparison

Overall Comparison. We quantitatively evaluate the effect of our method on DDPM (Ho et al., 2020) and unconditional latent diffusion model (LDM) (Rombach et al., 2022) on various datasets. Tab. 1 indicates that the diffusion models trained with our isometric regularizer exhibit substantial improvement in PPL, implying smoother transitions during latent traversal. Smaller mRTL, MCN, and VoR also signify that the encoder of score model gets closer to scaled-isometry with our method. On CelebA-HQ, LS significantly drops, indicating improved disentanglement of the latent space.

FID and Disentanglement Trade-off. As shown in Tab. 1, applying our regularizer appears to introduce some trade-off between FID and the disentanglement metrics. However, we emphasize that low FID and nice disentanglement are two distinct desired aspects of image generation tasks, and their importance may vary depending on the user's needs. For instance, let us assume that a generator f_{gen} has learned the exact distribution of the training dataset $\mathcal{D}$, $p_{\text{data}}(y) = \frac{1}{|\mathcal{D}|}\sum_{y_i \in \mathcal{D}} \delta(y - y_i)$, where $\delta(\cdot)$ denotes Dirac-delta function. That is, $f_{\text{gen}}(x) = y_i$ if $x \in X_i$, where $\{X_i\}$ is a partition of $\text{dom}(f_{\text{gen}})$, indicating a mode-collapsed generator. In this case, it would achieve the lowest FID, but this is not a desired generative model, as it would result a maximal entanglement in the latent space. Consequently, it could be evaluated as a poor generator for downstream tasks such as inversion, image editing, and interpolation. Our proposed method provides a systematic way for the users to efficiently adjust the relevant importance of these two aspects by setting the regularization coefficient λ_{iso}, according to their needs depending on the specific target task.

Table 2. **Isometric *vs.* Path length regularizers.** Ours with the correct Riemannian metric ($\mathbf{G}$) leads to better FID and PPL.

Regularizer	G	FID-10k↓	PPL-50k↓
-	-	15.89	648
$\mathcal{L}_{\text{pl}}$ (Path length reg.)	$\mathbf{I}$	20.04	552
$\mathcal{L}_{\text{iso}}$	$\mathbf{I}$	16.60	619
$\mathcal{L}_{\text{iso}}$ (Ours)	$\mathbf{G_s}$	16.18	**455**

Additionally, Karras et al. (2020) discovers correlation between the perceived image quality and PPL metric. They explain that FID cannot fully characterize the generation quality of a generative model and demonstrate qualitative comparisons, claiming that lower PPL with the same FID relates to higher image quality. This shows achieving a low PPL is also relevant to high quality of generated images.

Comparison with Path Length Regularizer. As mentioned in Sec. 2.4, EMA training of path length regularizer $\mathcal{L}_{\text{pl}}$ can be sub-optimal, while isometric regularizer $\mathcal{L}_{\text{iso}}$ is scale-free. Indeed, from Tab. 2, we observe that using $\mathcal{L}_{\text{pl}}$ slightly improves PPL from the baseline while significantly worsens FID. On the other hand, regularizing via $\mathcal{L}_{\text{iso}}$ with $\mathbf{G_s}$, considering the accurate geometry of the latent space, significantly improves PPL while maintaining FID. Also, as seen in Tab. 3 and Fig. 6, our method demonstrates superior performance in inversion and reconstruction downstream tasks. These experiments demonstrate that our isometric regularizer makes the training more stable and easier.

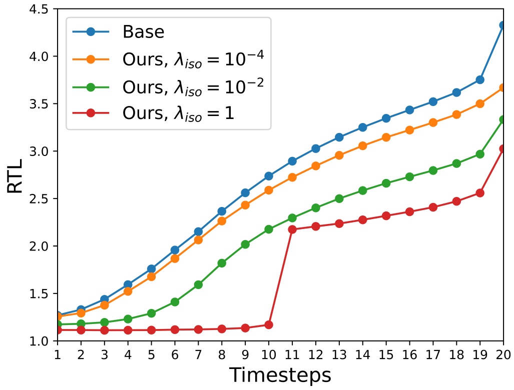


Figure 4. **RTL with various λ_{iso}.** A stronger regularization reduces the ratio to 1, flattening the trajectories in $\mathcal{H}$.

Table 3. **Quantitative comparisons of image inversion and reconstruction.** We employ DDIM inversion to convert source image to latent, and reconstruct the image with DDIM sampling. Note that low PPL relates to better inversion and reconstruction.

Regularizer	PPL-50k	MSE ↓	PSNR ↑	SSIM ↑	LPIPS ↓
-	401	0.00862	0.597	20.6	0.517
$\mathcal{L}_{\text{pl}}$ (Path length reg.)	368	0.00667	0.614	21.7	0.521
$\mathcal{L}_{\text{iso}}$ (Ours)	**340**	**0.00599**	**0.674**	**22.2**	**0.436**

Mean Relative Trajectory Length. Fig. 4 shows the measured Relative Trajectory Length (RTL) scores across the reverse timesteps in DDIM ($T = 20$). As the guidance of isometric loss gets larger with a larger λ_{iso}, the RTL tends to decrease, indicating the geodesic in $\mathcal{X}$ (Slerp) maps to geodesic in $\{\mathcal{H}_t\}$. We notice a significant drop when $t \leq 10$ especially with a larger λ_{iso}, where the isometric loss is applied. This indeed shows the isometric loss is accurately guiding the encoder of the score model to learn an isometric representation.

4.3. Analysis on the Disentanglement of Latent Space $\mathcal{X}$

We demonstrate that the disentangled latent space obtained with our method is advantageous in various downstream tasks such as interpolation, inversion, and linear editing.

Interpolation. We first conduct traversals on the latent space $\mathcal{X}$ between two points $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, illustrating the generated images from interpolated points between them in Fig. 5. We observe that with our isometric loss the latent space is better disentangled, resulting in smoother transitions without abrupt changes in gender. More examples are provided in Fig. VI in Appendix I.

Inversion and Reconstruction. In literature of GANs (Karras et al., 2020), achieving a lower PPL and consequently having a disentangled latent space is beneficial for image inversion and reconstruction. Achieving accurate inversion and reconstruction is particularly important for image editing with diffusion models because it consists of inverting the given image into a latent, and the editing happens in that latent space. Thus, we conduct similar experiments on inversion and reconstruction on diffusion, using DDIM (Song et al., 2021a) and ADM (Dhariwal & Nichol, 2021) trained on CelebA-HQ.

Tab. 3 reports the effect of our method on the image inversion and reconstruction tasks. Particularly, the PPL is a direct metric to measure disentanglement, and thus a lower PPL with our method strongly indicates better quality of image inversion. Fig. 6 qualitatively illustrates the advantage of our method in inversion and reconstruction.

Linearity. We also claim that the latent space $\mathcal{X}$ learned with our isometric loss has a property of *linearity*. Specifically, we compare the generated images with ours to baseline, where both are moved along the slerp in their latent spaces. For this, we find the editable direction following Jang et al. (2022), an unsupervised method for identifying semantic-factorizing directions in the latent space based on its local geometry, and perturb the latents through this direction both for baseline and our model. In this way, we discover the principal variations of the latent space in the neighborhood of the base latent code.

Fig. 7 demonstrates that a spherical perturbation on $\mathcal{X}$ with various intensity of $\Delta\boldsymbol{x}$ adds or removes specific attributes from the generated images accordingly. As seen in Fig. 7, the baseline often changes multiple factors (age, gender) abruptly and inconsistently with γ (*e.g.*, when $\gamma = -1$ on the right example, it suddenly shows a male-like output), while ours show disentangled changes.

Fig. 8 further illustrates the linearity of $\mathcal{X}$ with images manipulated in two directions in $\mathcal{X}$. For this, we follow Choi et al. (2022) to find the editing directions. Comparing the results of baseline and ours, we observe that our method better disentangles the concept of age and gender, successfully drawing a young male and an old female (marked with red boxes), where the baseline fails to. This indicates that the latent space trained with our approach is better disentangled, and they can be easily combined back with a linear combination.

4.4. Ablation Study

Tab. 4 shows the ablation study on the choice of optimal p and $\mathbf{G}$. We observe the best performance with $\gamma = 0.5$ and $\mathbf{G} = \mathbf{G}_s$, in FID and PPL. Note that $\gamma = 1$ denotes the original training of diffusion model. Also, using a proper Riemannian metric $\mathbf{G}$ of the latent space when calculating the isometric loss turns out to be important. This result supports our idea to model the latent space of diffusion model as a Riemannian manifold S^{n-1} is indeed reasonable.

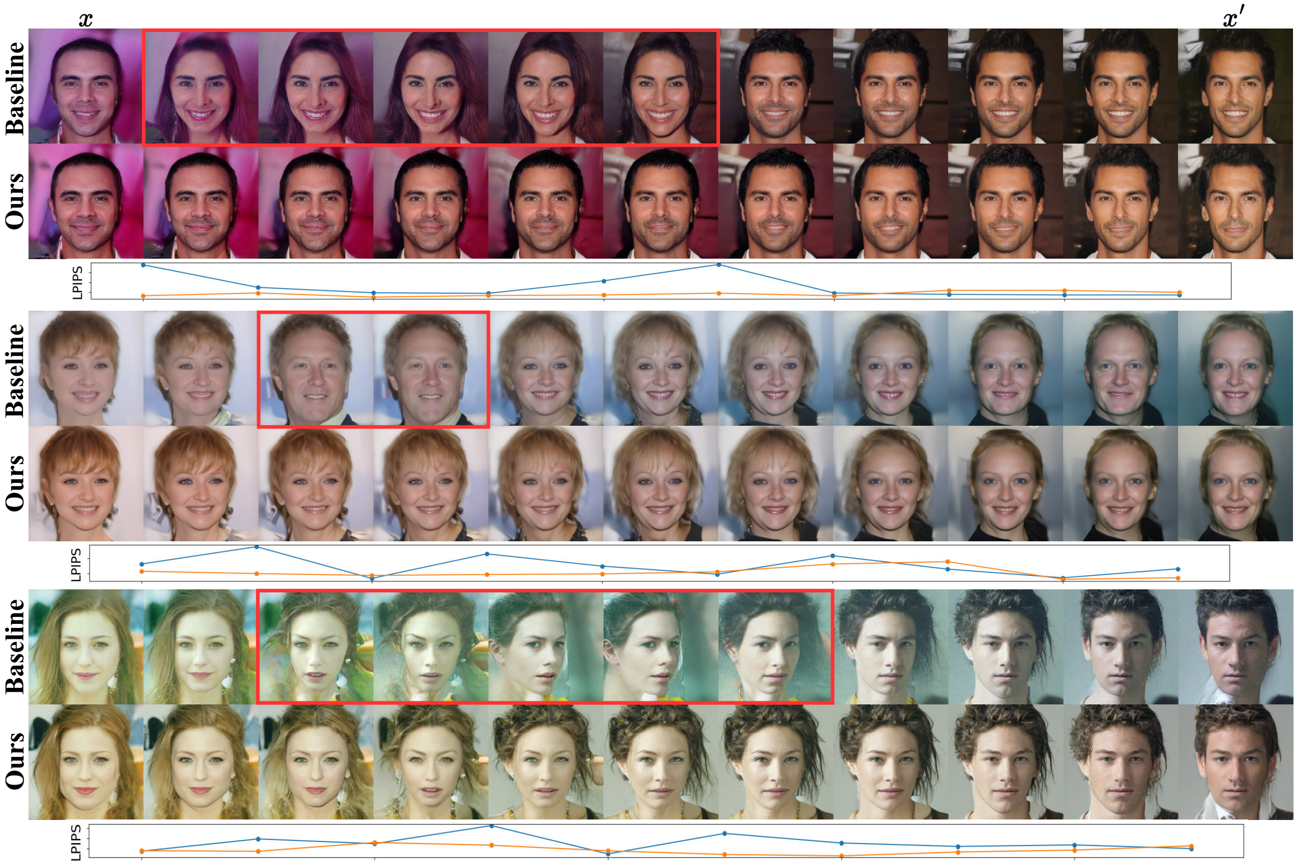


Figure 5. **Image interpolation.** Examples of latent traversal between two latents $\boldsymbol{x}$ and $\boldsymbol{x}'$ with DDPM (Ho et al., 2020), trained on 256 × 256 CelebA-HQ. We observe unnecessary changes of female → male in the baseline, while smoother transitions in ours. For quantitative support, we plot LPIPS distance between each adjacent frames (Blue: Base, Orange: Ours).

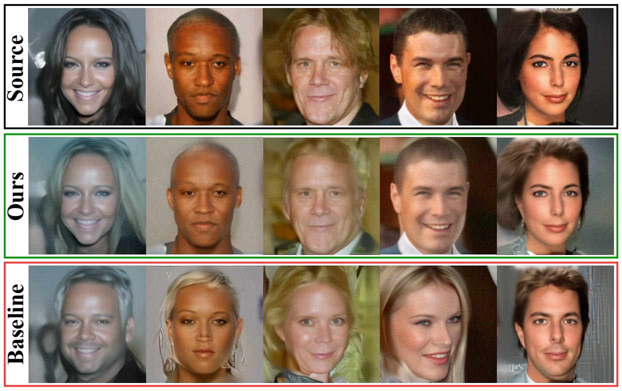


Figure 6. **Image inversion and reconstruction.** Baseline is ADM (Dhariwal & Nichol, 2021) trained on 256×256 CelebA-HQ.

5. Related Work

Latent Space of Generative Models. On Generative Adversarial Networks (GANs) (Goodfellow et al., 2014; Radford et al., 2015; Zhu et al., 2017; Choi et al., 2018; Ramesh et al., 2019; Härkönen et al., 2020; Abdal et al., 2021), StyleGAN (Karras et al., 2019) is a pioneering work on latent space analysis and improvement. In StyleGANv2 (Karras et al., 2020), a path length regularizer guides the generator to learn an isometric mapping from the latent space to the image space. Recently, additional studies on GANs (Shen et al., 2020a;b; Shen & Zhou, 2021) and VAEs (Hadjeres et al., 2017; Zheng & Sun, 2019; Zhou & Wei, 2020) have examined the latent spaces of generative models. Kwon et al. (2023) found that the internal feature space of U-Net in diffusion models, $\mathcal{H}$, plays the same role as a semantic latent space. Preechakul et al. (2022) discovered that using a semantic encoder enables the access to the semantic space of diffusion models. However, this method utilizes additional conditioning information, while our work proposes a method that can directly utilize the latent space without any condition.

Table 4. **Ablation study** on γ, the ratio of timesteps to skip applying isometric loss, and $\mathbf{G}$, the choice of Riemannian metric.

γ	$\mathbf{G}$	λ_{iso}	FID-10k ↓	PPL-50k ↓
1	-	-	15.89	653
0	$\mathbf{I}$	10^{-4}	24.07	447
0.5	$\mathbf{I}$	10^{-3}	30.28	441
0.5	$\mathbf{I}$	10^{-4}	16.60	619
0.5	$\mathbf{G}_s$	10^{-4}	16.18	455

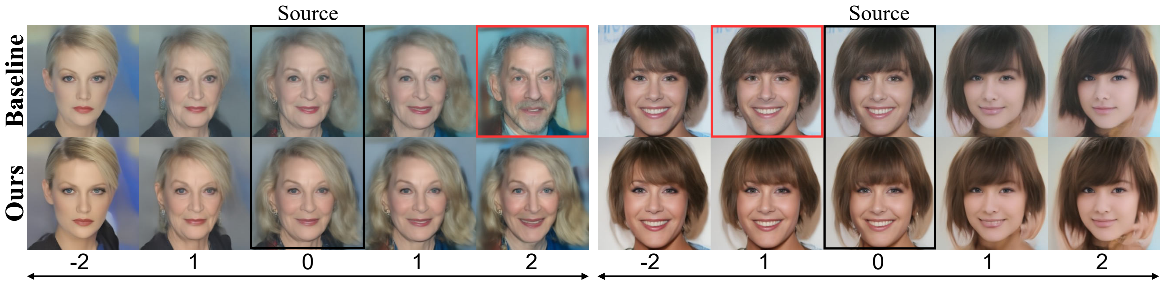
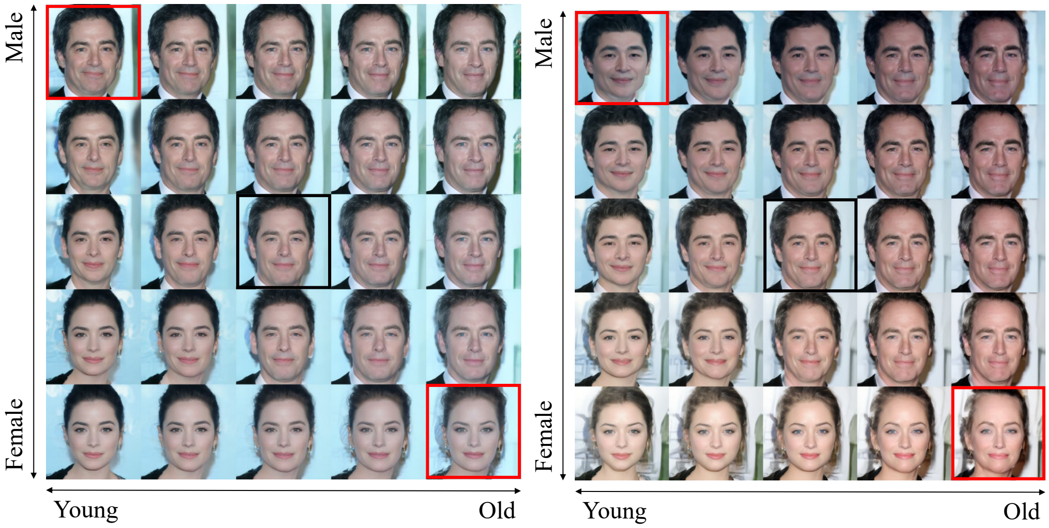


Figure 7. **Linearity.** Images generated from a source latent vector $\boldsymbol{x}$ and from slightly perturbed latents, $\boldsymbol{x} + \gamma\Delta\boldsymbol{x}$ with $\gamma \in \{-2, -1, 0, 1, 2\}$, where $\Delta\boldsymbol{x}$ corresponds to the change in age axis.



(a) DDPM (base) (b) *Isometric Diffusion* (ours)

Figure 8. **Linear combination in $\mathcal{X}$.** With ours, age and gender axes are better disentangled. Image generated with the source latent is marked with black box.

Riemannian Geometry for Generative Models. There exist some previous works on utilizing Riemannian geometry to understand the latent spaces. Chen et al. (2020) proposed that interpreting the latent space as Riemannian manifold and regularizing the Riemannian metric to be a scaled identity help VAEs learn a good latent representation. Lee et al. (2021) proposed an isometric regularization method for geometry-preserving latent space coordinates in scale-free and coordinate invariant form, arguing that an isometrically regularized autoencoder is advantageous in image retrieval task. Arvanitidis et al. (2018) claimed understanding Riemmanian geometry of the latent space and directly incorporating the pullback metric can improve analysis of representations as well as generative modeling. However, this method can be computationally heavy. Our method focuses on the reduction of computation cost at inference. See Appendix. E for further discussions.

6. Summary

In this work, we address a critical challenge in the field of generative models, particularly disentangling latent space for diffusion models. Despite the notable progress of diffusion models in generating photorealistic samples, there persists a substantial gap in comprehending and controlling their latent spaces.

Motivated from isometric representation learning, our *Isometric Diffusion* introduces a novel regularizer aimed at obtaining a more disentangled latent space for diffusion models. Through a mapping from latent space to data manifold being close to isometry, our approach demonstrates the attainment of a more intuitive and disentangled latent space for diffusion models, as evidenced both quantitatively and qualitatively. We demonstrate advantages of achieving disentangled and smoother latent space through extensive experiments of image interpolation, inversion and linear editing.

Our method will open up new possibilities for practical applications, including video generation with seamless transitional frames and easier manipulation of specific features, providing a high degree of control and customization. We believe our method can be applied to conditional generation, which will be a promising future work.

Acknowledgements

This work was supported by the New Faculty Startup Fund from Seoul National University, by Samsung Electronics Co., Ltd (IO230414-05943-01, RAJ0123ZZ-80SD), by Youlchon Foundation (Nongshim Corp.), and by National Research Foundation (NRF) grants (No. 2021H1D3A2A03038607/50%, RS-2024-00336576/10%, RS- 2023-00222663/5%) and Institute for Information & communication Technology Planning & evaluation (IITP) grants (No. RS-2024-00353131/25%, 2022-0-00264/10%), funded by the government of Korea.

Software and Data

Our source code is publicly available at `https://github.com/isno0907/isodiff`. Readers would be able to reproduce the reported results by running this code. We describe the detailed experimental settings including hyperparameters and hardware environments we use in Sec. 4.1 and 4.4.

Impact Statement

This paper proposes a method to enhance the underlying latent space of diffusion models to ease the image or video editing, selectively adjusting certain aspects of them as intended. Our work shares ethical issues of generative models that are currently known in research community; to name some, deep fake, fake news, malicious editing to manipulate evidence, and so on. We believe our work does not significantly worsen these concerns in general, but a better disentangled latent semantic space with our approach might ease these abuse cases as well. Also, other relevant ethical issues regarding potential discrimination caused by a biased dataset still remain the same with our approach, neither improving nor worsening ethical concerns in this aspect. A collective effort within the entire research community and society will be important to keep generative models beneficial.

References

Abdal, R., Zhu, P., Mitra, N. J., and Wonka, P. StyleFlow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (ToG)*, 40(3):1–21, 2021.

Apostol, T. *Mathematical Analysis*. Addison-Wesley series in mathematics. Addison-Wesley, 1974. ISBN 9780201002881.

Arvanitidis, G., Hansen, L. K., and Hauberg, S. Latent space oddity: on the curvature of deep generative models. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2018.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828, 2013.

Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S. W., Fidler, S., and Kreis, K. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Brown, R. G. Exponential smoothing for predicting demand, 1956.

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., and Lerchner, A. Understanding disentangling in β-VAE, 2017.

Chen, N., Klushyn, A., Ferroni, F., Bayer, J., and van der Smagt, P. Learning flat latent manifolds with VAEs. In *Proc. of the International Conference on Machine Learning (ICML)*, 2020.

Chen, R. T., Li, X., Grosse, R. B., and Duvenaud, D. K. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Cho, Y., Yi, S., Kim, S. K., Yoon, H., and Lee, J. Hybrid diffusions for stable molecular structure generation via explicit energy-based model. In *Proc. of the International Conference on Machine Learning (ICML)*, 2023.

Choi, J., Lee, J., Yoon, C., Park, J. H., Hwang, G., and Kang, M. Do not escape from the manifold: Discovering the local coordinates on the latent space of gans. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2022.

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., and Choo, J. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

do Carmo, M. *Riemannian Geometry*. Mathematics (Birkhäuser) theory. Birkhäuser Boston, 1992. ISBN 9780817634902.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

Hadjeres, G., Nielsen, F., and Pachet, F. GLSR-VAE: Geodesic latent space regularization for variational autoencoder architectures. In *IEEE symposium series on computational intelligence (SSCI)*, 2017.

Härkönen, E., Hertzmann, A., Lehtinen, J., and Paris, S. GANSpace: Discovering interpretable GAN controls. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., and Cohen-Or, D. Prompt-to-prompt image editing with cross attention control. *arXiv:2208.01626*, 2022.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in Neural Information Processing Systems (NIPS)*, 30, 2017.

Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., Botvinick, M. M., Mohamed, S., and Lerchner, A. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2017.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J., et al. Imagen video: High definition video generation with diffusion models. *arXiv:2210.02303*, 2022.

Huang, H., He, R., Sun, Z., Tan, T., et al. IntroVAE: Introspective variational autoencoders for photographic image synthesis. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018.

Hutchinson, M. F. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 18(3):1059–1076, 1989.

Jang, C., Lee, Y., Noh, Y.-K., and Park, F. C. Geometrically regularized autoencoders for non-euclidean data. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2022.

Jeong, J., Kwon, M., and Uh, Y. Training-free style transfer emerges from h-space in diffusion models. In *Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. Analyzing and improving the image quality of styleGAN. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., and Irani, M. Imagic: Text-based real image editing with diffusion models. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Kim, H. and Mnih, A. Disentangling by factorising. In *Proc. of the International Conference on Machine Learning (ICML)*, 2018.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2013.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2012.

Kwon, M., Jeong, J., and Uh, Y. Diffusion models already have a semantic latent space. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2023.

Lee, S. and Lee, J. PoseDiff: Pose-conditioned multimodal diffusion model for unbounded scene synthesis from sparse inputs. In *Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

Lee, Y., Yoon, S., Son, M., and Park, F. C. Regularized autoencoders for isometric representation learning. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2021.

Miranda, R. *Algebraic curves and Riemann surfaces*, volume 5. American Mathematical Soc., 1995.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *Proc. of the International Conference on Machine Learning (ICML)*, 2022.

Park, Y.-H., Kwon, M., Choi, J., Jo, J., and Uh, Y. Understanding the latent space of diffusion models through the lens of riemannian geometry. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *Proc. of the IEEE/CVF Conference on International Conference on Computer Vision (ICCV)*, 2023.

Preechakul, K., Chatthee, N., Wizadwongsa, S., and Suwajanakorn, S. Diffusion AutoEncoders: Toward a meaningful and decodable representation. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv:1511.06434*, 2015.

Ramesh, A., Choi, Y., and LeCun, Y. A spectral regularizer for unsupervised disentanglement. In *Proc. of the International Conference on Machine Learning (ICML)*, 2019.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with CLIP latents. *arXiv:2204.06125*, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Ronneberger, O., Fischer, P., and Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In *Proc. of the Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Shen, Y. and Zhou, B. Closed-form factorization of latent semantics in GANs. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Shen, Y., Gu, J., Tang, X., and Zhou, B. Interpreting the latent space of GANs for semantic face editing. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020a.

Shen, Y., Yang, C., Tang, X., and Zhou, B. InterfaceGAN: Interpreting the disentangled face representation learned by gans. *IEEE transactions on pattern analysis and machine intelligence*, 44(4):2004–2018, 2020b.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proc. of the International Conference on Machine Learning (ICML)*, 2015.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2021a.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2021b.

Vincent, P. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

Wang, L., Guo, S., Huang, W., Xiong, Y., and Qiao, Y. Knowledge guided disambiguation for large-scale scene classification with multi-resolution CNNs. *IEEE Transactions on Image Processing*, 26(4):2055–2068, 2017.

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. Aggregated residual transformations for deep neural networks. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Zheng, Z. and Sun, L. Disentangling latent space for vae by label relevant/irrelevant dimensions. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Zhou, D. and Wei, X.-X. Learning identifiable and interpretable latent models of high-dimensional neural activity using pi-VAE. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. of the IEEE/CVF Conference on International Conference on Computer Vision (ICCV)*, 2017.

A. Implementation Details

Our network architecture follows the backbone of DDPM (Ho et al., 2020), which uses a U-Net (Ronneberger et al., 2015) internally. If not specified, we train with batch size 32, learning rate 10^{-4}, $p = 0.5$, and $\lambda_{\text{iso}} = 10^{-4}$ for 10 epochs by default.

For all datasets and models, we initialize with pre-trained weights and further fine-tune them with each competing method until the lowest FID is achieved. All the scores reported in Tab. 1 have been achieved before 5000 iterations. We experimentally confirm that the results of training from the scratch and fine-tuned are almost identical.

We use Adam optimizer and exponential moving average (Brown, 1956) on model parameters with a decay factor of 0.9999. We use 4 NVIDIA A100 GPUs with 40GB memory for experiments.

B. Details on Evaluation Metrics

In this section, we provide further details of the evaluation metrics we use throughout this paper.

Linear separability (LS) (Karras et al., 2019) measures the degree of disentanglement of a latent space. Karras et al. (2019) argues that if a latent space is disentangled, it should be able to find a consistent direction that changes an image attribute independently, and thus the latent space labeled according to the specific attribute should be separable by a hyperplane. The formal definition of this metric is as follows:

$$\text{LS} = e^{\sum_i H(Y_i|X_i)}, \tag{9}$$

where i is the attribute index, $H(\cdot|\cdot)$ is conditional entropy, X are the classes predicted by SVM, and Y are the classes predicted by a pre-trained classifier. Intuitively, it measures how much additional information is needed to fully determine the label determined by the classifier, knowing the label predicted by SVM, hence indicating how much the latent space is separable by a hyperplane.

We train a classifier with ResNeXt (Xie et al., 2017) to predict the 40 attribute confidence scores with CelebA annotated for each image, and then follow the method in (Karras et al., 2019). We calculate it with SVMs using linear kernel and radial basis function kernel, regarding the spherical geometry of the latent space. We compute it with 1,000 images pruned after sorting with classifier confidence scores, from 2,000 images generated.

Mean condition number (MCN) and *variance of Riemannian metric* (VoR) are the metrics measuring how much a mapping is close to a scaled-isometry, proposed by (Lee et al., 2021). We measure MCN and VoR of the score models' encoders to measure how much our isometric regularizer has successfully guided the encoder to be isometric. Formally, the mean condition number (MCN) is defined as

$$\text{MCN} = \mathbb{E}_{\boldsymbol{x}_0}\mathbb{E}_{\boldsymbol{x}_t \sim p(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[\frac{\sigma_M(\boldsymbol{J}(\boldsymbol{x}_t))}{\sigma_m(\boldsymbol{J}(\boldsymbol{x}_t))}\right], \tag{10}$$

where σ_M, σ_m are the maximum and minimum singular values. MCN measures how isotropic the Riemannian metric is. Note that $\sigma_i(\boldsymbol{J}(\boldsymbol{x}_t)) = \lambda_i^2(\boldsymbol{J}^\top(\boldsymbol{x}_t)\boldsymbol{J}(\boldsymbol{x}_t))$, where λ_i is the i-th eigenvalue. The variance of Riemannian metric (VoR) is defined as

$$\text{VoR} = \sum_i \text{Var}_{\boldsymbol{x}_0, \boldsymbol{x}_t \sim p(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[\sigma_i(\boldsymbol{J}(\boldsymbol{x}_t))\right], \tag{11}$$

where we measure how homogeneous Riemannian metric is. Note that we slightly modify its definition to bypass the exact calculation of Jacobian by exploiting SVD. Satisfying both isotropicity and homogeneity of Riemannian metric, a mapping can be determined its proximity to isometry. We measure them with 1,000 images.

C. Advantages of Disentangled Latent Space

While there exists some topological discrepancy between Gaussian prior and the true image distribution, generative modeling have often modeled their latent spaces as Gaussian (*e.g.*, GANs, VAEs) and there have been studies on the advantages of geometric regularizing in learning a 'better' latent space modeled as Gaussian, even though the target distribution will be quite different from it. We believe that such geodesic preserving property is motivated from various literatures in generative models.

For example, StyleGAN2 (Karras et al., 2020) uses path length regularizer to guide the generator to become closer to isometry and achieves a smoother latent space. Their work shows that the path-length-regularized StyleGAN2 improves 1) to lower PPL (a consistency and stability metric in image generation), and 2) to have invertibility from image to its latent codes. We believe the latter is potentially related to the existence of smooth inverse function of the generator, which is an important feature for image manipulation. In diffusion models, this corresponds to DDIM inversion (Dhariwal & Nichol, 2021), and we believe our method can improve the inversion quality in diffusion models and hence contribute to high quality latent manipulations, with similar effects with that of path length regularized StyleGAN2.

Additionally, FMVAE (Chen et al., 2020) uses isometric regularizer to the decoder of VAE to learn a mapping from Gaussian latent space to image space close to isometry, obtaining advantages in downstream tasks using geometrically aligned latent space. As also illustrated in Karras et al. (2020) and Chen et al. (2020), we admit that it somehow penalizes the FID score, possibly due to the nature of regularizer. We leave the exploration of minimizing the tradeoff as a promising future work.

Also, disentangled latent space leads to improvement in image editing capabilities. First of all, disentangled representations make image editing more effective and intuitive, since it becomes easier to manipulate specific attributes of an image without affecting others when the underlying key factors are disentangled. For example, if a model has disentangled representations for pose and identity in images of faces, one could edit the pose of a face without altering its identity, or vice versa. We demonstrate in Fig. 6 that the advantages of disentangled latent space in the inversion and reconstruction task, which is particularly important for image editing with diffusion models. This is because image editing consists of inverting the given image into a latent, and the editing happens in that latent space.

D. On the Scalability of the Proposed Method

As discussed in Park et al. (2023), the complexity of $\mathcal{H}$ increases as the complexity of the training dataset increases. The work also explicitly reports the entanglement phenomena empirically discovered in Stable Diffusion (Rombach et al., 2022), marking as its limitation. While intervention of large-scale training data, latent encoder/decoder, and text encoder in latent diffusion models (LDM) or Stable Diffusion complicates the relation between the noise space ($\mathcal{X}$) and the semantic space ($\mathcal{H}$), Jeong et al. (2024) demonstrates the efficacy of $\mathcal{H}$ space also in Stable Diffusion in a text-conditioned setting, hence validating the method also in large-scale setting.

Therefore, we believe the method can be scaled up, and also can incorporate conditional models including text-to-image models such as Stable Diffusion, which can be an interesting direction for future work. As long as the $\mathcal{H}$ space is effective, our approach can be easily adopted to further regularize it with minimal additional cost.

E. On the Challenges of Directly Applying Pullback Metric to Diffusion Models

Under the setting of using VAE in Arvanitidis et al. (2018), pulling back the metric of the observed space could be straightforward, since the generator is explicitly defined with VAE. However, since the generative process of diffusion model is iterative, directly translating this method to diffusion models can be infeasible. Specifically, pulling back the metric of the observed space requires calculating the Riemmanian metric $\mathbf{G} = \frac{\partial f}{\partial \boldsymbol{x}}^{\top} \frac{\partial f}{\partial \boldsymbol{x}}$ for every point on the interested trajectory. This requires full calculation of Jacobian of $f = f_0 \circ \cdots \circ f_{N-1}$, where N is the number of reverse steps (*e.g.*, $N = 100$ in DDIM) and f_i is the i-th reverse step, resulting in a long chain of function compositions. This could be computationally expensive for heavy models such as high resolution diffusion models.

Also, in order to obtain geodesic, one needs to numerically solve a corresponding ODE or to directly optimize discretized trajectory, and this additional step also can be computationally expensive. Our method proposes to transfer this computation from inference time to training time, and this is beneficial in a sense that inference can be done many times while training will be done only once.

Furthermore, assuming calculation of the pull back metric in the diffusion model is feasible, directly utilizing the pullback metric and our method are not conflicting but complementary to each other. Our approach improves the latent space but can take further benefit by direct methods like Arvanitidis et al. (2018), by obtaining exact geodesics and fully reflecting the geometry of observed space to the latent space.

F. Stochastic Trace Estimator

F.1. Estimation Accuracy

In Eq. (7) of the main text, we explained that the second quality holds because of the stochastic trace estimator (Hutchinson, 1989) which is an algorithm to obtain such an estimate from matrix-vector products:

$$\mathrm{Tr}(A) = \mathbb{E}[\boldsymbol{v}^\top A \boldsymbol{v}] \simeq \frac{1}{N} \sum_{i=1}^{N} v_i^T A v_i, \tag{12}$$

where A is any square matrix and $\boldsymbol{v}$ is random vector such that $\mathbb{E}[\boldsymbol{v}\boldsymbol{v}^\top] = \boldsymbol{I}$.

As shown in Fig. I the error of stochastic trace estimator increases as the number of sample N. In this experiment, A follows $\mathcal{N}(0, \boldsymbol{I}) \in \mathbb{R}^{256\times256}$ and $\boldsymbol{v}$ follows $\mathcal{N}(0, \boldsymbol{I}) \in \mathbb{R}^{256\times1}$.

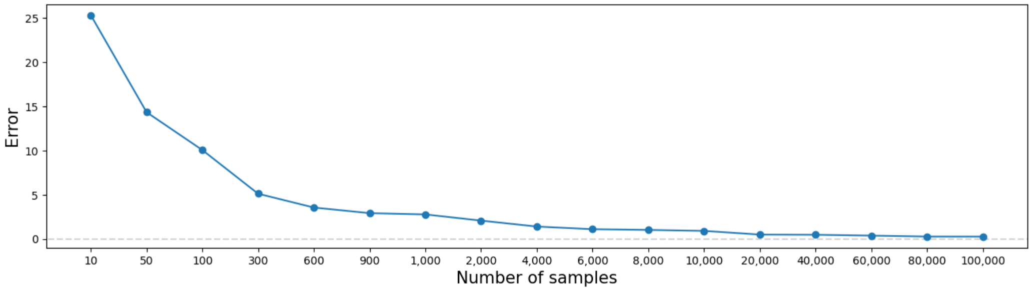
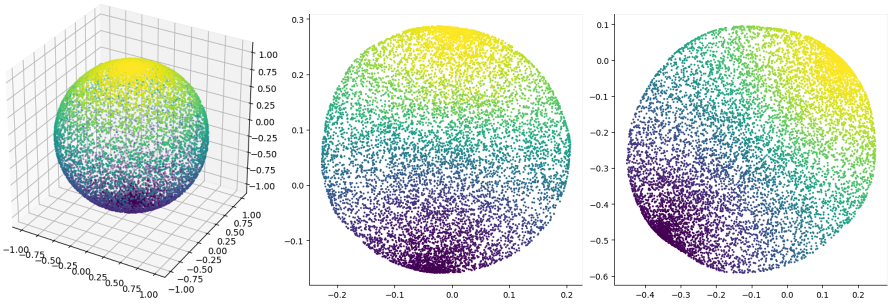


Figure I. **Approximation error of stochastic trace estimator against the number of samples.** Each point on the graph represents the error corresponding to a particular sample size.

Despite the inherent errors of estimator, we conduct a simple experiment in the setting similar to Fig. 3 to investigate whether optimizing with estimated trace converges similar to optimizing with exact trace. As shown in Fig. II, optimizing the model by approximating the trace of the matrix with the stochastic trace estimator yields similar results to those obtained by using the actual trace of the matrix. Furthermore, Fig. I demonstrates that the approximated trace exhibits a similar convergence pattern in loss over training time. These results suggest that the final convergence point is similar even when the loss function is optimized by estimating the trace of the matrix through stochastic trace estimator.

F.2. Computational Comparison

Given that $\mathcal{X} \subset \mathbb{R}^{256\times256\times3}$ and $\mathcal{H} \subset \mathbb{R}^{8\times8\times512}$, the encoder's Jacobian $\boldsymbol{J}$ contains 6,442,450,944 elements. With `float32` data type, the Jacobian matrix uses approximately 24 GB of memory. The computation time for a single Jacobian takes 202.77 seconds under our environment using NVIDIA A100 40GB.



(a) S^2 manifold (b) $\mathcal{L}_{\text{iso}}$ with estimator (c) $\mathcal{L}_{\text{iso}}$ with exact trace

Figure II. (a) Illustration of the input S^2 manifold. (b) latent coordinates learned with isometric regularizer, estimated with the stochastic trace estimator. (c) latent coordinates learned with exact isometric regularizer.

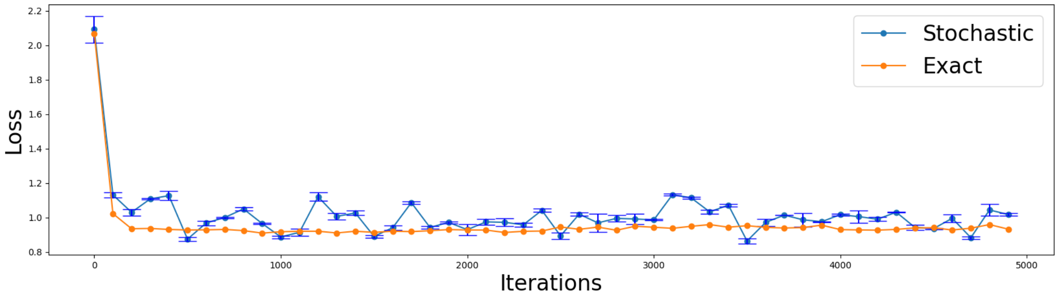
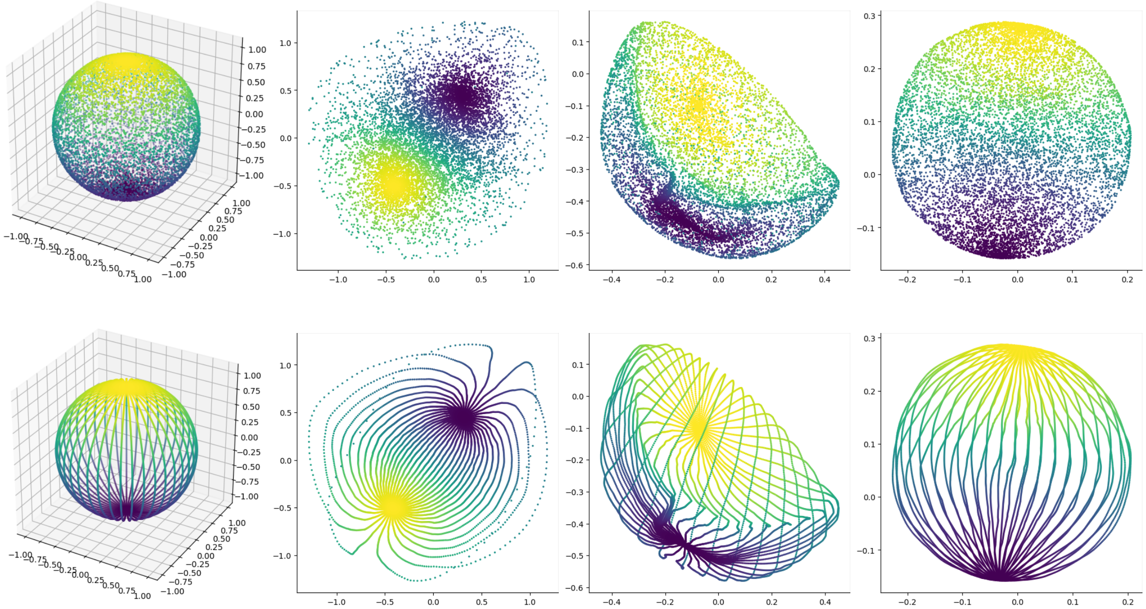


Figure III. **Loss plot during training of the toy model.** Loss calculated with trace estimator successfully converges compared to that calculated with the exact trace value. Loss calculated with trace estimator was repeated 5 times.

In contrast, the Jacobian Vector Product (JVP) does not explicitly calculate the entire Jacobian matrix, but it directly computes the product of the Jacobian matrix with a specific vector, requiring only $(256 \times 256 \times 3 + 8 \times 8 \times 512) \times 4$ = 91,750 bytes, which is approximately 0.875MB of memory. In our isometry loss, we utilize three times of JVPs for estimating the trace of a Jacobian. The computation time for a single JVP takes 0.6 seconds under our environment.

G. Illustration of the Isometric Loss

In Fig. IV, we provide more illustrations of the latent space of an autoencoder, regularized with isometric loss.



(a) S^2 manifold (b) Reconstruction only (c) $\mathcal{L}_{\text{iso}}$ with $\mathbf{G} = \boldsymbol{I}$ (Euclidean) (d) $\mathcal{L}_{\text{iso}}$ with full Eq. (7)

Figure IV. (a) Illustration of the input S^2 manifold. (b–d) Mapped contours in latent coordinates learned by an autoencoder; (b) with reconstruction loss only, (c) with isometric loss assuming naive Euclidean geometry, and (d) with our isometric loss considering S^2 geometry.

H. Preservation of $\mathcal{H}$ after Isometric Training

Trained with our isometric loss acting as a regularizer to the denoising score matching loss, it is not trivial if the model eventually learns the semantic space in $\mathcal{H}$. However, Kwon et al. (2023) argues that $\mathcal{H}$ exists in the bottleneck layer of the

U-Net, for all pretrained diffusion models. Hence, it is reasonable to deduce that $\mathcal{H}$ space exists given that the denoising score matching (DSM) loss has converged. Therefore, it can be inferred that $\mathcal{H}$-space exists if the DSM loss converges to a similar point, even when the isometric loss is added.

We observe that the addition of the isometry loss does not significantly alter the convergence point of the diffusion loss and still shows comparable FID scores. From this, we can naturally conclude that $\mathcal{H}$-space also still exists in our model.

As empirical evidence, we provide some qualitative results of image editing with the $\mathcal{H}$ in Fig. V. We aim to edit the image $\boldsymbol{x}_0$ to the direction toward $\boldsymbol{x}'_0$, manipulating only the content of the image while preserving the person's identity. Specifically, we first calculate features $\{\boldsymbol{h}_t\}$ and $\{\boldsymbol{h}'_t\}$ corresponding to $\boldsymbol{x}_t$ and $\boldsymbol{x}'_t$, respectively, where t is the DDIM time steps. Then, we use $\{\boldsymbol{h}_t + 1.5\boldsymbol{h}'_t\}$ to inject contents during the reverse process starting from $\boldsymbol{x}_T$, following Jeong et al. (2024). Note that the leftmost image for each row is $\boldsymbol{x}_0$, and other images in the same row are the edited ones.

I. Latent Traversal Examples

We provide additional examples to compare the latent traversals with the baseline (DDPM) and with our model trained with isometric loss, trained on CelebA-HQ, LSUN-Bedroom, and LSUN-Church datasets. The image resolution is 256×256 for all datasets. Fig. VI–VIII extend Fig. 5 with more examples.

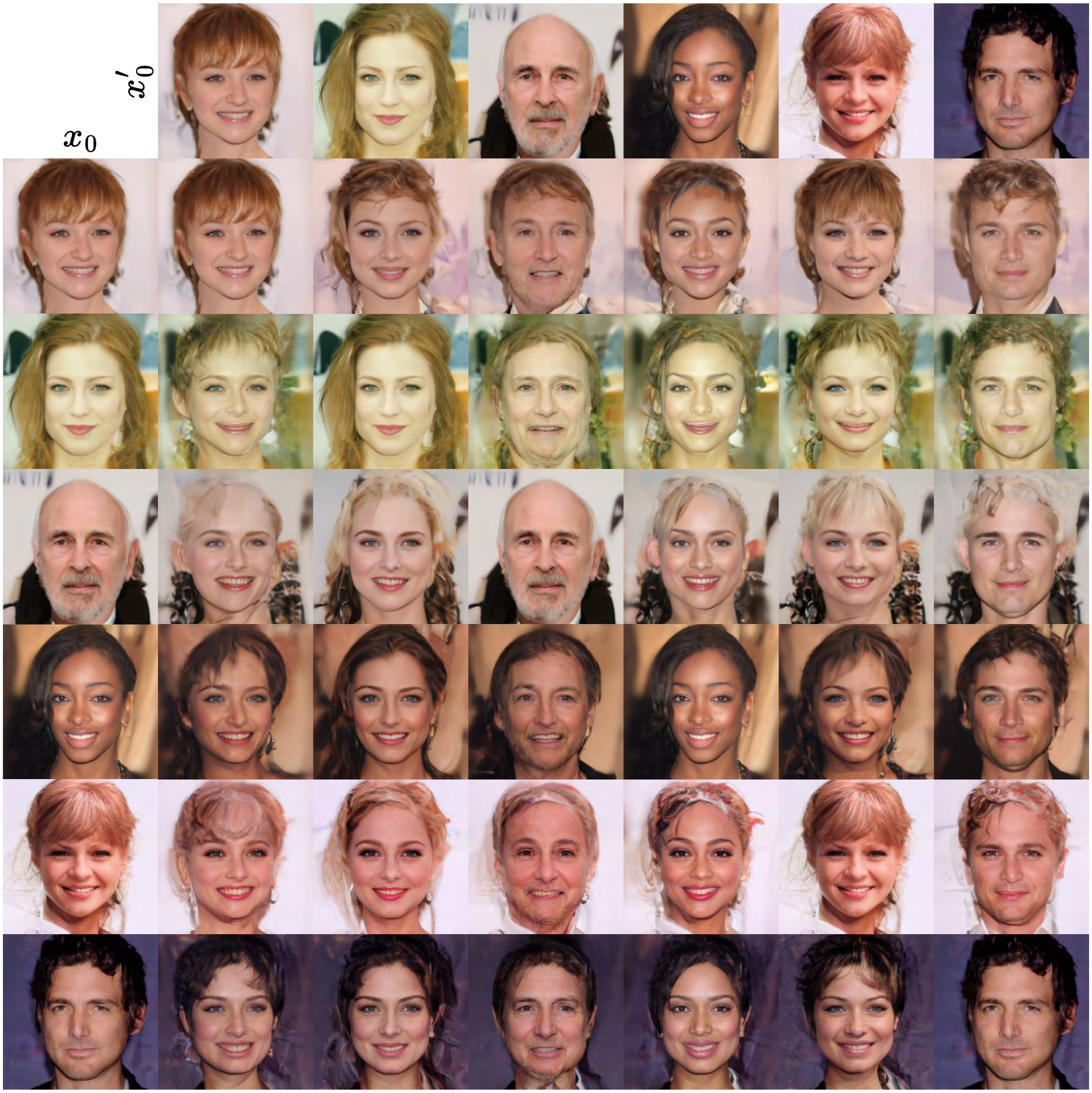


Figure V. **Empirical observation regarding existence of $\mathcal{H}$ in our model.** Images in the same row share the original image $\boldsymbol{x}_0$, images in the same column share the source image $\boldsymbol{x}_0'$ for editing direction $\{h_t'\}$.

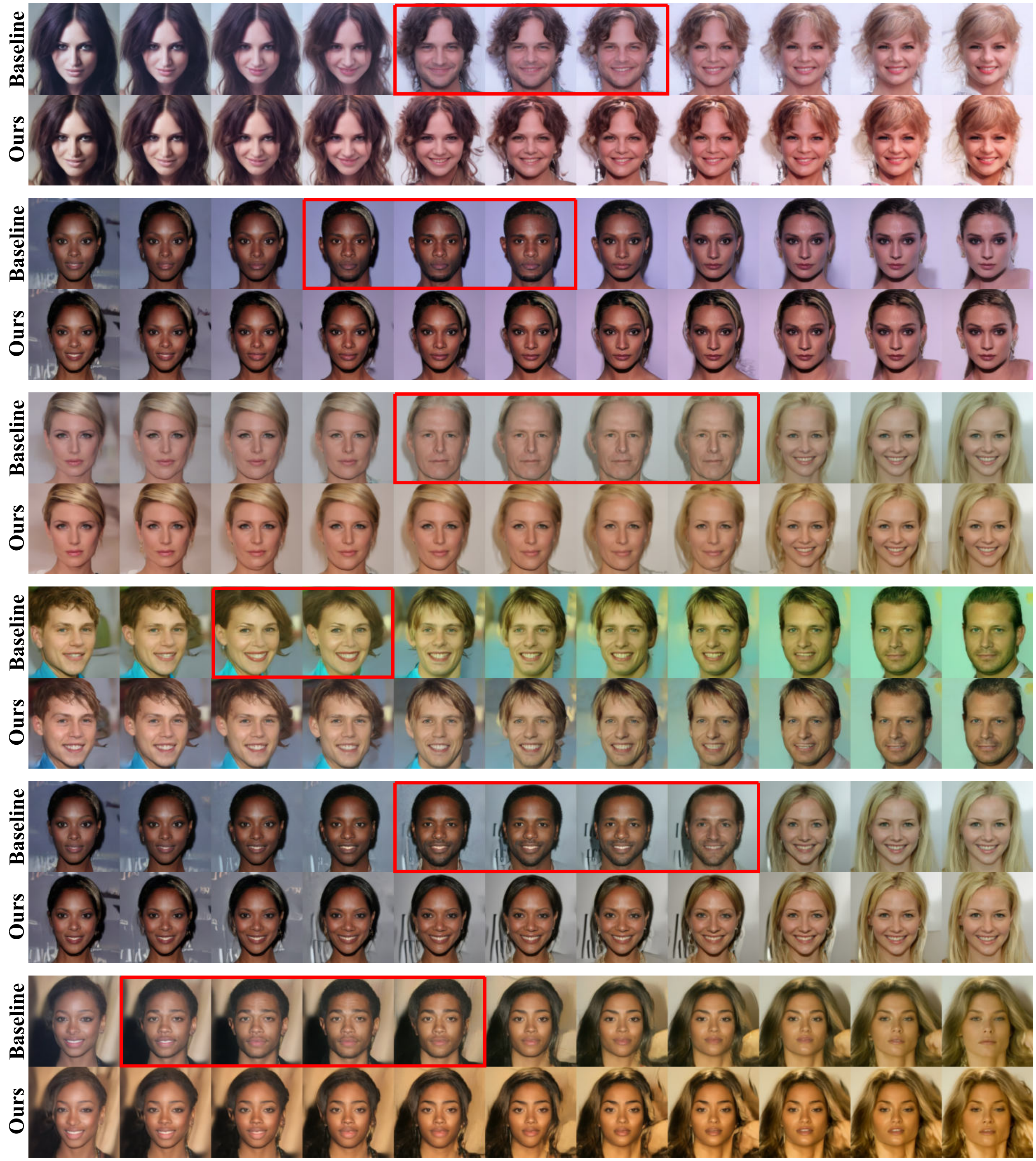


Figure VI. **Additional examples of latent traversal** between two images with DDPM and ours trained with isometric regularizer, trained on 256 × 256 CelebA-HQ.

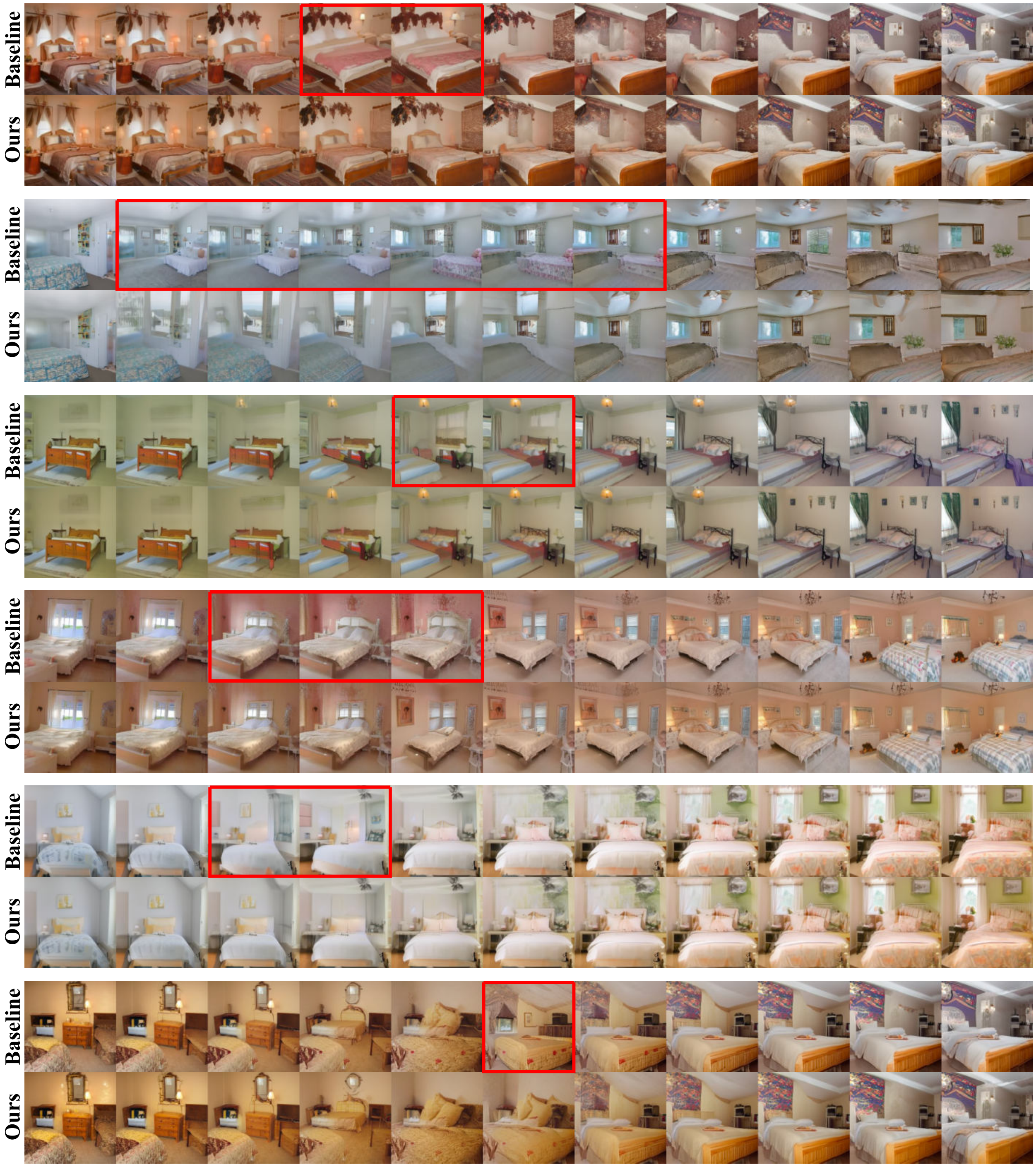


Figure VII. **Additional examples of latent traversal** between two images with DDPM and ours trained with isometric regularizer, trained on 256 × 256 LSUN-Bedroom.

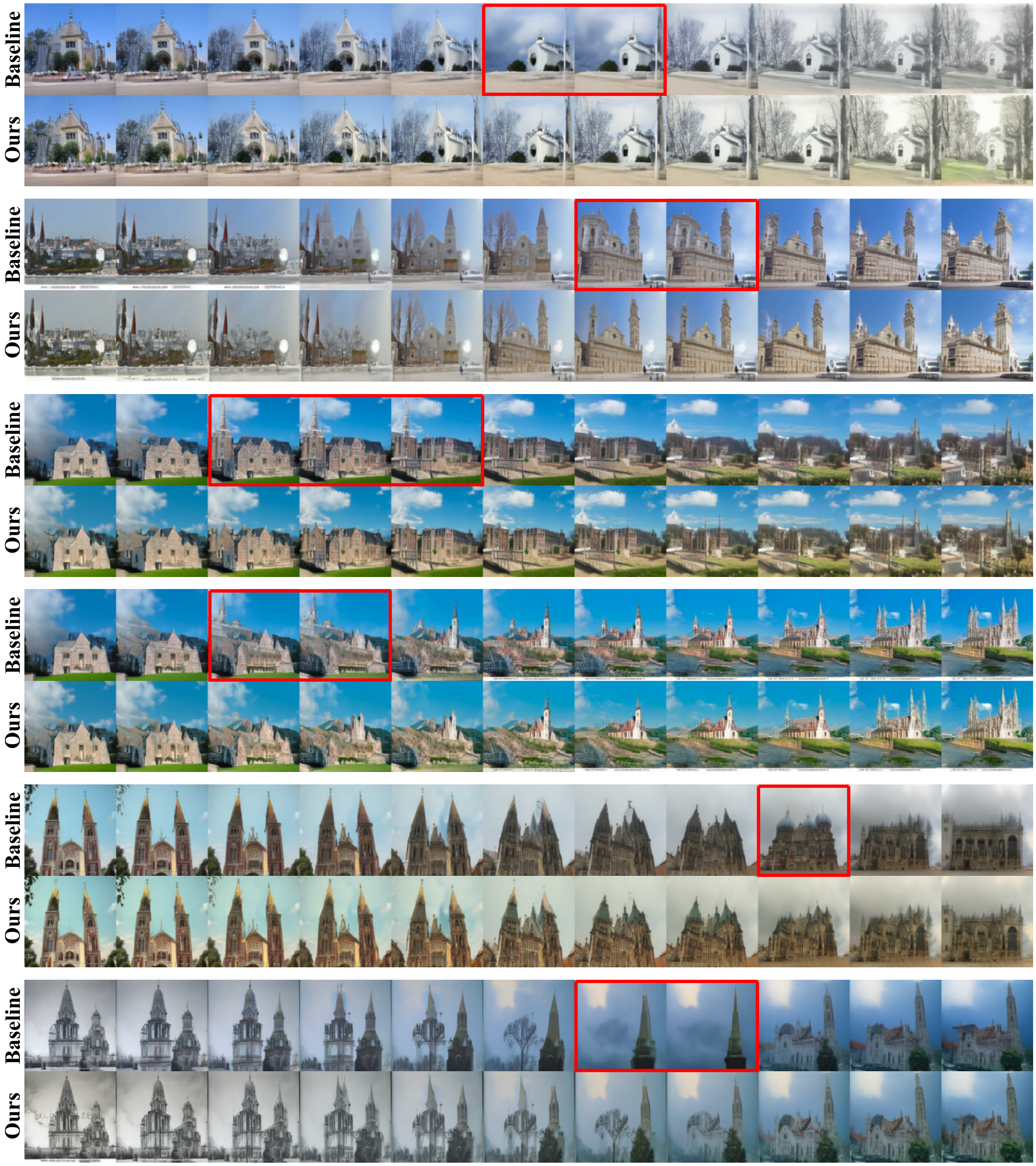


Figure VIII. **Additional examples of latent traversal** between two images with DDPM and ours trained with isometric regularizer, trained on 256 × 256 LSUN-Church.