Unsupervised Discovery of Semantic Latent Directions in Diffusion Models

Yong-Hyun Park [*1] **Mingi Kwon** [*2] **Junghyo Jo** [1] **Youngjung Uh** [2]

Abstract

Despite the success of diffusion models (DMs), we still lack a thorough understanding of their latent space. While image editing with GANs builds upon latent space, DMs rely on editing the conditions such as text prompts. We present an unsupervised method to discover interpretable editing directions for the latent variables $\mathbf{x}_t \in \mathcal{X}$ of DMs. Our method adopts Riemannian geometry between $\mathcal{X}$ and the intermediate feature maps $\mathcal{H}$ of the U-Nets to provide a deep understanding over the geometrical structure of $\mathcal{X}$. The discovered semantic latent directions mostly yield disentangled attribute changes, and they are globally consistent across different samples. Furthermore, editing in earlier timesteps edits coarse attributes, while ones in later timesteps focus on high-frequency details. We define the curvedness of a line segment between samples to show that $\mathcal{X}$ is a curved manifold. Experiments on different baselines and datasets demonstrate the effectiveness of our method even on Stable Diffusion. Our source code will be publicly available for the future researchers.

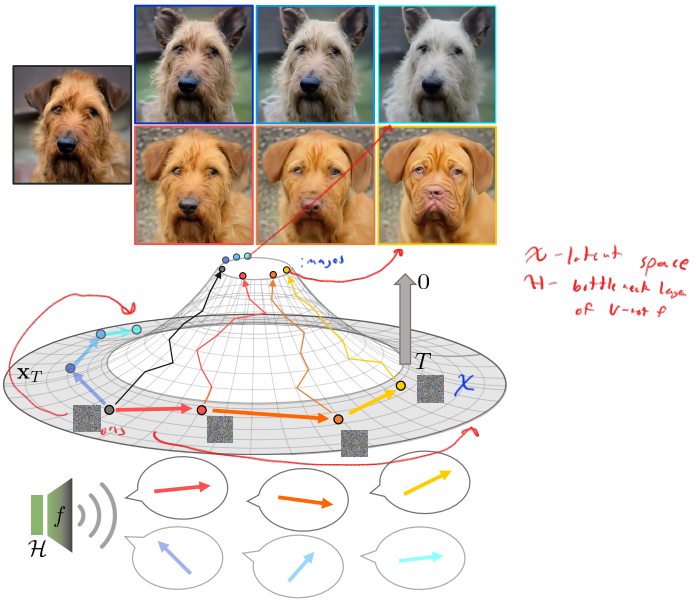


Figure 1. **Conceptual illustration of our method.** We find semantic directions in the latent space $\mathbf{x}_t$ in an unsupervised fashion relying on Riemannian geometry between $\mathbf{x}_t$ and $\mathbf{h}_t$. $\mathcal{H}$ denotes the bottleneck layer and f indicates the frozen encoder of a U-Net. The found directions manipulate the semantics of the resulting images.

1. Introduction

Diffusion models (DMs) are highly powerful generative models that have shown great performance (Ho et al., 2020; Song et al., 2020a;b; Dhariwal & Nichol, 2021; Nichol & Dhariwal, 2021). To control the generative process, existing methods have introduced conditional DMs, especially for text-to-image synthesis (Ramesh et al., 2022; Rombach et al., 2022; Balaji et al., 2022; Nichol et al., 2021), or mixing the latent variables $\mathbf{x}_t$ of different sampling processes (Choi et al., 2021a; Meng et al., 2021; Avrahami et al., 2022b; Liew et al., 2022; Kawar et al., 2022; Avrahami et al., 2022a).

Despite their success, the research community still lacks a clear understanding of what the latent variables or intermediate features of the models are embedded or how they are reflected in the resulting images. We attribute it to the characteristic iterative process of the DMs which involves a sequence of noisy images and subtle noises, i.e., the embeddings are not directly connected to the final images. In contrast, arithmetic operations in the latent space of generative adversarial networks (GANs) lead to semantic changes in the resulting images (Goodfellow et al., 2020). This property has been one of the key factors in developing GANs for real-world applications. We suppose that a better understanding of the latent space of DMs will boost similar development.

Kwon et al. (2022) adopt the intermediate feature space of the diffusion kernel as a semantic latent space, namely $\mathcal{H}$, paired with a designated asymmetric sampling process.

[*]Equal contribution [1]Department of Physics Education, Seoul National University, Seoul, Korea [2]Department of Artificial Intelligence, Yonsei University, Seoul, Korea. Authors: Yong-Hyun Park <enkeejunior1@snu.ac.kr>, Mingi Kwon <kwonmingi@yonsei.ac.kr>. Correspondence to: Youngjung Uh <yj.uh@yonsei.ac.kr>, Junghyo Jo <jojunghyo@snu.ac.kr>.

They revealed the local linearity of $\mathcal{H}$, adding to our understanding of the latent space of DMs. However, they do not directly deal with the latent variables $\mathbf{x}_t$ but rely only on a proxy, $\mathbf{h}$. Furthermore, they require external supervision such as Contrastive Language-Image Pretraining (CLIP) to find editable directions. (Radford et al., 2021)

In this paper, we introduce useful intuitions about latent space $\mathcal{X}$ to deepen our understanding of how we can control *pretrained and frozen* diffusion models. First, we identify semantic latent directions in $\mathcal{X}$ which manipulate the resulting images using Riemannian geometry in an unsupervised manner. The directions come from the singular value decomposition of the Jacobian of the mapping from $\mathcal{X}$ to $\mathcal{H}$, the intermediate feature space of the model. Figure 1 illustrates the main concept of our method.

Second, we find global semantic directions by exploiting the homogeneity of $\mathcal{H}$. It removes cumbersome per-sample Jacobian computation and allows general controllability. It follows the course of generative adversarial networks: extending per-sample editing directions (Ramesh et al., 2018; Patashnik et al., 2021; Abdal et al., 2021; Shen & Zhou, 2021) to global editing directions (Härkönen et al., 2020; Shen & Zhou, 2021; Yüksel et al., 2021).

Last but not least, we show interesting properties of the diffusion models. Spherical linear interpolation in $\mathcal{X}$ leads to smooth interpolation between samples because it is approximately geodesic in $\mathcal{H}$. That is, $\mathcal{X}$ is a warped space. The early timesteps generate low frequency components and the later timesteps generate high frequency components. Although it is indirectly shown in existing works Choi et al. (2022), we explicitly reveal it via power spectral density.

In the experiments, we demonstrate that the directions found in an unsupervised manner indeed lead to semantic changes in the images. We note that discovering the editing directions in the latent variables of diffusion models has not been tackled. Furthermore, we provide thorough quantitative and qualitative analyses on the aforementioned properties. Our method even works on stable diffusion (Rombach et al., 2022).

2. Related Works

Recent advances in DMs have resulted in the development of a universal approach known as DDPMs (Ho et al., 2020). Song et al. (2020b) have facilitated the unification of DMs with score-based models using SDEs. However, further studies still remain to fully understand and utilize the capabilities of DMs.

An important subject is the introduction of gradient guidance, including classifier-free guidance, to control the generative process (Dhariwal & Nichol, 2021; Sehwag et al., 2022; Avrahami et al., 2022b; Liu et al., 2021; Nichol et al., 2021; Rombach et al., 2022). Choi et al. (2021a) and Meng et al. (2021) have attempted to manipulate the resulting images of DMs by replacing latent variables, allowing the generation of desired random images. However, due to the lack of semantics in the latent variables of DMs, current approaches have critical problems with semantic image editing.

Alternative approaches have explored the potential of using the feature space within the U-Net for semantic image manipulation. For example, Baranchuk et al. (2021) and Tumanyan et al. (2022) use the feature map of the U-Net for semantic segmentation and maintaining the structure of generated images. Kwon et al. (2022) have shown that the bottleneck of the U-Net can be used as a semantic latent space. The experimental observation lacks a theoretical understanding of the feature map of DMs.

The study of latent spaces has gained significant attention in recent years. In the field of Generative Adversarial Networks (GANs), researchers have proposed various methods to manipulate the latent space to achieve the desired effect in the generated images. For example, local latent space manipulation techniques such as (Ramesh et al., 2018; Patashnik et al., 2021; Abdal et al., 2021) have been developed, as well as global manipulation techniques such as (Härkönen et al., 2020; Shen & Zhou, 2021; Yüksel et al., 2021). More recently, several studies (Zhu et al., 2021; Choi et al., 2021b) have examined the geometrical properties of latent space in GANs and utilized these findings for image manipulations. These studies bring the advantage of better understanding the characteristics of the latent space and facilitating the analysis and utilization of GANs. In contrast, the latent space of DMs remains poorly understood, making it difficult to fully utilize their capabilities.

Some studies have applied Riemannian geometry to analyze the latent spaces of deep generative models, such as Variational Autoencoders (VAEs) and GANs. (Arvanitidis et al., 2017; Shao et al., 2018; Chen et al., 2018; Arvanitidis et al., 2020) Shao et al. (2018) proposed a pullback metric on the latent space from image space Euclidean metric to analyze the latent space's geometry. This method has been widely used in VAEs and GANs because it only requires a differentiable map from latent space to image space. However, it has limitations such as a lack of evidence for applying the Euclidean metric in image space and the absence of a global semantic direction for manipulating arbitrary samples. Moreover, no studies have investigated the geometry of latent space of DMs utilizing the pullback metric.

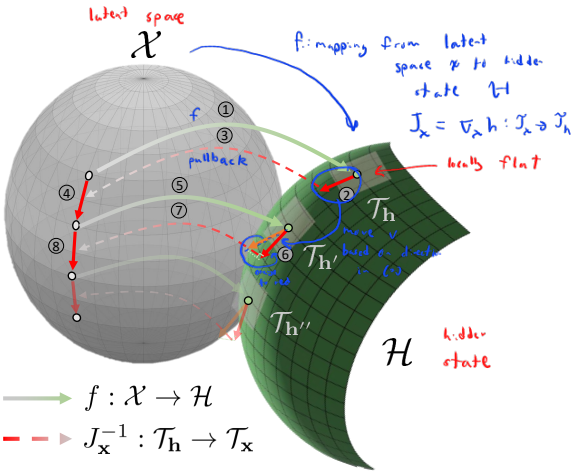


Figure 2. **Conceptual illustration of our editing procedure.** It consists of two parts: (① ∼ ④) discovering semantic latent directions using pullback metric and (⑤ ∼ ⑧) editing samples multiple times through geodesic shooting. ① Map a sample in $\mathcal{X}$ into a tangent space $\mathcal{T}_\mathbf{h}$ in $\mathcal{H}$. ② Choose a direction in $\mathcal{T}_\mathbf{h}$. ③ Find its corresponding direction in $\mathcal{X}$ using ${J_\mathbf{x}}^{-1}$. ④ Edit the sample by adding the discovered direction after normalizing to a predefined length. ⑤ Map the edited sample to a new tangent space $\mathcal{T}_{\mathbf{h}'}$ in $\mathcal{H}$ for multiple editing. ⑥ Using parallel transport, move the direction chosen in ② to the new tangent space $\mathcal{T}_{\mathbf{h}'}$. ⑦-⑧ Repeat ③-④. And then repeat ⑤-⑧.

3. Editing with semantic latent directions

This section explains how we extract the interpretable directions in the latent space of DMs using differential geometry. First, we adopt the local Euclidean metric of $\mathcal{H}$ to identify semantic directions for individual samples in $\mathcal{X}$. Second, we find global semantic directions by averaging the local semantic directions of individual samples. Then, we use the global directions to manipulate any sample to have the same interpretable features. Finally, we introduce a normalization technique to prevent distortion.

3.1. Pullback metric

We consider a curved manifold, $\mathcal{X}$, where our latent variables $\mathbf{x}_t$ exist. The differential geometry represents $\mathcal{X}$ through patches of tangent spaces, $\mathcal{T}_\mathbf{x}$, which are vector spaces defined at each point $\mathbf{x}$. Then, all the geometrical properties of $\mathcal{X}$ can be obtained from the metric of $||dx||^2 = \langle dx, dx \rangle_\mathbf{x}$ in $\mathcal{T}_\mathbf{x}$. However, we do not have any knowledge of $\langle dx, dx \rangle_\mathbf{x}$. It is definitely not a Euclidean metric. Furthermore, samples of $\mathbf{x}_t$ at intermediate timesteps of DMs include inevitable noise, which prevents finding semantic directions in $\mathcal{T}_\mathbf{x}$.

Fortunately, Kwon et al. (2022) observed that $\mathcal{H}$, defined by the bottleneck layer of the U-Net, exhibits local linearity. This allows us to adopt the Euclidean metric on $\mathcal{H}$. In differential geometry, when a metric is not available on a space, *pullback metric* is used. If a smooth map exists between the original metric-unavailable space and a metric-available space, the pullback metric of the mapped space is used to measure the distances in the original space. Our idea is to use the pullback Euclidean metric on $\mathcal{H}$ to define the distances between the samples in $\mathcal{X}$.

DMs are trained to infer the noise ϵ_t from a latent variable $\mathbf{x}_t$ at each diffusion timestep t. Each $\mathbf{x}_t$ has a different internal representation $\mathbf{h}_t$, the bottleneck representation of the U-Net, at different t's. The differentiable map between $\mathcal{X}$ and $\mathcal{H}$ is denoted as $f : \mathcal{X} \to \mathcal{H}$. Hereafter, we refer to $\mathbf{x}_t$ as $\mathbf{x}$ for brevity unless it causes confusion. It is important to note that our method can be applied at any timestep in the denoising process. The differential geometry then defines a linear map between the tangent space $\mathcal{T}_\mathbf{x}$ at $\mathbf{x}$ and corresponding tangent space $\mathcal{T}_\mathbf{h}$ at $\mathbf{h}$. The linear map can be described by the *Jacobian* $J_\mathbf{x} = \nabla_\mathbf{x}\mathbf{h}$ which determines how a vector $\mathbf{v} \in \mathcal{T}_\mathbf{x}$ is mapped into a vector $\mathbf{u} \in \mathcal{T}_\mathbf{h}$ by $\mathbf{u} = J_\mathbf{x}\mathbf{v}$. In practice, the Jacobian can be computed from automatic differentiation of the U-Net. However, since the Jacobian of too many parameters is not tractable, we use a sum-pooled feature map of the bottleneck representation as our $\mathcal{H}$.

Using the local linearity of $\mathcal{H}$, we assume the metric, $||d\mathbf{h}||^2 = \langle d\mathbf{h}, d\mathbf{h} \rangle_\mathbf{h} = d\mathbf{h}^\top d\mathbf{h}$ as a usual dot product defined in the Euclidean space. To assign a geometric structure to $\mathcal{X}$, we use the pullback metric of the corresponding $\mathcal{H}$. The pullback norm of $\mathbf{v} \in \mathcal{T}_\mathbf{x}$ is defined as follows:

$$||\mathbf{v}||^2_{\text{pb}} \triangleq \langle \mathbf{u}, \mathbf{u} \rangle_\mathbf{h} = \mathbf{v}^\top J_\mathbf{x}^\top J_\mathbf{x} \mathbf{v}. \tag{1}$$

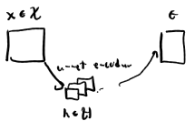
3.2. Extracting the semantic directions and editing

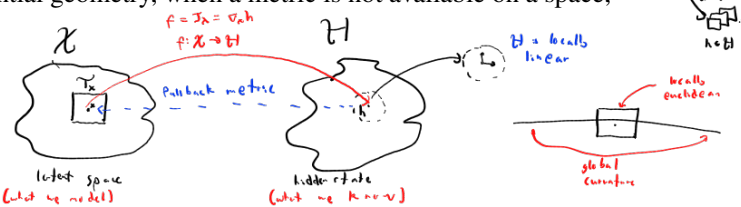
This subsection describes how we extract semantic latent directions using the pullback metric, and how we edit samples for multiple times given the meaningful directions by geodesic shooting. The overall process is illustrated in Figure 2.

Semantic latent directions Using the pullback metric, we can extract semantic directions of $\mathbf{v} \in \mathcal{T}_\mathbf{x}$ that show large variability of the corresponding $\mathbf{u} \in \mathcal{T}_\mathbf{h}$. We find a unit vector $\mathbf{v}_1$ that maximizes $||\mathbf{v}||^2_{\text{pb}}$. In practice, $\mathbf{v}_1$ corresponds to the first right singular vector from the singular value decomposition of $J_\mathbf{x} = U\Lambda V^\top$. It can be interpreted as the first eigenvector of $J_\mathbf{x}^\top J_\mathbf{x} = V\Lambda^2 V^\top$. By maximizing $||\mathbf{v}||^2_{\text{pb}}$ while remaining orthogonal to $\mathbf{v}_1$, one can obtain the second unit vector $\mathbf{v}_2$. This process can be repeated to have n semantic directions of $\{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n\}$ in $\mathcal{T}_\mathbf{x}$.

Using the linear transformation between $\mathcal{T}_\mathbf{x}$ and $\mathcal{T}_\mathbf{h}$ via the Jacobian $J_\mathbf{x}$, one can also obtain semantic directions in $\mathcal{T}_\mathbf{h}$:

$$\mathbf{u}_i = \frac{1}{\lambda_i} J_\mathbf{x} \mathbf{v}_i. \tag{2}$$





Here, we normalize $\mathbf{u}_i$ by dividing the i-th singular value λ_i of Λ to preserve the Euclidean norm $||\mathbf{u}_i|| = 1$. After selecting the top n (e.g. $n = 50$) directions of large eigenvalues, we can approximate any vector in $\mathcal{T}_\mathbf{h}$ with finite basis, $\{\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_n\}$. When we refer to a tangent space henceforth, it means the n-dimensional low-rank approximation of the original tangent space.

Iterative editing with geodesic shooting Now, we edit a sample with the i-th semantic direction through $\mathbf{x} \rightarrow \mathbf{x}' = \mathbf{x} + \gamma\mathbf{v}_i$, where γ is a hyper-parameter that controls the size of the editing. If we want to increase the editing strength, we need to repeat the same operation. However, this would not work because $\mathbf{v}_i$ may escape from the tangent space $\mathcal{T}_{\mathbf{x}'}$. Thus, it is necessary to relocate the extracted direction to a new tangent space. To achieve this, we use *parallel transport* that projects $\mathbf{v}_i$ onto the new tangent space $\mathcal{T}_{\mathbf{x}'}$. Parallel transport moves a vector without changing its direction as much as possible, while keeping the vector tangent on the manifold (Shao et al., 2018). It is notable that the projection significantly modifies the original vector $\mathbf{v}_i$, because $\mathcal{X}$ is a curved manifold. However, $\mathcal{H}$ is relatively flat. Therefore, it is beneficial to apply the parallel transport in $\mathcal{H}$.

To project $\mathbf{v}_i$ onto the new tangent space $\mathcal{T}_{\mathbf{x}'}$, we use parallel transport in $\mathcal{H}$. First, we convert the semantic direction $\mathbf{v}_i$ in $\mathcal{T}_\mathbf{x}$ to the corresponding direction of $\mathbf{u}_i$ in $\mathcal{T}_\mathbf{h}$. Second, we apply the parallel transport $\mathbf{u}_i \in \mathcal{T}_\mathbf{h}$ to $\mathbf{u}'_i \in \mathcal{T}_{\mathbf{h}'}$, where $\mathbf{h}' = f(\mathbf{x}')$. The parallel transport has two steps. The first step is to project $\mathbf{u}_i$ onto a new tangent space. This step keeps the vector tangent to the manifold. The second step is to normalize the length of the projected vector. This step preserves the size of the vector. Third, we obtain $\mathbf{v}'_i$ by transforming $\mathbf{u}'_i$ into $\mathcal{X}$. Using this parallel transport of $\mathbf{v}_i \rightarrow \mathbf{v}'_i$ via $\mathcal{H}$, we can realize the multiple feature editing of $\mathbf{x} \rightarrow \mathbf{x}' = \mathbf{x} + \gamma\mathbf{v}_i \rightarrow \mathbf{x}'' = \mathbf{x}' + \gamma\mathbf{v}'_i$. Based on the definition of Jacobian, this editing process can be viewed as a movement in $\mathcal{H}$ with the corresponding direction, i.e., $\mathbf{h} \rightarrow \mathbf{h}' = \mathbf{h} + \delta\mathbf{u}_i \rightarrow \mathbf{h}'' = \mathbf{h}' + \delta\mathbf{u}'_i$. This iterative editing procedure is called *geodesic shooting*, since it naturally forms a geodesic (Shao et al., 2018). Figure 2 summarizes the above procedure. See Appendix D for details.

3.3. Global semantic directions

We extracted meaningful directions for editing $\mathbf{x}_t$. However, the semantic latent directions are *local*, and thus are applicable only to individual samples of $\mathbf{x}_t$. Thus, we need to obtain *global* semantic directions that have the same semantic meaning for every sample. In this study, we observed a large overlap between the latent directions of individual samples. This observation motivates us to hypothesize that $\mathcal{H}$ has global semantic directions. To verify this hypothesis, we investigate whether, for any $\mathbf{u}_i^{(1)} \in \mathcal{T}_{\mathbf{h}^{(1)}}$, there exists

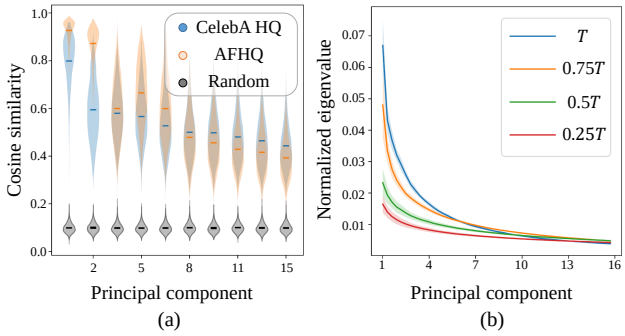


Figure 3. **(a) Homogeneity across local directions from different images.** A distribution represents the statistics of maximum cosine similarities between the principal directions of pairs of 100 samples in $\mathbf{x}_T$. The top principal directions better align than the rest. The comparison with random directions (black) confirms that the similarity does not arise by chance. **(b) Eigenvalue spectrum of $J_\mathbf{x}$ at different timesteps.** Early timesteps ($t \approx T$) have larger top eigenvalues implying fewer but more eminent directions than later timesteps.

$\mathbf{u}_j^{(2)} \in \mathcal{T}_{\mathbf{h}^{(2)}}$ that has a large overlap with $\mathbf{u}_i^{(1)}$. Then, we compare latent directions of $\mathbf{u}_i^{(1)}$ between many samples of $\mathbf{x}_t$. For the dominant directions of $\mathbf{u}_i^{(1)}$ and $\mathbf{u}_j^{(2)}$ with large eigenvalues of $\lambda_i^{(1)}$ and $\lambda_j^{(2)}$, we always found a good pair of (i, j) that showed a significant overlap between the two unit vectors when $t = T$ (Figure 3 (a)). Thus, we define global semantic directions, $\bar{\mathbf{u}}_i$, by averaging the closest latent directions in $\mathcal{H}$ of individual samples of $\mathbf{x}_T$. The global direction can be used to edit any sample $\mathbf{x}$. Note that $\bar{\mathbf{u}}_i$ can sometimes escape from the local tangent space of $\mathcal{T}_\mathbf{h}$. To mitigate this escape, we project $\bar{\mathbf{u}}_i$ into $\mathcal{T}_\mathbf{h}$. Since our method edits the sample in $\mathcal{X}$, we transform $\bar{\mathbf{u}}_i$ into the corresponding direction $\bar{\mathbf{v}}_i$ in $\mathcal{T}_\mathbf{x}$ via the Jacobian.

However, it is cautious to apply our hypothesis when we consider $\mathbf{x}_t$ for small t. We compared eigenvalue spectra between different t, and observed that they become flatter as t is closer to 0 (Figure 3 (b)). This shows that a few dominant feature directions exist for $\mathbf{x}_T$, whereas diverse feature directions exist for $\mathbf{x}_t$ with small t. Then, it is difficult to define global directions based on the homogeneity of local feature directions.

3.4. Normalizing distortion due to editing

DMs generate images by iteratively denoising $\mathbf{x}_T \rightarrow \mathbf{x}_{T-1} \rightarrow \cdots \rightarrow \mathbf{x}_0$. Suppose that we edit an image of $\mathbf{x}_t$ at a time step t with $\mathbf{x}_t \rightarrow \mathbf{x}_t + \gamma\mathbf{v}_i$. The editing signal of $\mathbf{v}_i$ is propagated and amplified throughout the denoising process. The amplification may lead to unexpected artifacts in generating $\mathbf{x}_0$. To avoid this problem, some normalization of $\mathbf{x}_t$ is necessary after the editing. However, it is difficult to

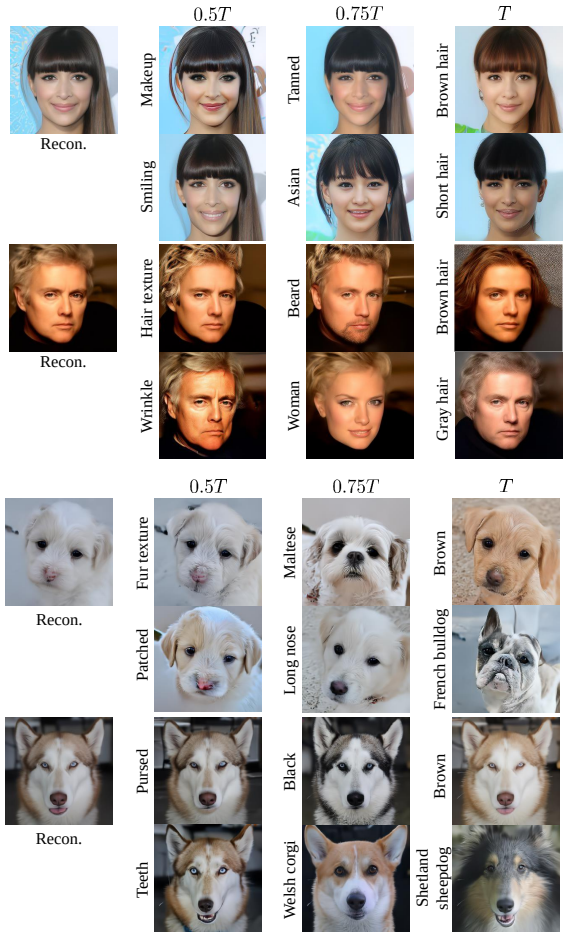


Figure 4. **Example edited results by the semantic latent directions.** "Recon." denotes the reconstruction of real images through DDIM inversion. The attributes are manually interpreted because the directions are not supervised. Different columns are edited at different denoising timesteps ($0.5T$, $0.75T$, and T).

normalize only the signal inside $\mathbf{x}_t$ that is mixed with white noise. Here, we propose an improved editing method.

Denoising diffusion implicit models (DDIM) computes $\mathbf{x}_0$ from $\mathbf{x}_t$ with predicted noise $\boldsymbol{\epsilon}_t^\theta(\mathbf{x}_t)$ (Song et al., 2020a):

$$\sqrt{\alpha_t}\mathbf{x}_0 = \mathbf{x}_t - \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_t^\theta(\mathbf{x}_t). \tag{3}$$

With a little abuse of notation, let $\mathbf{x}_0(\mathbf{x}_t)$ be a function of $\mathbf{x}_t$. In an ideal scenario, $\mathbf{x}_0(\mathbf{x}_t)$ can be assumed to contain only the signal of $\mathbf{x}_t$, which simplifies the regularization process (Zhang et al., 2022). Our improved editing method consists of three steps. First, we edit the original image as $\mathbf{x}_t \rightarrow \mathbf{x}_t + \gamma\mathbf{v}_i$. Second, we regularize $\mathbf{x}_0(\mathbf{x}_t + \gamma\mathbf{v}_i)$ to preserve its signal after the edition. Regularization is implemented by normalizing the pixel-to-pixel standard deviation of $\mathbf{x}_0(\mathbf{x}_t + \gamma\mathbf{v}_i)$, while keeping it's mean pixel values fixed. We denote the normalized $\mathbf{x}_0(\mathbf{x}_t + \gamma\mathbf{v}_i)$ as $\mathbf{x}'_0$. Third, we solve the DDIM equation for $\mathbf{x}'_t$, $\sqrt{\alpha_t}\mathbf{x}'_0 = \mathbf{x}'_t - \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_t^\theta(\mathbf{x}'_t)$, to obtain a corresponding edited sample which may be derived from $\mathbf{x}_t + \gamma\mathbf{v}_i$. Using the first-order Taylor expansion, $\boldsymbol{\epsilon}_t^\theta(\mathbf{x}'_t) \approx \boldsymbol{\epsilon}_t^\theta(\mathbf{x}_t) + \nabla_{\mathbf{x}_t}\boldsymbol{\epsilon}_t^\theta(\mathbf{x}_t)\cdot(\mathbf{x}'_t - \mathbf{x}_t)$, we have an updated equation:

$$\mathbf{x}'_t = \mathbf{x}_t + \frac{\sqrt{\alpha_t}}{1-\kappa\sqrt{1-\alpha_t}}(\mathbf{x}'_0 - \mathbf{x}_0(\mathbf{x}_t)), \tag{4}$$

where we use $\kappa = 0.99$. See Appendix B for a detailed derivation.

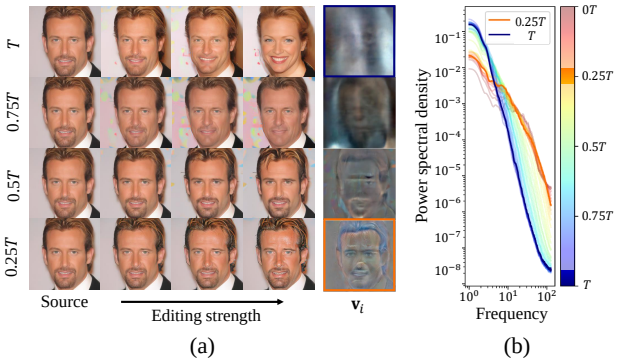


Figure 5. **Comparison of the directions at different timesteps.** (a) Qualitative comparison showing that the directions in earlier timesteps edit coarse attributes while ones in later timesteps focus on high-frequency components. (b) Power spectral density (PSD) of $\mathbf{v}_i$. The PSD at $t = T$ (blue line) shows a larger portion of low-frequency signals, whereas the PSD at smaller t (orange line) shows a larger portion of high-frequency signals.

4. Experiments

Thorough experiments demonstrate the usefulness of our method in various aspects. The editing latent directions in $\mathcal{X}$ found by our method include semantic changes and exhibit coarse-to-fine behavior (§ 4.1). $\mathcal{X}$ is a spherically curved space (§ 4.2). Our method generalizes to stable diffusion (§ 4.3). Both the finding directions and the editing equation contribute to the nice properties of our method (§ 4.4). Our method outperforms the existing methods (§ 4.5).

Implementation details We validate our method and provide analyzes in CelebA-HQ (Karras et al., 2018) for DDPM++ (Ho et al., 2020; Meng et al., 2021), AFHQ-dog (Choi et al., 2018) for iDDPM (Nichol & Dhariwal, 2021). All input images are from test sets in 256^2 resolution. Quantitative results are from CelebA-HQ, unless otherwise noted. For Stable Diffusion (Rombach et al., 2022), we use "Cyberpunk city" and "Painting of Van Gogh" as text prompts to showcase the versatility of our method. We use the official codes and pre-trained checkpoints for all baselines and keep the parameters *frozen*. Further implementation details are deferred to Appendix A. The source code for our experi-

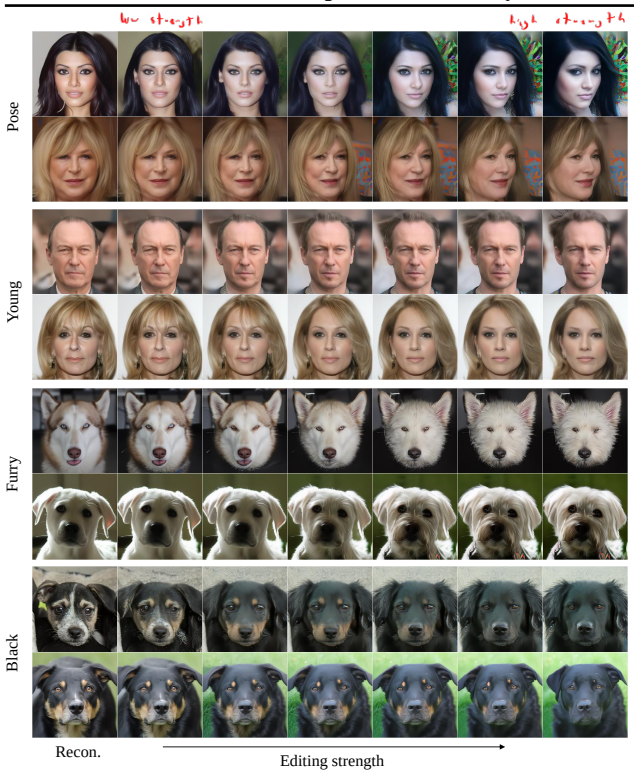


Figure 6. **Example images edited with global semantic directions.** Consistent semantic changes in two rows validate the global semantic direction. The attributes are manually interpreted because the directions are not supervised.

ments is included in the supplementary materials, and will be publicly available upon publication.

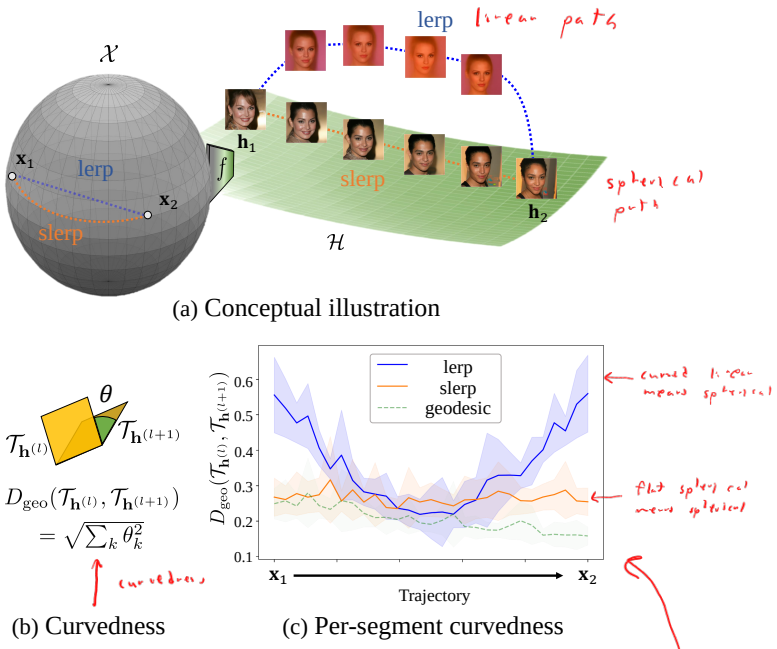
4.1. Image manipulation

Semantic latent directions Figure 4 illustrates the example results edited by the directions found by our method *without supervision* such as CLIP or a classifier. The directions clearly contain semantics such as gender, age, ethnicity, facial expression, breed, and texture. Interestingly, editing at timestep T leads to coarse changes such as hair color, hair length, far breed. On the other hand, editing at the timestep $0.5T$ leads to fine changes such as make-up, hair texture, wrinkles, facial expression, and close breed. Appendix E.1 provides more examples.

Editing timing We further investigate the coarse-to-fine editing along the generative process from timestep T to 0. Figure 5 (a) shows the example directions $\mathbf{v}_i$ across different timesteps. At T, $\mathbf{v}_i$ leads to coarse attribute changes in $\mathbf{x}_0$ by blurry change in $\mathbf{x}_T$. At $0.25T$, $\mathbf{v}_i$ edits high-frequency details in both $\mathbf{x}_0$ and $\mathbf{x}_t$. Figure 5 (b) shows the power spectral density (PSD) of $\mathbf{v}_i$. We compute the PSD by taking $\mathbf{v}_1, ..., \mathbf{v}_{10}$ from 20 samples. The early timesteps contain a larger portion of low frequency than the later timesteps and the later timesteps contain a larger portion of high frequency. This phenomenon agrees with the tendency in the edited images. This results strengthens the common understanding of the timesteps (Kwon et al., 2022; Choi et al., 2022; Daras & Dimakis, 2022).

Table 1. Semantic path length for lerp, slerp, and geodesic paths in CelebA-HQ for DDPM++.

Path	Semantic Path Length ($\mu \pm \sigma$)
lerp	10.29 ± 1.11
slerp	7.69 ± 0.87
geodesic	5.98 ± 0.76



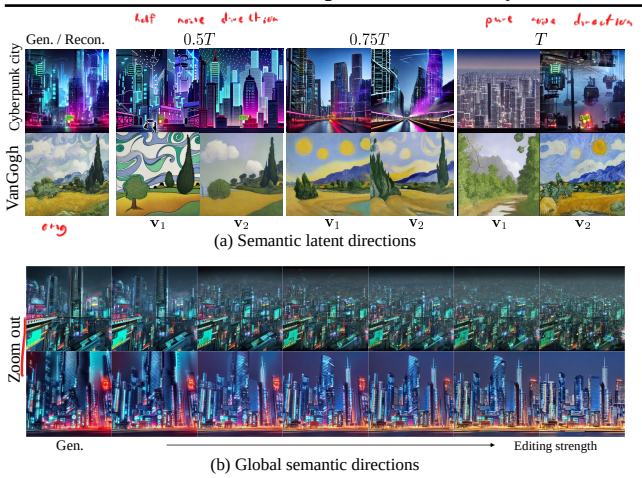
(a) Conceptual illustration

(b) Curvedness

(c) Per-segment curvedness

Figure 7. $\mathcal{X}$ **is a curved manifold.** (a) Conceptual illustration of a linear path (lerp) and a spherical path (slerp) on the manifolds. (b) Curvedness of a line segment. (c) Per-segment distribution of curvedness along different paths. The lerp paths roughly have higher curvedness than slerp and reach similar curvedness to slerp. The slerp paths are closer to the geodesic shooting paths. It implies that $\mathcal{X}$ is a curved manifold. The shades depict $\pm$ 0.5 standard deviation. We use 50 segments for each path between $\mathbf{x}_1$ and $\mathbf{x}_2$.

Global semantic directions Figure 6 demonstrates that the global directions in $\mathbf{x}_t$ lead to the same semantic changes, such as rotation, age, furriness, or color in different samples. It confirms that $\mathcal{X}$ inherits the homogeneity of $\mathcal{H}$ via the pullback metric although $\mathcal{X}$ is a metric-less space. Appendix E.2 provides more examples.



(a) Semantic latent directions

(b) Global semantic directions

Figure 8. **Generalization to Stable Diffusion.** (a) Semantic latent directions successfully edit the images conditioned on "Cyberpunk" and "painting of Van Gogh". The leftmost images are the originals. Different directions and timesteps edit different attributes. (b) A global semantic direction consistently zooms out the different generated images.

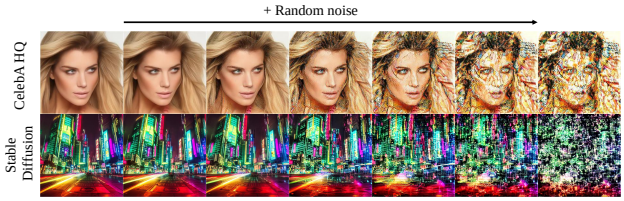


Figure 9. **Importance of the discovered semantic directions.** Adding random directions instead of semantic directions severely distorts the resulting images.

Figure 10. **Importance of the normalization in Eq. (4).** Removing the normalization leads to excessive saturation.

4.2. Curved manifold of DMs

We present empirical grounds for the assumption in § 3.2: $\mathcal{X}$ is a curved manifold. Semantic path length between two points on a manifold is defined by the sum of the local warpage of the line segments which connects them along the manifold. We use *geodesic metric* (Choi et al., 2021b; Ye & Lim, 2016) to define the curvedness of a line segment $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}\}$ as the angle between two tangent spaces centered at $\{\mathbf{h}^{(1)}, \mathbf{h}^{(2)}\}$:

$$D_{\text{geo}}(\mathcal{T}_{\mathbf{h}^{(1)}}, \mathcal{T}_{\mathbf{h}^{(2)}}) = \sqrt{\sum_k \theta_k^2}, \quad (5)$$

where $\theta_k = \cos^{-1}(\sigma_k)$ denotes the k-th principle angle between $\mathcal{T}_{\mathbf{h}^{(1)}}$ and $\mathcal{T}_{\mathbf{h}^{(2)}}$. The angle is visualized in Figure 7 (b). Then, the semantic path length becomes $\sum_l D_{\text{geo}}(\mathcal{T}_{\mathbf{h}^{(l)}}, \mathcal{T}_{\mathbf{h}^{(l+1)}})$, where l denotes the segment index in the path. We set the number of segments to 30. Then, the semantic path length increases as the path deviates further from the manifold.

To verify the assumption, we compare the semantic path lengths of different paths, e.g., linear path, spherical path, and geodesic shooting path. Figure 7 (a) visualizes the manifold, linear path (lerp), and spherical path (slerp) and their corresponding path on $\mathcal{H}$ mapped by the function f. We computed the semantic path lengths for 50 randomly selected pairs of images. Table 1 shows that the semantic path length of slerp is smaller than lerp, indicating that the slerp path lies closer to the manifold than lerp, i.e., the manifold is curved. Figure 7 (b) shows the distribution of the length of the segments along the path. Interestingly, the length of the lerp is high at the ends and shrinks to that of geodesics near the center. We suppose that the lerp path moves away from the original manifold and moves along another manifold.

Our semantic path length resembles the perceptual path length (PPL, Karras et al. (2019)) regarding the summation along the interpolation path. PPL measures LPIPS (Zhang et al., 2018) distance between resulting images along the path. Higher PPL between two latent variables indicates spikier interpolation of images accompanying artifacts. On the other hand, semantic path length measures how drastically the geometric structure changes between neighboring tangent spaces.

4.3. Stable diffusion

This section demonstrates that our method is generalized to Stable Diffusion (Rombach et al., 2022). Our method extracts latent directions in the learned latent space $\mathbf{z}_t$ using the same procedure. Figure 8 (a) shows the edited images along different directions on various timesteps. The phenomena are similar to the image-based DMs: editing at $t = T$ provides coarse changes, and editing at later timesteps provides more fine texture-ish changes such as cartoonization.

Furthermore, Figure 8 (b) shows that a global semantic direction leads to the same *zoom-out* effect on different samples. Contrary to the global directions in the image-based DMs, the global directions are found within a text prompt, i.e., each text prompt has its own global directions. Appendix E provides more examples where we find some odd cases indicating that the learned latent space may not follow the same assumptions of the image-based DMs or the text guidance somehow twists the manifold.

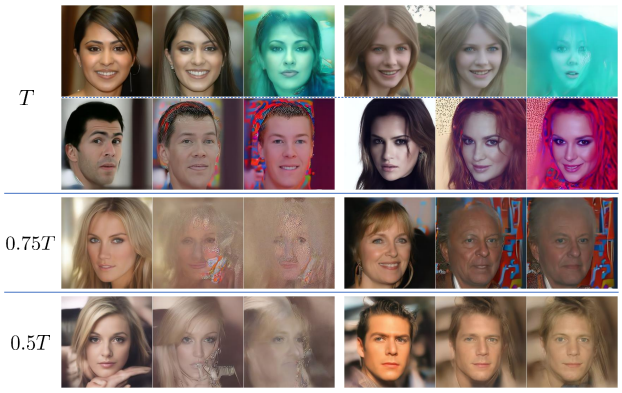


Figure 11. **Inferiority of GANSpace on $\mathcal{H}$.** The GANSpace directions accompany severe distortion or entanglement while somewhat altering the attributes such as expression, rotation, and age.

4.4. Ablation study

We provide ablation studies that include alternative approaches. First, we edit images by applying random directions instead of semantic latent directions. Figure 9 shows that random directions seriously degrade the images. This experiment validates the excellence of the latent directions found by our method.

Figure 10 demonstrates the necessity of normalization in Eq. (4). While our full method produces plausible edited images even with extreme changes, removing the normalization leads to excessive saturation.

4.5. Comparison to other editing methods

As we introduce the first unsupervised editing in DMs, we compare our method with GANSpace (Härkönen et al., 2020) considering the mapping from $\mathcal{X}$ to $\mathcal{H}$ instead of $\mathcal{Z}$ to $\mathcal{W}$ in GANs. Accordingly, we find directions in $\mathcal{H}$ using PCA. Figure 11 shows their effects: they somewhat alter the attributes but accompany severe distortion or entanglement. On the contrary, our method finds the directions with the largest changes in $\mathcal{H}$ considering the geometrical structure leading to decent manipulation as shown in earlier results. Appendix C describes more details for GANSpace.

5. Discussion

In this section, we provide additional intuitions and implications. It is interesting that our semantic latent directions usually convey disentangled attributes even though we do not adopt attribute annotation to enforce disentanglement. We suppose that decomposing the Jacobian of the encoder in the U-Nets naturally yields disentanglement to some extent. It grounds on the linearity of the intermediate feature space $\mathcal{H}$ in the U-Nets (Kwon et al., 2022). However, it does not guarantee the perfect disentanglement and some directions are entangled. For example, the direction for long hair converts the male subject to female as shown in Figure 12 (a). This kind of entanglement often occurs in other editing methods due to the dataset prior: there are few male faces with long hair.

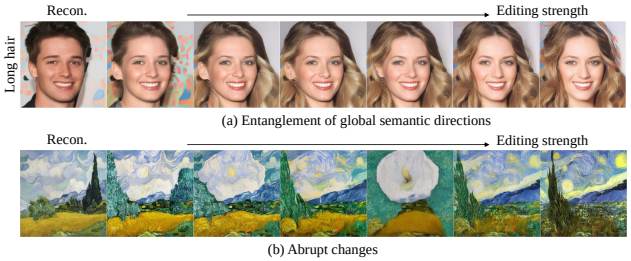


Figure 12. **Limitations.** (a) Entanglement between attributes due to the dataset prior. (b) Abrupt changes in Stable Diffusion.

Although we have shown that our method is also valid to Stable Diffusion, we still need more observation. It discovers less number of semantic latent directions and few directions occasionally convey abrupt changes during the editing procedure in Stable Diffusion as shown in Figure 12 (b). We suppose that its learned latent space may have a more complex manifold than the image space (Arvanitidis et al., 2017). Alternatively, the conditional DMs with classifier-free guidance or the cross-attention mechanism may add complexity on the manifold. Our future work includes analyzing the latent directions in the conditions such as text prompts or segmentation labels.

Despite these limitations, our method provides a significant advance in the field of image editing for DMs, and potential applications in a wide range of tasks.

6. Conclusion

In this work, we have proposed an unsupervised approach to extract semantic latent directions in $\mathcal{X}$, the latent space of diffusion models (DMs). Decomposing the Jacobian of the encoder in the U-Nets discovers the directions that manipulate mostly disentangled attributes. Our detailed analyses provide in-depth understanding of DMs: 1) different samples share the same latent directions in local tangent space leading to global semantic directions, 2) the generative process produces low-frequency components and adds high-frequency details, 3) the latent variable $\mathbf{x}_t$ lives in a curved manifold, and 4) Stable Diffusion shares the similar intuitions with image-based DMs in the learned latent space.

Furthermore, we believe that better understanding the latent space of DMs will open up new possibilities for the development of DMs in useful applications, similar to how the arithmetic operations on the latent space of GANs has led to various follow-up research.

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.

Arvanitidis, G., Hansen, L. K., and Hauberg, S. Latent space oddity: on the curvature of deep generative models. *arXiv preprint arXiv:1710.11379*, 2017.

Arvanitidis, G., Hauberg, S., and Schölkopf, B. Geometrically enriched latent spaces. *arXiv preprint arXiv:2008.00565*, 2020.

Avrahami, O., Hayes, T., Gafni, O., Gupta, S., Taigman, Y., Parikh, D., Lischinski, D., Fried, O., and Yin, X. Spatext: Spatio-textual representation for controllable image generation. *arXiv preprint arXiv:2211.14305*, 2022a.

Avrahami, O., Lischinski, D., and Fried, O. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18208–18218, 2022b.

Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

Baranchuk, D., Rubachev, I., Voynov, A., Khrulkov, V., and Babenko, A. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021.

Chen, N., Klushyn, A., Kurle, R., Jiang, X., Bayer, J., and Smagt, P. Metrics for deep generative models. In *International Conference on Artificial Intelligence and Statistics*, pp. 1540–1550. PMLR, 2018.

Choi, J., Kim, S., Jeong, Y., Gwon, Y., and Yoon, S. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021a.

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. *arXiv preprint arXiv:2106.06959*, 2021b.

Choi, J., Lee, J., Shin, C., Kim, S., Kim, H., and Yoon, S. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11472–11481, 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8789–8797, 2018.

Daras, G. and Dimakis, A. G. Multiresolution textual inversion. *arXiv preprint arXiv:2211.17115*, 2022.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Härkönen, E., Hertzmann, A., Lehtinen, J., and Paris, S. Ganspace: Discovering interpretable gan controls. *Advances in Neural Information Processing Systems*, 33: 9841–9850, 2020.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models. *arXiv preprint arXiv:2206.00364*, 2022.

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. *arXiv preprint arXiv:2210.09276*, 2022.

Kwon, M., Jeong, J., and Uh, Y. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022.

Liew, J. H., Yan, H., Zhou, D., and Feng, J. Magicmix: Semantic mixing with diffusion models. *arXiv preprint arXiv:2210.16056*, 2022.

Liu, X., Park, D. H., Azadi, S., Zhang, G., Chopikyan, A., Hu, Y., Shi, H., Rohrbach, A., and Darrell, T. More control for free! image synthesis with semantic diffusion guidance. *arXiv preprint arXiv:2112.05744*, 2021.

Meng, C., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

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. *arXiv preprint arXiv:2112.10741*, 2021.

Nichol, A. Q. and Dhariwal, P. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171. PMLR, 2021.

Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., and Lischinski, D. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2085–2094, 2021.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Ramesh, A., Choi, Y., and LeCun, Y. A spectral regularizer for unsupervised disentanglement. *arXiv preprint arXiv:1812.01161*, 2018.

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

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and 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.

Sehwag, V., Hazirbas, C., Gordo, A., Ozgenel, F., and Canton, C. Generating high fidelity data from low-density regions using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11492–11501, 2022.

Shao, H., Kumar, A., and Thomas Fletcher, P. The riemannian geometry of deep generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 315–323, 2018.

Shen, Y. and Zhou, B. Closed-form factorization of latent semantics in gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1532–1540, 2021.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Tumanyan, N., Geyer, M., Bagon, S., and Dekel, T. Plug-and-play diffusion features for text-driven image-to-image translation. *arXiv preprint arXiv:2211.12572*, 2022.

Ye, K. and Lim, L.-H. Schubert varieties and distances between subspaces of different dimensions. *SIAM Journal on Matrix Analysis and Applications*, 37(3):1176–1197, 2016.

Yüksel, O. K., Simsar, E., Er, E. G., and Yanardag, P. Latentclr: A contrastive learning approach for unsupervised discovery of interpretable directions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14263–14272, 2021.

Zhang, Q., Tao, M., and Chen, Y. gddim: Generalized denoising diffusion implicit models. *arXiv preprint arXiv:2206.05564*, 2022.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Zhu, J., Feng, R., Shen, Y., Zhao, D., Zha, Z.-J., Zhou, J., and Chen, Q. Low-rank subspaces in gans. *Advances in Neural Information Processing Systems*, 34:16648–16658, 2021.

A. Implementation Details

Table A1 summarizes various hyperparameter settings in our experiments. Specific details not covered in the main text are discussed in the following paragraphs.

Inversion step To obtain the latent code of a given image, we compute the latent code $\mathbf{x}_T$ using DDIM inversion. (Song et al., 2020a) The inversion step hyperparameter refers to the number of DDIM steps used to calculate the latent code. For stable-diffusion, we use classifier-free guidance.

Low dimensional approximation (n) In our work, we employ a low-dimensional approximation of the tangent space. Rather than fixing the dimensionality at n, we determined to dynamically choose n based on the distribution of eigenvalues. More specifically, we approximated the tangent space with dimensions corresponding to eigenvalues with cumulative density below a given threshold. As such, Table 1 presents the threshold rather than the dimensionality n. It worth note that, despite being determined dynamically, the actual values of n has stable for various images. For example, for $t = T, 0.75T, 0.5T, 0.25T$, the values of n were approximately 25, 50, 75, and 100, respectively.

Quality boosting (t_{boost}) While DDIM alone is capable of generating high-quality images, Karras et al. (2022) showed that the inclusion of stochasticity improves image quality, and Kwon et al. (2022) suggested the technique of adding stochasticity at the end of the generative process. We employ this technique in our experiments on CelebA-HQ and Stable-Diffusion after t_{boost}.

Stable-Diffusion In order to mitigate the influence of classifier-free guidance, the strength of the guidance, denoted as w, was set to zero, utilizing only the text-conditional model. (Ho & Salimans, 2022) When generating the original Cyberpunk city images, we set the guidance strength as $w = 7.5$. The prompts utilized for the Cyberpunk city images were "Cyberpunk city" and for the Van Gogh paintings, the prompt used was "painting of Van Gogh." Through the process of DDIM inversion, latent codes $\mathbf{x}_T$, were generated given the appropriate prompts for each image, with the guidance strength also set to zero (i.e., guidance scale $= 1$ in the code).

B. Improved Editing Equation

Song et al. (2020a) derived the following equation:

$$\sqrt{\alpha_t}\mathbf{x}_0 = \mathbf{x}_t - \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_t^\theta(\mathbf{x}_t). \tag{6}$$

Given inferred $\mathbf{x}_0(\mathbf{x}_t + \mathbf{v}_i)$ under feature edition, our object is to find corrected $\mathbf{x}'_t$ that satisfies the above equation, $\sqrt{\alpha_t}\mathbf{x}'_0 = \mathbf{x}'_t - \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_t^\theta(\mathbf{x}'_t)$. As stated in the main text, it is difficult to have exact $\mathbf{x}'_t$. However, we can decompose the solution as $\mathbf{x}'_t = \mathbf{x}_t + d\mathbf{x}_t$, and then obtain the solution for the small $d\mathbf{x}_t$. This approximation follows as

$$\begin{aligned}
\sqrt{\alpha_t}\mathbf{x}'_0 &= \mathbf{x}'_t - \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_t^\theta(\mathbf{x}'_t) && (7)\\
\Rightarrow \sqrt{\alpha_t}(\mathbf{x}_0(\mathbf{x}_t) + d\mathbf{x}_0) &= \mathbf{x}_t + d\mathbf{x}_t - \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_t^\theta(\mathbf{x}_t + d\mathbf{x}_t) && (8)\\
\Rightarrow \sqrt{\alpha_t}(\mathbf{x}_0(\mathbf{x}_t) + d\mathbf{x}_0) &= \mathbf{x}_t + d\mathbf{x}_t - \sqrt{1-\alpha_t}\big(\boldsymbol{\epsilon}_t^\theta(\mathbf{x}_t) + \nabla_{\mathbf{x}_t}\boldsymbol{\epsilon}_t^\theta \cdot d\mathbf{x}_t\big) && (9)\\
\Rightarrow \sqrt{\alpha_t}d\mathbf{x}_0 &= d\mathbf{x}_t - \sqrt{1-\alpha_t}\nabla_{\mathbf{x}_t}\boldsymbol{\epsilon}_t^\theta \cdot d\mathbf{x}_t && (10)\\
\Rightarrow \sqrt{\alpha_t}d\mathbf{x}_0 &= d\mathbf{x}_t - \kappa\sqrt{1-\alpha_t}d\mathbf{x}_t && (11)\\
\therefore d\mathbf{x}_t &= \frac{\sqrt{\alpha_t}}{1-\kappa\sqrt{1-\alpha_t}}d\mathbf{x}_0 && (12)
\end{aligned}$$

In the third line of the derivation, we used a first-order Taylor expansion of $\boldsymbol{\epsilon}_t^\theta$. In the fourth line, we eliminated dominant terms on both sides using Eq.(6). In the fifth line, we approximated the Jacobian matrix as an identity matrix, $\nabla_{\mathbf{x}_t}\boldsymbol{\epsilon}_t^\theta \approx \kappa\mathrm{I}$. This approximation enables us to obtain $\mathbf{x}'_t$. it is important to note that the Jacobian is multiplied by $\sqrt{1-\alpha_t}$. Therefore, the component works only for t close to T, because otherwise $\sqrt{1-\alpha_t}$ vanishes. Then, it is sufficient to show that our

Table A1. Hyper-parameter settings.

Experiment	t_{edit}	γ	inversion step	threshold (n)	t_{boost}	guidance strength (w)
CelebA-HQ	T	0.0025	40	0.5	$0.15T$	$\times$
	$0.75T$	0.0125	40	0.5	$0.15T$	$\times$
	$0.5T$	0.2500	40	0.5	$0.15T$	$\times$
	$0.25T$	2.5000	40	0.5	$0.15T$	$\times$
AFHQ-dog	T	0.0025	80	0.5	$\times$	$\times$
	$0.75T$	0.0100	80	0.5	$\times$	$\times$
	$0.5T$	0.2500	80	0.5	$\times$	$\times$
	$0.25T$	2.5000	80	0.5	$\times$	$\times$
Stable-Diffusion	T	0.025	80	0.25	$0.15T$	0
	$0.75T$	0.100	80	0.25	$0.15T$	0
	$0.5T$	0.500	80	0.25	$0.15T$	0
	$0.25T$	2.5000	80	0.25	$0.15T$	0

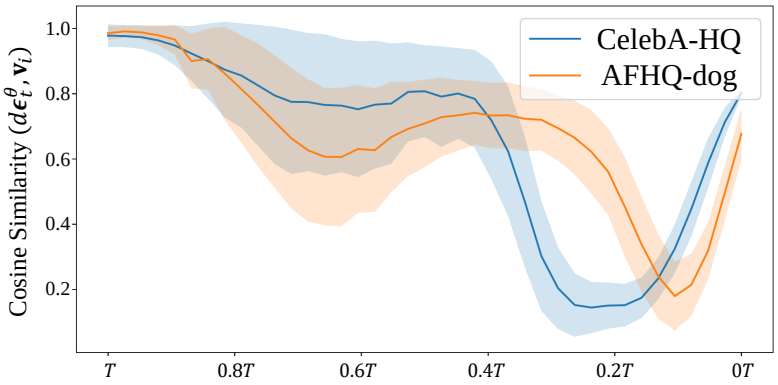


Figure A1. Cosine similiarity between $(d\boldsymbol{\epsilon}_t^\theta, \mathbf{v}_i)$. The shaded region represents the mean $\pm$ standard deviation of the measurements. The results depicted in the figure were obtained by measuring 100 samples from $\mathbf{x} \sim \mathcal{N}(0, \mathrm{I})$.

approximation works well in the range of t close to T. We examined the validity of our approximation numerically through the following equation,

$$\nabla_{\mathbf{x}_t} \boldsymbol{\epsilon}_t^\theta \approx \frac{\boldsymbol{\epsilon}_t^\theta(\mathbf{x}_t + d\mathbf{x}_t) - \boldsymbol{\epsilon}_t^\theta(\mathbf{x}_t)}{||d\mathbf{x}_t||} \tag{13}$$

where $d\mathbf{x}_t = \mathbf{v}_i$. Then, the approximation of $\nabla_{\mathbf{x}_t} \boldsymbol{\epsilon}_t^\theta \approx \kappa \mathrm{I}$ represents a good alignment between two vectors of $d\boldsymbol{\epsilon}_t^\theta$ and $\mathbf{v}_i$. We confirmed the good alignment using their cosine similarity (Figure A1). We use $\kappa = 0.99$, since it is a representative value in the range $t \approx T$. This value also serves to prevent the vector $d\mathbf{x}$ from becoming excessively small, when $\sqrt{\alpha_t} \approx 0$.

C. Comparison Details

Since unsupervised editing is not available for DMs, we consider GANSpace for image editing. The spaces of $\mathcal{Z}$ and $\mathcal{W}$ of GAN correspond to $\mathcal{X}$ and $\mathcal{H}$ of DM, respectively. We use 1k random images with DDIM generative process for GANSpace. Note that the GANSpace method is obtaining directions in $\mathcal{W}$ thus we used GANSpace to add directions directly to $\mathcal{H}$. In addition to what § 4.5 provides, the editing direction, extracted by the GANSpace, primarily alters colors in images. This suggests that simply collecting every $\mathbf{h}$ in $\mathcal{H}$ and extracting their principal axes may find poor feature directions that may control just overall color.

D. Algorithms

Algorithm 1 Feature Direction

Require: latent variable $\mathbf{x}$, timestep t, U-Net encoder $f : \mathcal{X} \times T \to \mathcal{H}$, Feature direction index i

J = Jacobian$(f(\cdot, t))(\mathbf{x})$
$U, S, V^\top$ = SingularValueDecomposition(J)
$\mathbf{v}_i, \mathbf{u}_i = V^\top[i, :], U[:, i]$
Return $\mathbf{v}_i, \mathbf{u}_i$

Algorithm 2 Global Feature Direction

Require: latent variable $\mathbf{x}$, U-Net encoder $f : \mathcal{X} \times T \to \mathcal{H}$, low-dimensional approximation n, number of local bases L, Global feature direction i

for $l = 1$ **to** L **do**
  $\mathbf{x} \sim \mathcal{N}(0, \mathrm{I})$
  J = Jacobian$(f(\cdot, T))(\mathbf{x})$
  $U^{(l)}, \cdot, \cdot$ = SingularValueDecomposition(J)
  $U^{(l)} = U^{(l)}[:, :n]$
end for
for $l = 1$ **to** L **do**
  if $l = 1$ **then**
    $\bar{U} = U^{(l)}$
  else
    for $m = 1$ **to** n **do**
      $U^{(l)}$ = SortBySimilarity$(U^{(l)}, \bar{U})$
      $\bar{U} = \bar{U} + \sum_k \mathrm{sign}(\sum_c U^{(l)}_{ck} \bar{U}_{ck}) U^{(l)}_{\cdot k}$
    end for
  end if
end for
$\bar{U} = \frac{1}{L} \bar{U}$
$\bar{\mathbf{u}}_i = \bar{U}[:, i]$
Return $\bar{\mathbf{u}}_i$

Algorithm 3 Tangent Space Projection

Require: latent variable $\mathbf{x}$, timestep t, U-Net encoder $f : \mathcal{X} \times T \to \mathcal{H}$, $\mathcal{H}$ direction $\mathbf{u}_i$, low-dimensional approximation n.

$J_\mathbf{x}$ = Jacobian$(f(\cdot, t))(\mathbf{x})$ Get target space of x in H
$U, S, V^\top$ = SingularValueDecomposition$(J_\mathbf{x})$
$U, V^\top = U[:, :n], V^\top[:n, :]$ ← Get n principal components
$\mathbf{u}'_i = UU^\top \mathbf{u}_i$
$\mathbf{u}'_i \mathrel{/}= ||\mathbf{u}'_i||$
$\mathbf{v}'_i = VU^\top \mathbf{u}'_i$
Return $\mathbf{v}'_i, \mathbf{u}'_i$

Algorithm 4 Eq. (4)

Require: latent variable $\mathbf{x}$, timestep t, map from $\mathbf{x}_t$ to predicted $\mathbf{x}_0$ $P : \mathcal{X}_t \to \mathcal{X}_0$ edit step size γ, feature direction $\mathbf{v}_i$

$\mathbf{x}'_0, \mathbf{x}_0 = P(\mathbf{x}_0 + \gamma \mathbf{v}_i), P(\mathbf{x}_0)$

$\mu_{\mathbf{x}'_0}, \mu_{\mathbf{x}_0}$ = Mean($\mathbf{x}'_0$), Mean($\mathbf{x}_0$)

$\mathbf{x}'_0 = \mu_{\mathbf{x}'_0} + (\mathbf{x}'_0 - \mu_{\mathbf{x}'_0}) \frac{\text{StandardDeviation}(\mathbf{x}_0 - \mu_{\mathbf{x}_0})}{\text{StandardDeviation}(\mathbf{x}'_0 - \mu_{\mathbf{x}'_0})}$

$d\mathbf{x} = \frac{\sqrt{\alpha_t}}{1 - 0.99\sqrt{1-\alpha_t}} (\mathbf{x}'_0 - \mathbf{x}_0)$

Return $d\mathbf{x}$

Algorithm 5 Total Editing Process

Require: latent variable $\mathbf{x}$, timestep t, U-Net encoder $f : \mathcal{X} \times T \to \mathcal{H}$, (Global) feature direction index i, low-dimensional approximation n, edit step size γ, edit iteration N_{iter}

if Use Global Feature Direction **then**

  $\mathbf{u}_i$ = GlobalFeatureDirection($\mathbf{x}$, f, n, L, i)

else

  $\mathbf{u}_i$ = FeatureDirection($\mathbf{x}$, t, f, i)

end if

for $edit = 1$ **to** N_{iter} **do**

  $\mathbf{v}_i, \mathbf{u}_i$ = TangentSpaceProjection($\mathbf{x}$, t, f, $\mathbf{u}_i$, n)

  $d\mathbf{x}$ = Eq. (4)($\mathbf{x}$, t, P, γ, $\mathbf{v}_i$)

  $\mathbf{x} = \mathbf{x} + d\mathbf{x}$

end for

Return $\mathbf{x}$

E. Additional results

E.1. Feature direction

Figure A2. $\boldsymbol{t = T}$**.** A selection of interpretable edits discovered by our feature direction in CelebA-HQ. The image on the far left represents the reconstructed original image, while the subsequent images demonstrate the interpretable edits that have been made to it.

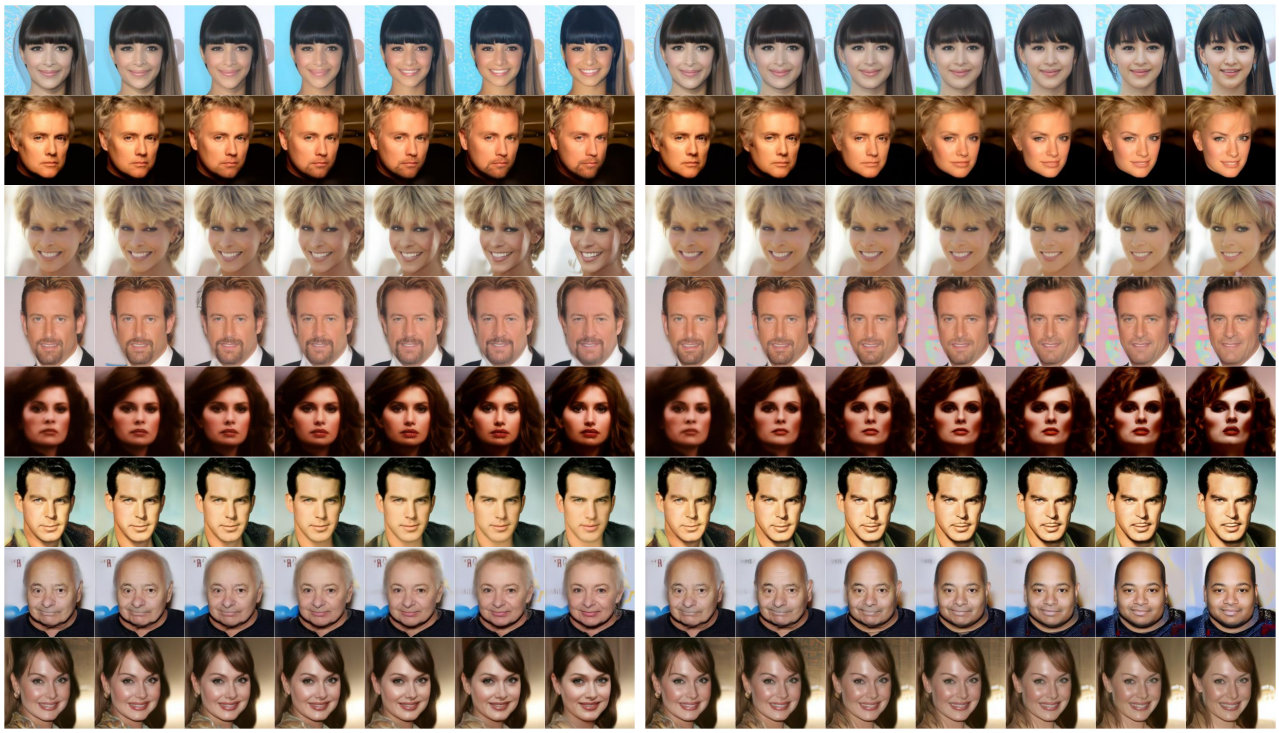


Figure A3. $t = 0.75T$. A selection of interpretable edits discovered by our feature direction in CelebA-HQ. The image on the far left represents the reconstructed original image, while the subsequent images demonstrate the interpretable edits that have been made to it.

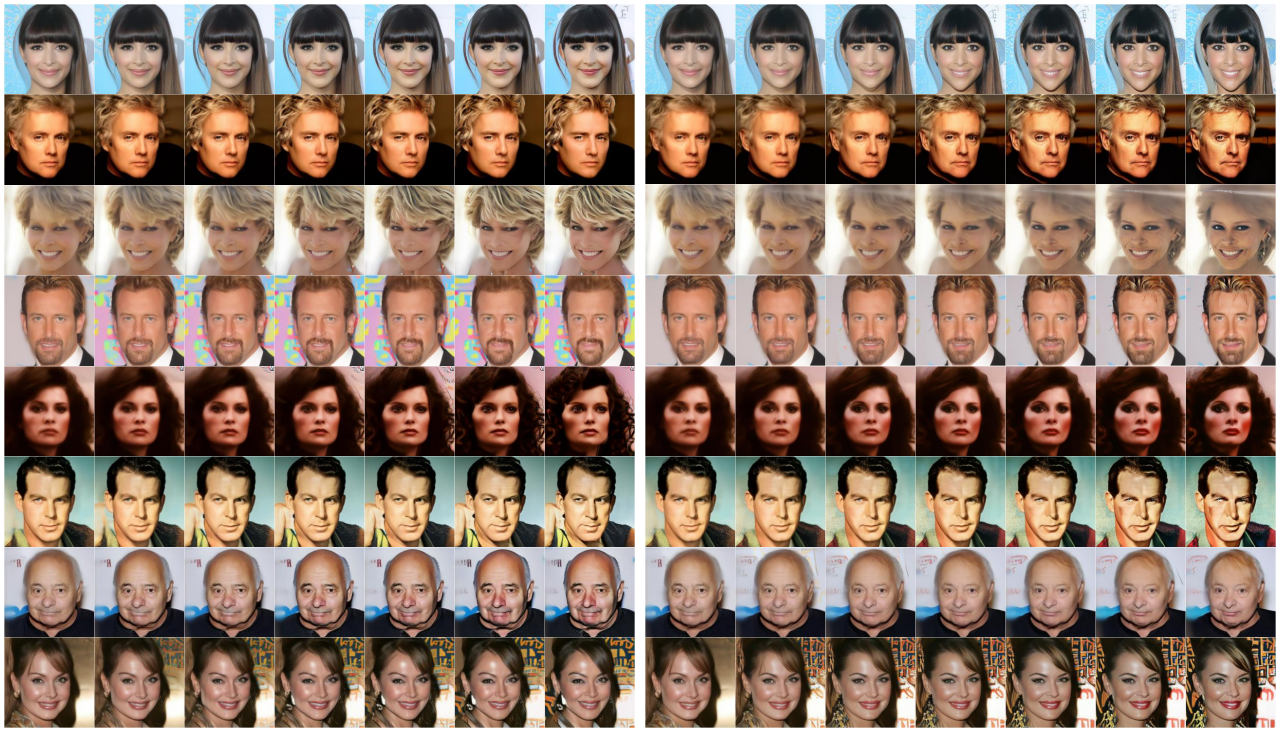


Figure A4. $t = 0.5T$. A selection of interpretable edits discovered by our feature direction in CelebA-HQ. The image on the far left represents the reconstructed original image, while the subsequent images demonstrate the interpretable edits that have been made to it.

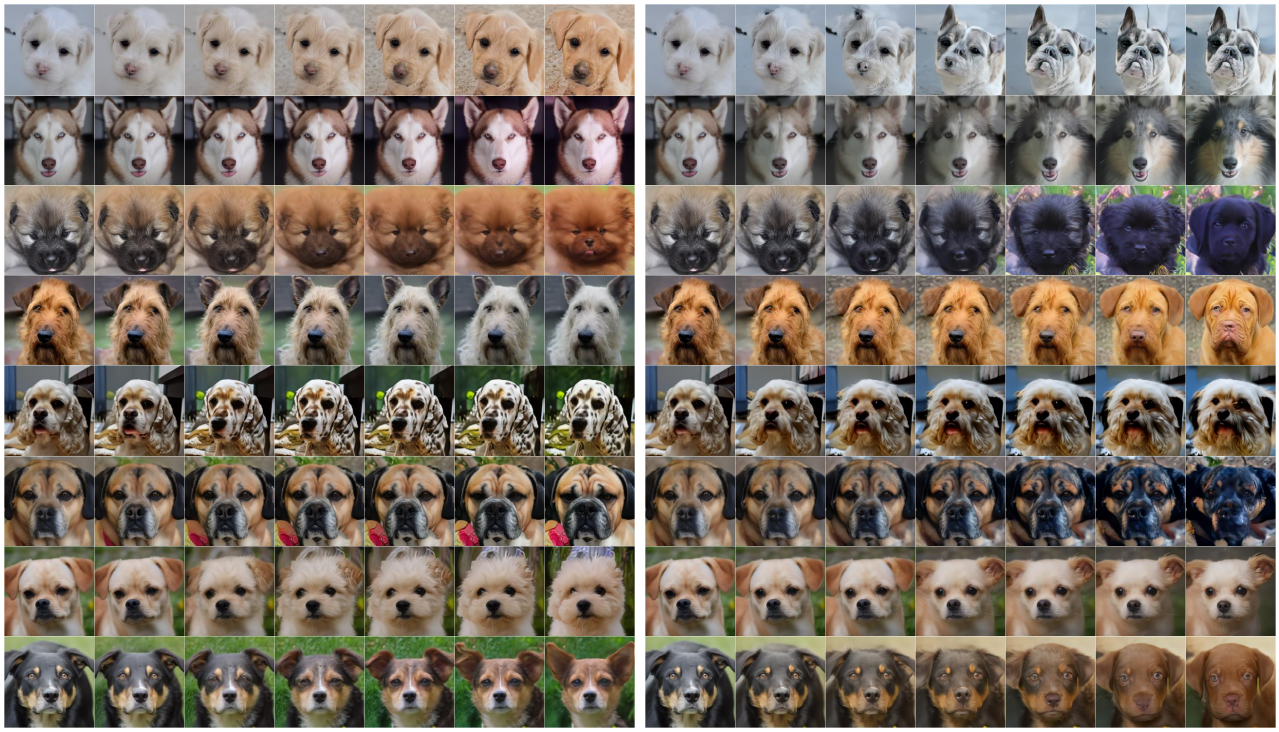


Figure A5. $t = T$. A selection of interpretable edits discovered by our feature direction in AFHQ. The image on the far left represents the reconstructed original image, while the subsequent images demonstrate the interpretable edits that have been made to it.

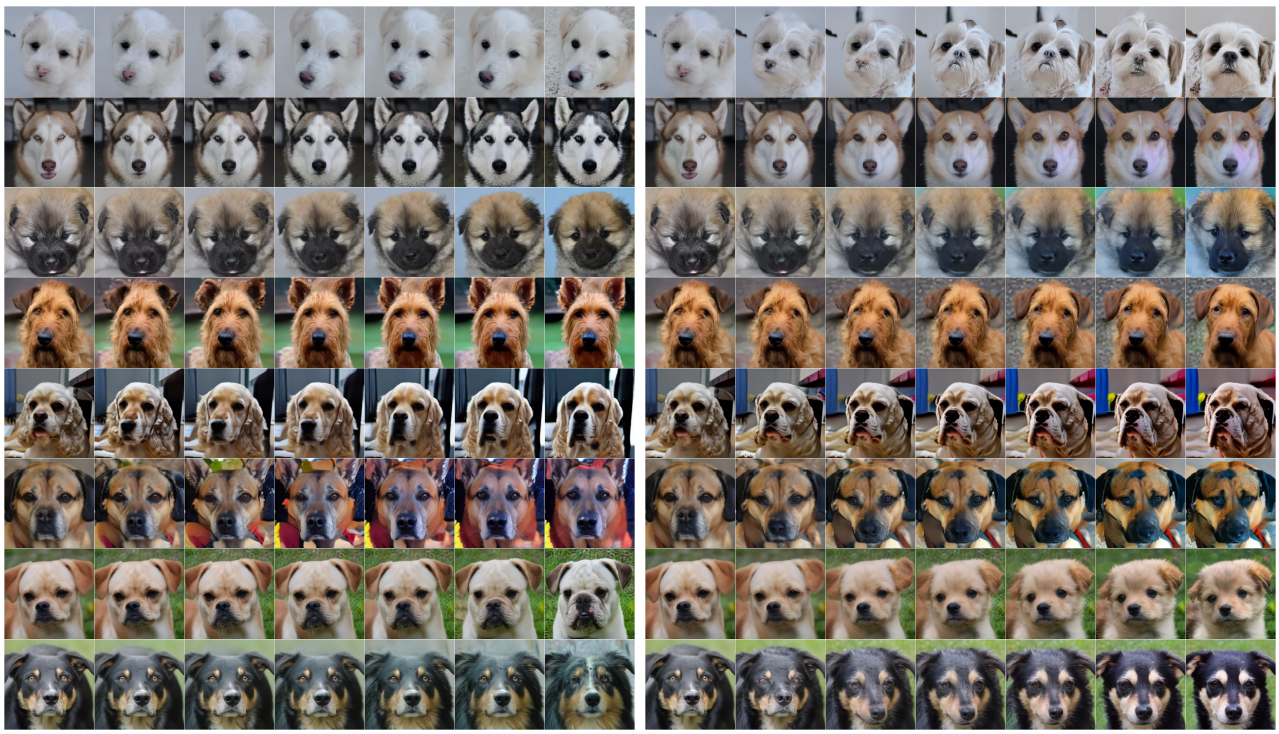


Figure A6. $t = 0.75T$. A selection of interpretable edits discovered by our feature direction in AFHQ. The image on the far left represents the reconstructed original image, while the subsequent images demonstrate the interpretable edits that have been made to it.

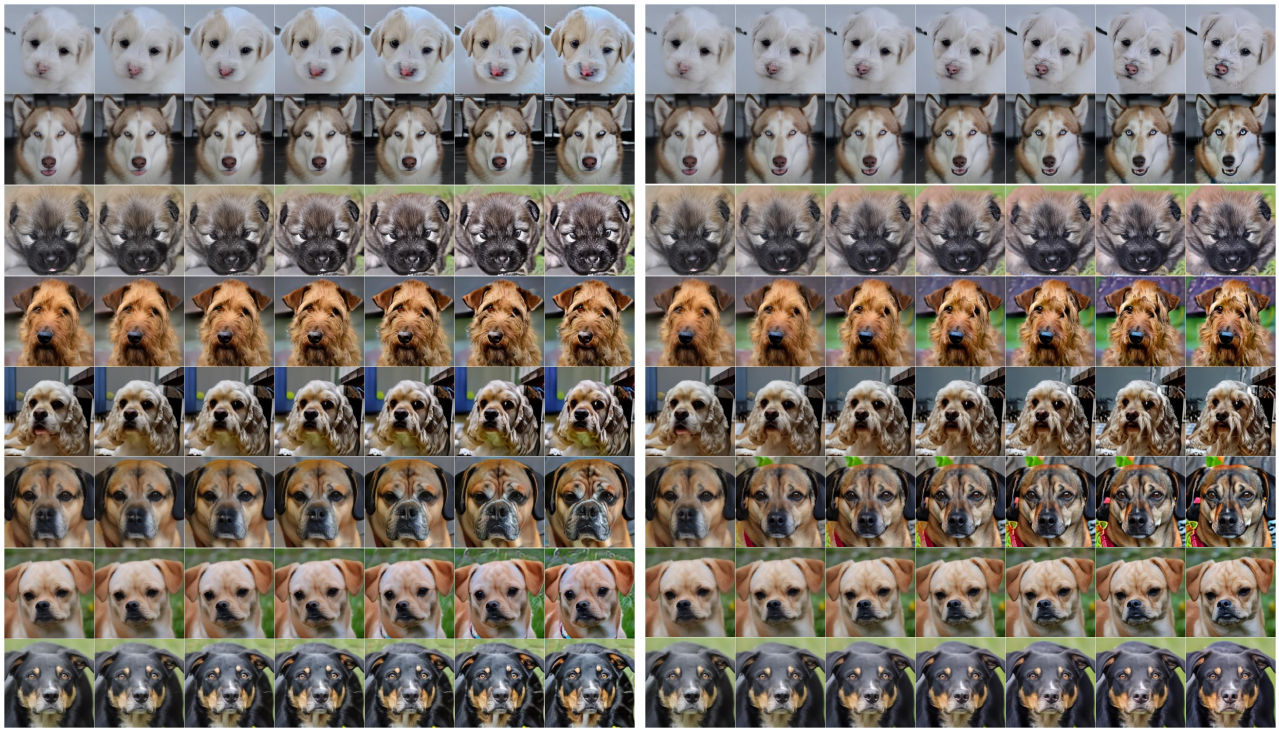


Figure A7. $t = 0.5T$. A selection of interpretable edits discovered by our feature direction in AFHQ. The image on the far left represents the reconstructed original image, while the subsequent images demonstrate the interpretable edits that have been made to it.

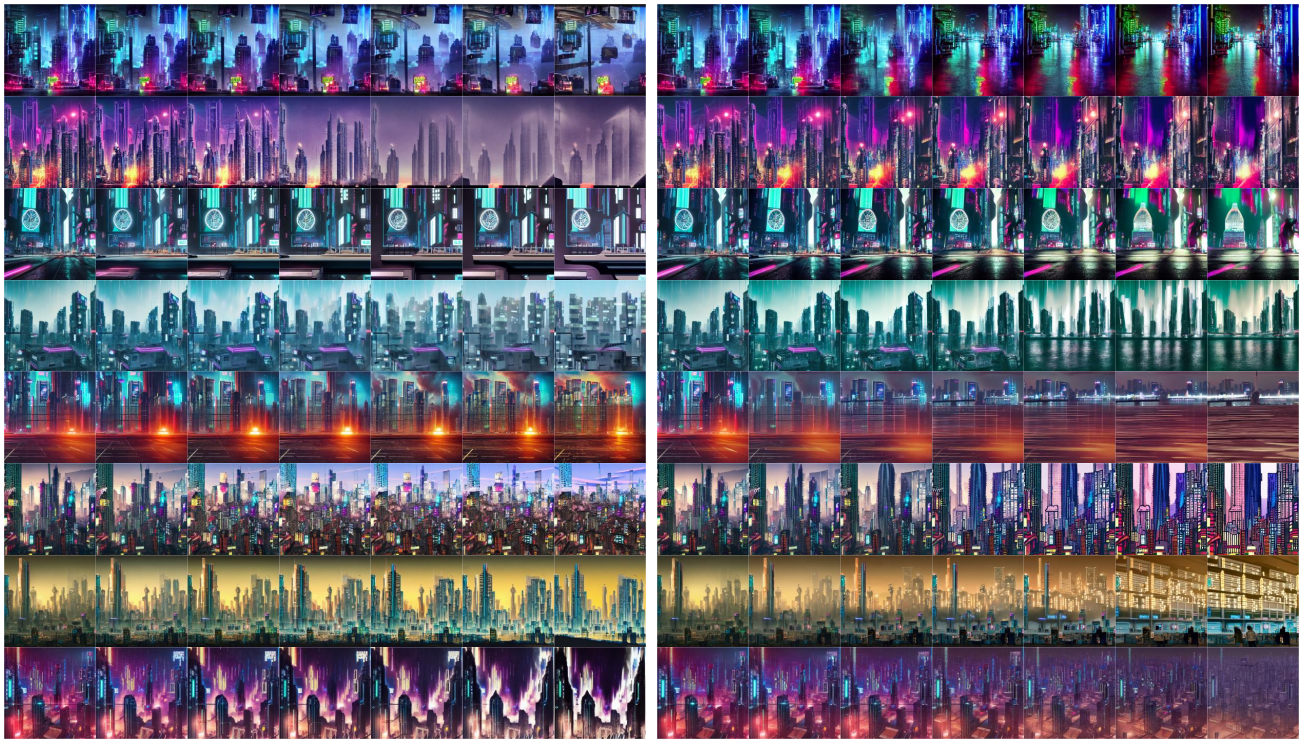


Figure A8. $t = T$. A selection of interpretable edits discovered by our feature direction in LDM with "Cyberpunk city". The image on the far left represents the reconstructed original image, while the subsequent images demonstrate the interpretable edits that have been made to it.

Figure A9. $t = T$. A selection of interpretable edits discovered by our feature direction in LDM with "Painting of VanGogh". The image on the far left represents the reconstructed original image, while the subsequent images demonstrate the interpretable edits that have been made to it.

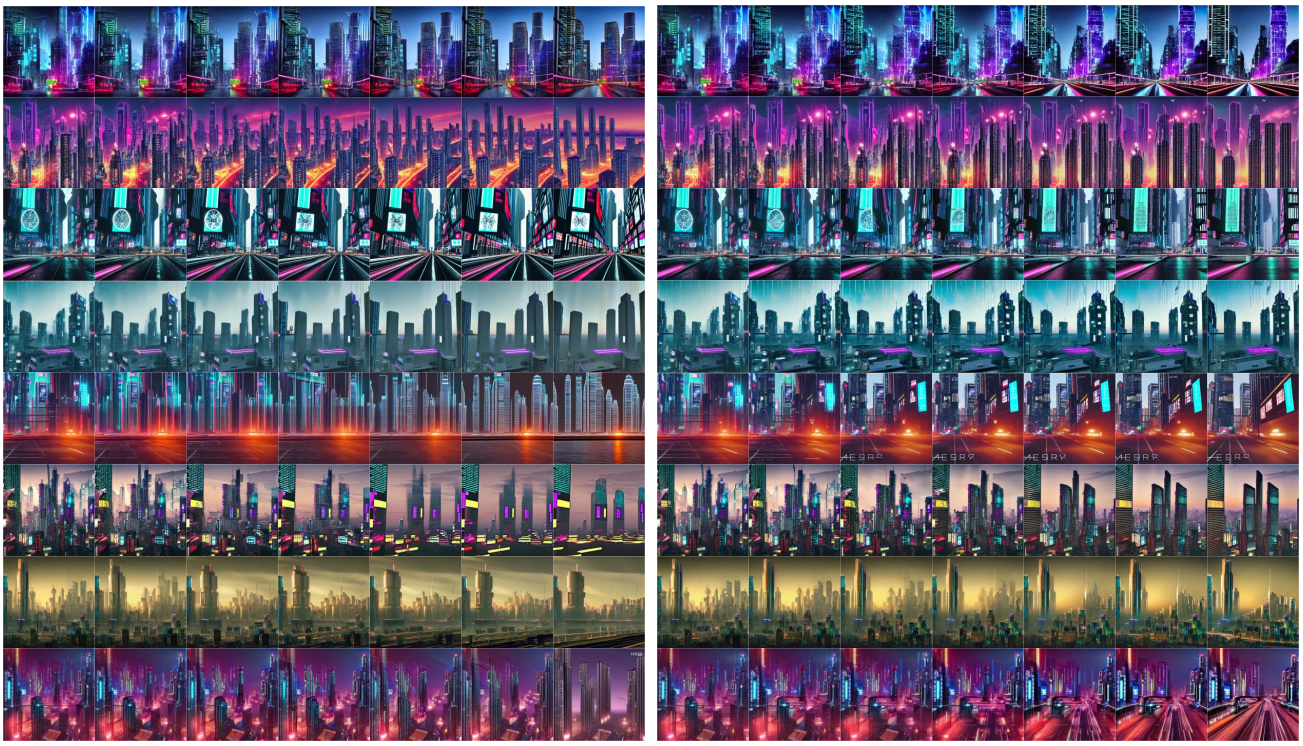


Figure A10. $t = 0.75T$. A selection of interpretable edits discovered by our feature direction in LDM with "Cyberpunk city". The image on the far left represents the reconstructed original image, while the subsequent images demonstrate the interpretable edits that have been made to it.

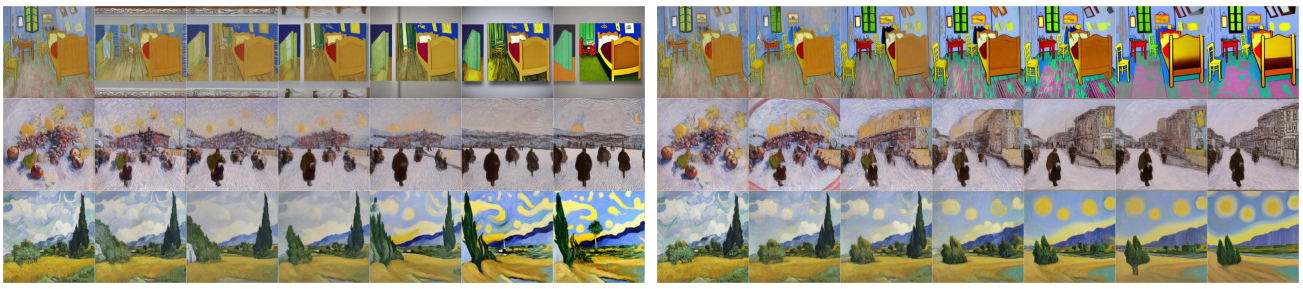


Figure A11. $t = 0.75T$. A selection of interpretable edits discovered by our feature direction in LDM with "Painting of VanGogh". The image on the far left represents the reconstructed original image, while the subsequent images demonstrate the interpretable edits that have been made to it.

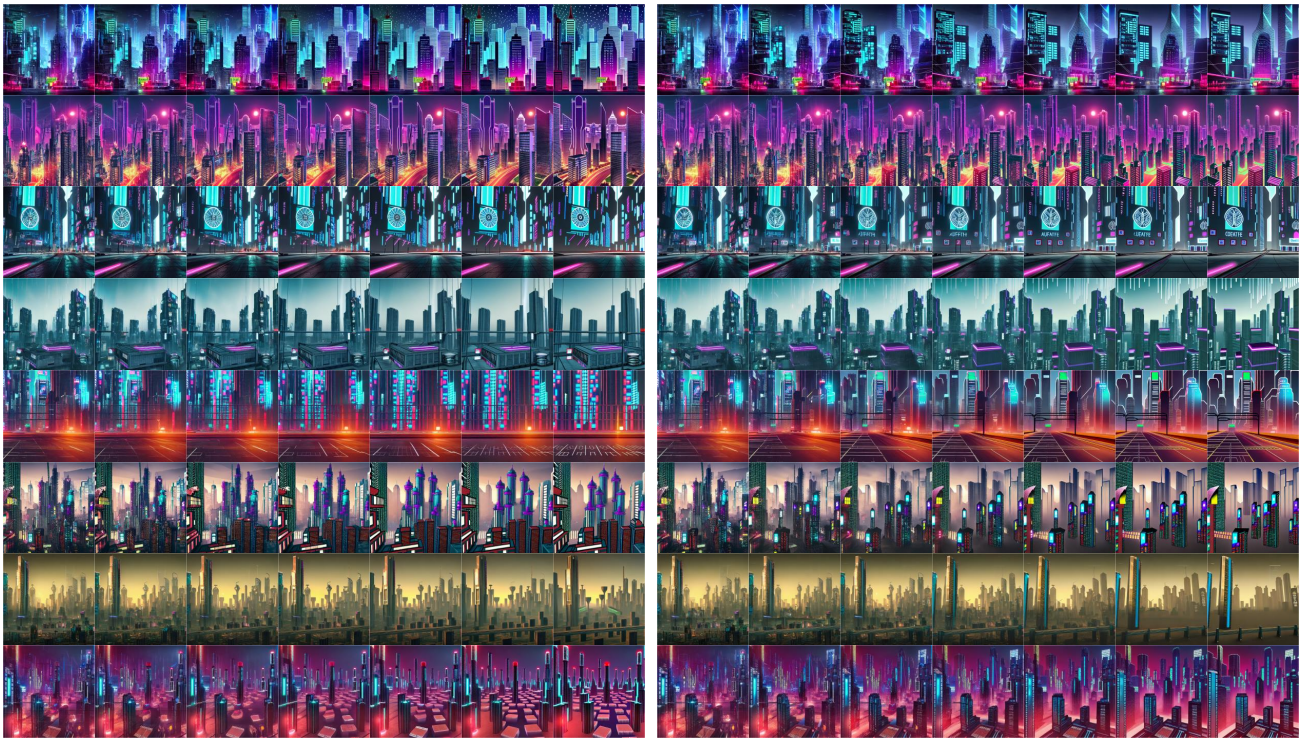


Figure A12. $t = 0.5T$. A selection of interpretable edits discovered by our feature direction in LDM with "Cyberpunk city". The image on the far left represents the reconstructed original image, while the subsequent images demonstrate the interpretable edits that have been made to it.

Figure A13. $t = 0.5T$. A selection of interpretable edits discovered by our feature direction in LDM with "Painting of VanGogh". The image on the far left represents the reconstructed original image, while the subsequent images demonstrate the interpretable edits that have been made to it.

E.2. Global feature direction

Figure A14. A selection of interpretable edits discovered by our global feature direction in CelebA-HQ. The image on the far left represents the reconstructed original image, while the subsequent images demonstrate the interpretable edits that have been made to it.

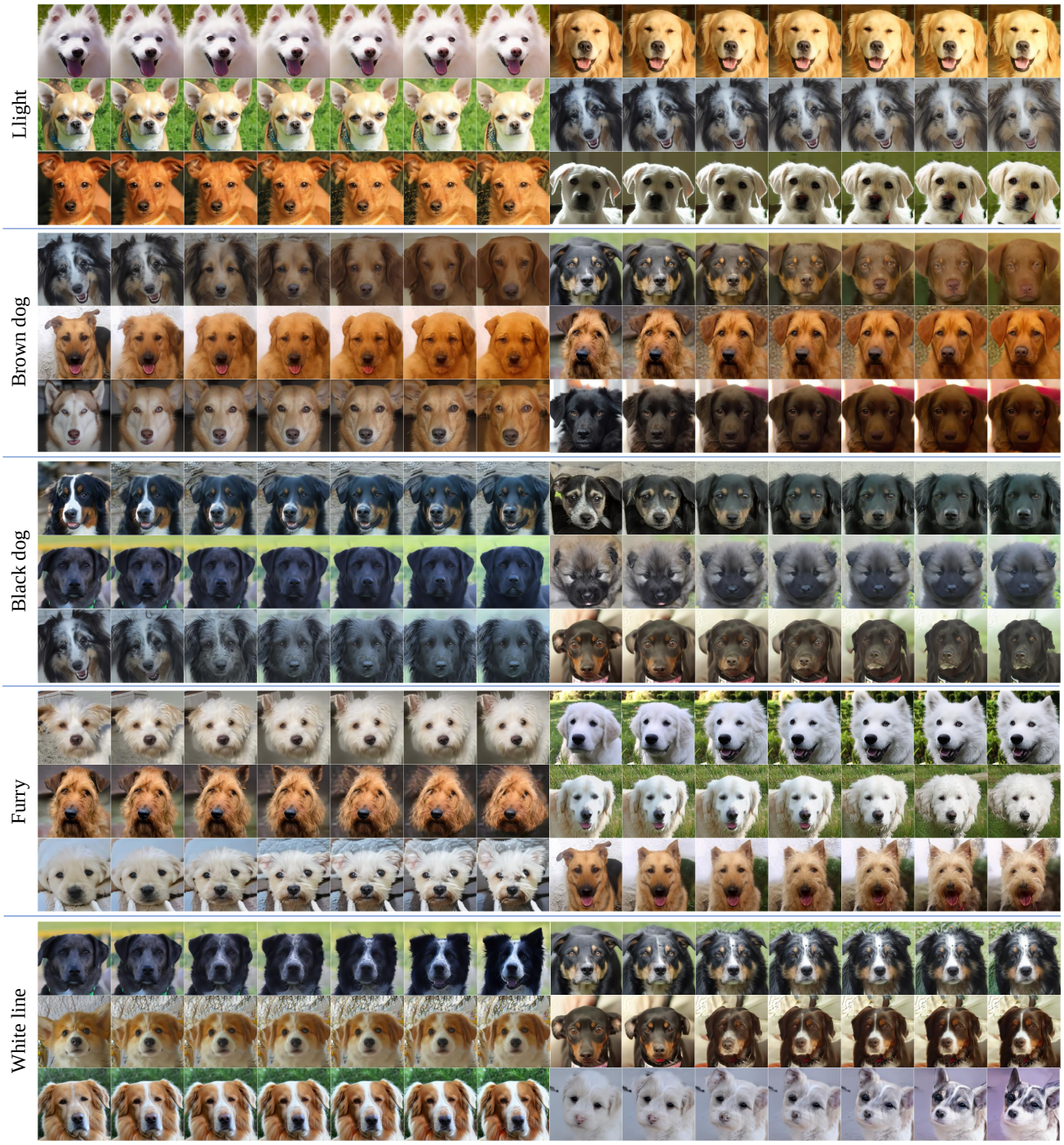


Figure A15. A selection of interpretable edits discovered by our global feature direction in AFHQ. The image on the far left represents the reconstructed original image, while the subsequent images demonstrate the interpretable edits that have been made to it.

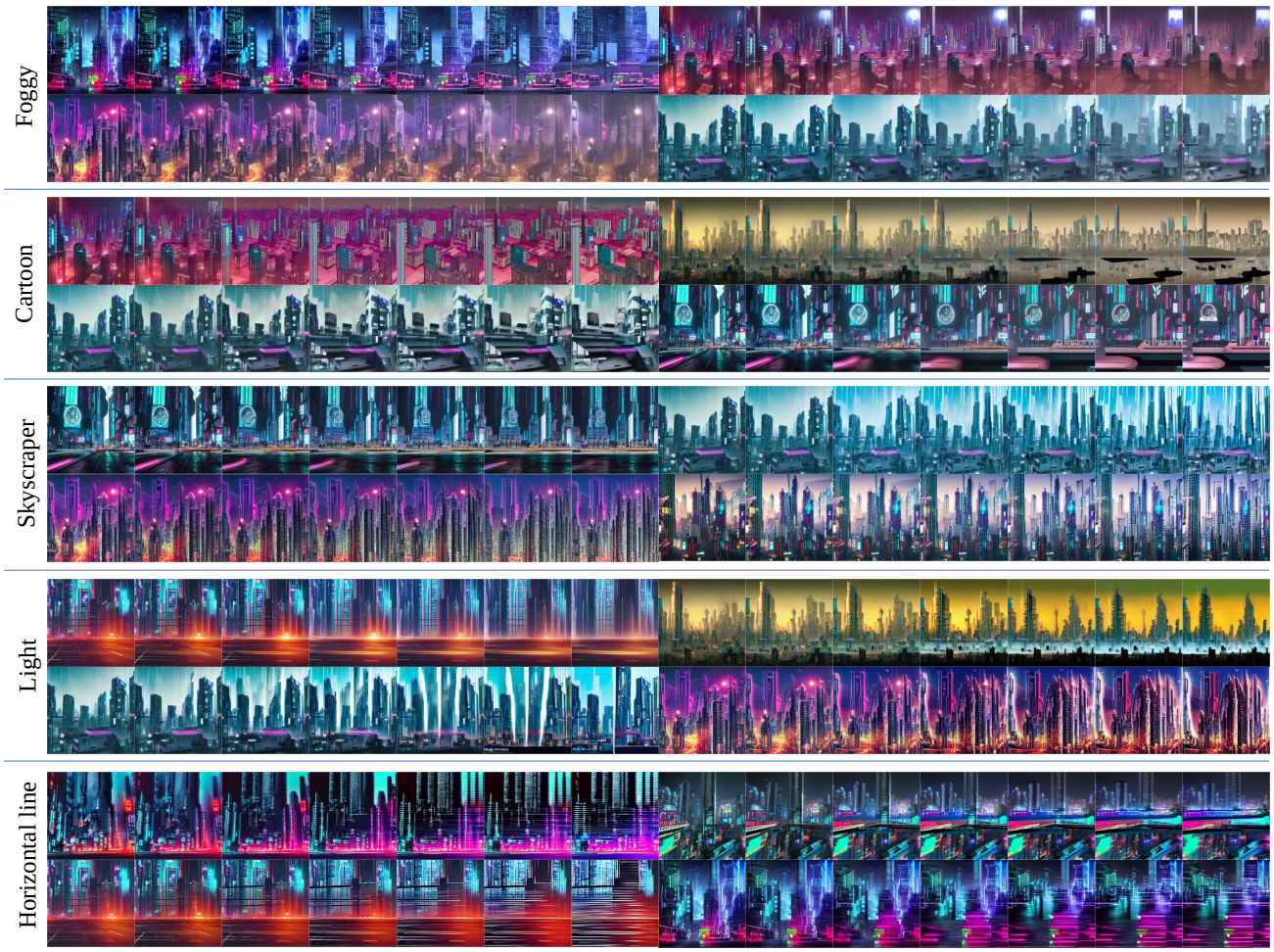


Figure A16. A selection of interpretable edits discovered by our global feature direction in LDM with ”Cyberpunk city”. The image on the far left represents the reconstructed original image, while the subsequent images demonstrate the interpretable edits that have been made to it.