Understanding the Generalization of Pretrained Diffusion Models on Out-of-Distribution Data

Sai Niranjan Ramachandran[1*†], **Rudrabha Mukhopadhyay**[2*], **Madhav Agarwal**[2*], **C.V. Jawahar**[2], **Vinay Namboodiri**[3]

[1]Indian Institute of Science, Bangalore
[2]International Institute of Information Technology, Hyderabad
[3]University of Bath

Abstract

This work tackles the important task of understanding out-of-distribution behavior in two prominent types of generative models, i.e., GANs and Diffusion models. Understanding this behavior is crucial in understanding their broader utility and risks as these systems are increasingly deployed in our daily lives. Our first contribution is demonstrating that diffusion spaces outperform GANs' latent spaces in inverting high-quality OOD images. We also provide a theoretical analysis attributing this to the lack of prior holes in diffusion spaces. Our second significant contribution is to provide a theoretical hypothesis that diffusion spaces can be projected onto a bounded hypersphere, enabling image manipulation through geodesic traversal between inverted images. Our analysis shows that different geodesics share common attributes for the same manipulation, which we leverage to perform various image manipulations. We conduct thorough empirical evaluations to support and validate our claims. Finally, our third and final contribution introduces a novel approach to the few-shot sampling for out-of-distribution data by inverting a few images to sample from the cluster formed by the inverted latents. The proposed technique achieves state-of-the-art results for the few-shot generation task in terms of image quality. Our research underscores the promise of diffusion spaces in out-of-distribution imaging and offers avenues for further exploration. Please find more details about our project at http://cvit.iiit.ac.in/research/projects/cvit-projects/diffusionOOD

Introduction

In the last decade, generative models such as Generative Adversarial Networks (GANs) (Goodfellow et al. 2020; Karras, Laine, and Aila 2019; Karras et al. 2020, 2021; Brock, Donahue, and Simonyan 2019) and Diffusion models (Preechakul et al. 2022; Sinha* et al. 2021; Rombach et al. 2021a; Ho et al. 2022) have significantly advanced image synthesis. These models excel in generating realistic images, but their handling of out-of-distribution (OOD) data — data that is substantially different from their training sets remains a challenge. OOD robustness is essential in deep learning for applications with limited resources, and it helps address issues like bias and distribution shifts. For example an autonomous vehicle trained on a dataset with limited extreme weather images must still perform safely under such conditions.

The research community has extensively analyzed and used the StyleGAN2 model (Karras et al. 2020) for various downstream tasks in particular for several face-related tasks using the StyleGAN2 trained on the FFHQ dataset (Karras et al. 2020). Its latent space is known for offering control over diverse aspects like facial expressions and hairstyles. Despite its strengths, StyleGAN2's limitations become evident with non-face images or faces that do not align with the FFHQ dataset's characteristics which are out-of-distribution (OOD) examples for StyleGAN. When applied to OOD images such as animals or inanimate objects, StyleGAN's image generation quality deteriorates. This limitation points to the challenges in training models on non-face datasets, given the high data demands and slow convergence. Thus exploring generative models that can better handle a variety of OOD images is critical for progress in fields like computer vision, graphics, and entertainment.

Our Contributions Diffusion models, known for their mathematical clarity and probabilistic insight, have prompted us to explore their potential in handling out-of-distribution (OOD) images compared to GANs' latent spaces. The main contributions of our work are as follows: **(1)** Empirical analysis reveals that diffusion models outperform GANs in inverting high-quality OOD images. This superiority is theoretically linked to the absence of "prior holes" in diffusion spaces, a flaw seen in GANs where inverted latents' distribution misaligns with the actual distribution. **(2)** We hypothesize that the Gaussian structure of diffusion spaces enables projection onto a hypersphere, streamlining the traversal between inverted images via geodesics as well as non-linear traversal. We present diverse image manipulations, assessing the universality of different diffusion spaces. **(3)** Capitalizing on the geometry of diffusion spaces, we introduce a method for few-shot generation, showcasing state-of-the-art results. In summary, our work underscores the advantages of diffusion models' latent spaces in addressing the challenges posed by GANs for

*These authors contributed equally.
†This work was done as a research assistant at International Institute of Information Technology, Hyderabad

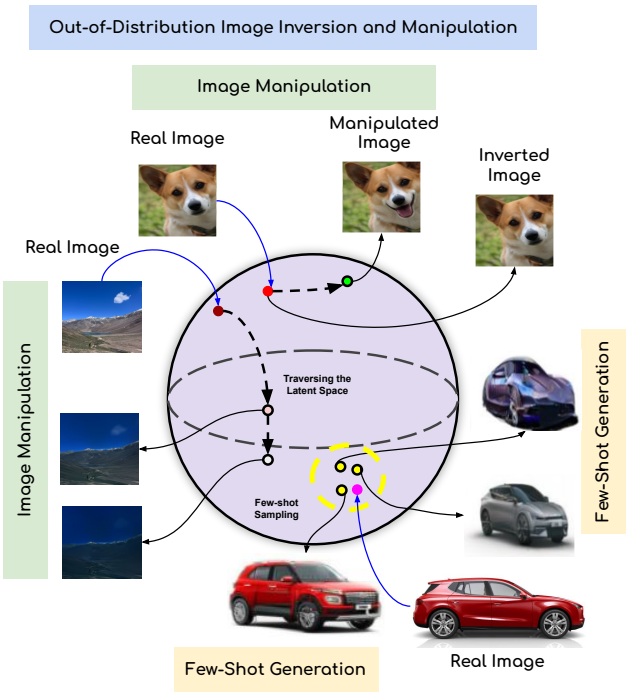

Figure 1: We hypothesize that diffusion spaces match the prior distribution perfectly and are devoid of any "prior holes" helping in out-of-distribution inversion. Their spherical, Gaussian nature facilitates traversal, aiding in OOD image manipulation. Their semantic richness leads to clustering in latent spaces, enabling few-shot image generation.

OOD images. Figure 1 illustrates an illustrative summary of our findings.

Background

GAN dissection and analyzing latent space of GANs A slew of research has delved into understanding GANs' latent spaces. Härkönen et al.'s GANSpace (Härkönen et al. 2020) offers interpretable controls for image synthesis, while Bau et al.'s framework (Bau et al. 2018) provides insights into GAN layers. Others have dissected GAN architectures to grasp their intricacies (Karras et al. 2019; Voynov and Babenko 2020; Shen et al. 2019). The power of GANs is partly attributed to their use of latent spaces (Radford, Metz, and Chintala 2015), yet their latent spaces exhibit flaws, especially with GAN-inversion on out-of-distribution data (Abdal, Qin, and Wonka 2019). This spurred developments like (Tov et al. 2021; Abdal et al. 2021; Shen and Zhou 2020; Subramanyam et al. 2022), with SPHInX (Subramanyam et al. 2022) being a standout for OOD image inversion in StyleGAN, serving as our study's baseline.

Diffusion Models Diffusion models,construct a Markov chain of noising steps, termed the "forward process," to progressively add noise until the data resembles noise. The learnt "backward process" reverses this to generate desired data samples from the noise. Compared to GANs and VAEs, these models yield superior samples and cover broader distribution modes (Xiao, Kreis, and Vahdat 2021; Dhariwal and Nichol 2021). They typically employ a Gaussian Diffusion process for sampling (Sinha* et al. 2021; Preechakul et al. 2022; Rombach et al. 2021a; Dhariwal and Nichol 2021; Xiao, Kreis, and Vahdat 2021; Song et al. 2021a; Bansal et al. 2022). We explore how diffusion models facilitate OOD inversion and image manipulation.

Multiple diffusion models covered in this work We posit that if forward processes in diffusion models share similarities, specific attributes should consistently manifest across diverse architectures, as the model dynamics are unchanged (Song et al. 2021b). We investigate the behavior of three different diffusion models: Diffusion Autoencoders (Preechakul et al. 2022) (DAE), Diffusion-Denoising Models for Few-shot Conditional Generation (Sinha* et al. 2021) (D2C), and Latent Diffusion Model (Rombach et al. 2021a) (LDM). DAE and D2C are both autoencoder-based architectures that utilize diffusion processes for encoding and decoding. DAE employs an additional semantic encoder network to condition the diffusion process, while D2C fits a conditional diffusion model on the latent space of a pretrained VAE. LDM is the state-of-the-art among diffusion models and is the backbone for many popular variants, such as Stable Diffusion (Rombach et al. 2021b). Note that while we do not analyze stable diffusion as it is trained on the very large LAION dataset (Schuhmann et al. 2021), we do analyze their unconditional latent diffusion model trained on a specific dataset, FFHQ, which we denote as LDM (FFHQ). Our OOD analysis results, therefore, should also apply to the stable diffusion model.

Inverting Out-of-Distribution Images

We use model-specific inversion techniques for inverting images onto the diffusion space. We obtain DAE's semantic and stochastic latents by following the procedure outlined in (Preechakul et al. 2022). For D2C, we use the pretrained variational encoder presented in (Sinha* et al. 2021) to extract features from the input image. Subsequently, we pass the features through the pretrained conditional latent diffusion process, where we set the condition value to be 0 to reflect that we need an unmodified image. For LDM (Rombach et al. 2021a), we pass the input image through the pretrained vector quantized (van den Oord, Vinyals, and Kavukcuoglu 2017) encoder and then apply the forward process to obtain the corresponding latent. We can reconstruct the output image by using the conditional reverse process for DAE (Preechakul et al. 2022) or a decoding step in the case of D2C (Sinha* et al. 2021) and LDM (Rombach et al. 2021a).

What do we consider as out-of-distribution? For a given dataset $\mathcal{D}$ on which a model is trained, we introduce the concept of an *extended in-distribution dataset*. This is formed by augmenting $\mathcal{D}$ with images of analogous characteristics, such as similar poses and alignments. For models trained on the FFHQ dataset (Karras et al. 2020), we add the dataset with images matching FFHQ's preprocessing criteria, including those from CelebA (Liu et al. 2015). Images not ad-

hering to FFHQ's properties are deemed out-of-distribution (OOD). While pretrained frameworks like StyleGAN (Karras, Laine, and Aila 2019; Karras et al. 2020) proficiently invert in-distribution images, they falter with non-standard face images or non-face visuals. To precisely categorize OOD datasets, we employ the "Proxy A-Distance" (PAD) metric (Ben-David et al. 2006). CelebA, aligned with FFHQ, has a PAD of 1.39. In contrast, diverse datasets like ImageNet (Deng et al. 2009) and ImageNet Sketches (Wang et al. 2019) register PAD values between 1.98 and 2.28, marking them as ideal OOD examples. Based on our analysis, we suggest a PAD threshold of 1.50 to label a dataset as OOD. Refer to Table 1 for PAD values related to FFHQ.

Dataset	PAD
ImageNet (Deng et al. 2009)	1.98
ImageNet Sketches (Wang et al. 2019)	2.28
DSPRITES (Matthey et al. 2017)	2.26
QuickDraw (Ha and Eck 2017)	2.28
CelebA (Liu et al. 2015)	1.39

Table 1: Proxy A-Distance values for various datasets with respect to FFHQ (Karras, Laine, and Aila 2019). For our analysis, a Proxy A-Distance value above 1.50 is considered indicative of Out-of-Distribution (OOD) data, as evidenced by the values in this table.

Comparison with state-of-the-art out-of-distribution inversion in GANs We establish multiple baselines. Our primary candidate for comparison is SPHInX (Subramanyam et al. 2022), the current SOTA in OOD inversion. However, we select several different GANs instead of relying solely on StyleGAN to obtain a more comprehensive understanding of GAN behavior. The chosen models are BigGAN (Brock, Donahue, and Simonyan 2019) and StyleGANXL (Sauer, Schwarz, and Geiger 2022), trained on ImageNet that can generate diverse natural images. Although ImageNet is now included in the distribution for these two models, the other selected datasets remain out of distribution. These models also have varied architectures, enabling us to evaluate the generic behavior of GANs. For StyleGANXL (Sauer, Schwarz, and Geiger 2022) and BigGAN (Brock, Donahue, and Simonyan 2019), we use a simple optimization-based inversion method and set a uniform number of 2000 iterations for each technique. We evaluate the reconstruction quality using the PSNR and SSIM (Wang et al. 2004) metrics.

Empirical observations Table 2 and Figure 2 show that diffusion models achieve high-quality OOD inversion with high PSNR and SSIM values. On the other hand, we observe a significant drop in quality for OOD inversion using GAN-based models consistently across datasets. We also observe the current state-of-the-art GAN-based OOD inversion, SPHInX (Subramanyam et al. 2022), to perform relatively well among the GAN inversion baselines. However, even SPHInX consistently generates inferior results compared to all the diffusion-based inversion techniques across all the datasets. Interestingly, in the case of BigGAN and StyleGANXL, even in-distribution inversion is hard, and their performance is inferior to that of the diffusion models. Empirical evidence suggests that GAN Inversion struggles more with sketch-like datasets such as QuickDraw (Ha and Eck 2017) and ImageNet Sketches (Wang et al. 2019) as these datasets require a large number of iterations to obtain visually meaningful content. As optimization is necessary to achieve the best results, GAN Inversion is significantly slower than inversion in diffusion models. Overall, we provide strong empirical evidence on diffusion models' ability to invert and successfully represent OOD images in their respective latent spaces. Due to limited space, only the results obtained from DAE are presented visually in the main paper, as it achieved the highest quantitative scores for OOD inversion. For additional information, kindly refer to our project page, the link to which can be found in the abstract. But why do diffusion models perform better in representing OOD images, and what could prevent GANs from doing the same? We justify this theoretically below.

Why Does OOD Inversion Perform Better for Diffusion Models but Is Objectively Much Harder for GANs?

We attempt to justify this phenomenon by analyzing the inherent geometry of the latent space of diffusion models and GANs. We assume that our set of images drawn from diverse datasets is a subset of the distribution of "all natural images". From the manifold hypothesis (De Bortoli 2022), we assume this lies on a low-dimensional manifold. This dimension is often much smaller than the dimension of latent spaces used in most generative models (Pope et al. 2021; Sauer, Schwarz, and Geiger 2022). Our equivalent problem is whether a given diffusion model can embed a given low-dimensional manifold. As (De Bortoli 2022) shows that convergence holds in the 1-Wasserstein metric for the SDE formulation used in Equation 1 i.e., we can embed low dimensional manifolds whose dimension is smaller than the latent dimension and theoretically recover the said distribution. This is a plausible justification for the models' capability of inverting OOD images.

$$dx = f(x,t)dt + g(t)dw \tag{1}$$

Mathematically defining inversion The problem of inversion involves finding the latent representation z corresponding to a given image $x \in \mathbb{R}^{H\times W}$ that best approximates the given image in the space of some generative model G. In general, if we denote the latent space of a model by $\mathcal{Z}$ then z is the solution to the constrained optimization problem given by $z = argmin\|G(z) - x\|_R, z \in \mathcal{Z}$, where R. is some suitable reconstruction metric. Diffusion models, on the other hand, typically obtain z using the forward process, whereas GANs lack an explicit mechanism to get z, making the resulting optimization problem practically challenging to solve. Our empirical observations show that the inverted distributions learned by GANs often do not match the prior of G, which limits their usefulness for downstream tasks. This discrepancy between the learned and actual priors is the crux of the difference in inversion settings of GAN and diffusion models.

	ImageNet		ImageNet Sketches		DSPRITES		QuickDraw	
Models	**PSNR (↑)**	**SSIM (↑)**	**PSNR (↑)**	**SSIM (↑)**	**PSNR (↑)**	**SSIM (↑)**	**PSNR (↑)**	**SSIM (↑)**
SPHInX (FFHQ)	27.63	0.8924	21.79	0.7708	27.69	0.9013	23.06	0.8016
StyleGANXL (ImageNet)	26.83	0.8718	20.85	0.6924	24.87	0.8517	22.68	0.7963
BigGAN (ImageNet)	23.37	0.8126	19.63	0.7015	24.72	0.8542	22.23	0.7924
DAE (FFHQ)	**32.14**	**0.9823**	**31.77**	**0.9587**	**32.09**	**0.9662**	**32.04**	**0.9645**
D2C (FFHQ)	31.68	0.9564	30.82	0.9413	32.03	0.9634	31.86	0.9592
LDM (FFHQ)	31.84	0.962	30.96	0.9431	31.67	0.957	32.01	0.9627

Table 2: We compare OOD image inversion for multiple models across multiple datasets. The training dataset for each model is also reported within brackets. We observe diffusion-based models to outperform GANs for OOD inversion significantly.

Defining prior holes We define an encoder model as $E = q_\phi(\mathbf{z}|\mathbf{x})$ (E represents any process used to invert an image, whether by optimization or by leveraging a trained network etc.), and a prior model as $p_\theta(\mathbf{z})$, where $\mathbf{z}$ is a latent variable and ϕ and θ are parameters. The issue of mismatch arises between $q_\phi(\mathbf{z})$ and $p_\theta(\mathbf{z})$, where q_ϕ represents the aggregate posterior, which is defined as $q_\phi(\mathbf{z}) = \mathbb{E}_{\mathbf{p_{data}}(\mathbf{x})}[\mathbf{q}_\phi(\mathbf{z}|\mathbf{x})]$. The mismatch between the prior and aggregate posterior distributions in GANs can result in "holes" in the prior that the aggregate posterior fails to cover. This can lead to worse generation quality as the inversion process may traverse through points that do not lie in the manifold. Indeed, as (Choi et al. 2022) shows, deviations from the manifold result in poor visual quality for StyleGAN. *We formally define the prior hole as follows: let $p(\mathbf{z})$ and $q(\mathbf{z})$ be two distributions with $supp(q) \subseteq supp(p)$ such that the probability measures are well defined. q has an (ϵ, δ) prior hole with respect to p for $\epsilon \in (0, 1)$ and $\delta \in (\epsilon, 1)$ if there exists an $S \in supp(p)$ such that $\int_S p(\mathbf{z})\mathbf{dz} \geq \delta$ and $\int_S q(\mathbf{z})\mathbf{dz} \leq \epsilon$.* In other words, the probability mass of the aggregate posterior falls short of the probability mass of the prior in certain regions if prior holes exist.

Observation regarding Prior Holes in (Sinha* et al. 2021) *Let $p_\theta(z) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. For any $\epsilon > 0$, $\exists$ a distribution $q_\phi(\mathbf{z})$ for any $\epsilon > 0, \delta < 0.5$ with an (ϵ, δ) prior hole such that $D_{KL}(q_\phi||p_\theta)$ and $W_2(q_\phi, p_\theta) < \gamma$ for any $\gamma > 0$.* Here, D_{KL} refers to the KL Divergence, and W_2 refers to the 2-Wasserstein distance. Thus, divergence objectives fail to solve the prior holes for normal priors. Therefore, inverted embeddings are not guaranteed to cover the prior if the prior is an isotropic Gaussian, which is the latent prior assumed for various GAN models. In this work, we extend this to a broad class of priors by the following hypothesis to reason the generic difficulty of GAN-inversion.Such a scenario becomes important for latent spaces such as $\mathcal{W}, \mathcal{W}^+$ in the StyleGAN family (Karras, Laine, and Aila 2019; Karras et al. 2020, 2021).

Lemma 1: Our extended hypothesis on prior holes for GAN-inversion Let $p_\theta(z)$ be any prior distribution such that p_θ is *absolutely continuous* over $\mathbb{R}^n$. For any $\epsilon > 0$, $\exists$ a distribution $q_\phi(\mathbf{z})$ for any $\epsilon > 0, \delta < 0.5$ with an (ϵ, δ) prior hole such that $D_{KL}(q_\phi||p_\theta)$ and $W_2(q_\phi, p_\theta) < \gamma$ for any $\gamma > 0$. The intuition behind this hypothesis follows from the fact that we can obtain a large class of priors from the Isotropic Gaussian using a simple transformation. Therefore, we can now reason why inversion is hard in models like StyleGAN, which can easily generate high-quality in-distribution images. While in GANs, posterior and prior matching ensures sampling is good, learning an encoder or even directly optimizing a latent for inversion is hard as the "embedded" latent space might be forming holes in the prior.

Lemma 1.1: Prior Holes are eliminated in Diffusion Models We now consider the case of diffusion models and claim that prior holes are eliminated for diffusion model-based inversion by construction. The prior for a diffusion model can be regarded as $\mathcal{N}(\mathbf{0}, \mathbf{I})$ as this is the distribution used for sampling. From our discussion earlier, the analog to the encoder, in this case, is the forward process. Observe that as (Song et al. 2021a) show, the forward diffusion process can be represented in terms of a Stochastic Differential Equation (SDE) (Song et al. 2021a; De Bortoli 2022) according to Equation 1. The SDE describing the diffusion process of $\mathbf{x}(\mathbf{t})$ is a standard Wiener process $\mathbf{w}$ with the drift coefficient $\mathbf{f}(, \mathbf{t})$, and the diffusion coefficient $g : \mathbb{R} \to \mathbb{R}$. Due to the process's nature, as demonstrated by (Song et al. 2021a), the forward process converges to $\mathcal{N}(\mathbf{0}, \mathbf{I})$ as $t \to \infty$, given a small enough step size. Therefore, we can conclude that the learned distribution matches precisely with the prior distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, which is used for sampling. Thus, our analysis offers a compelling explanation for the empirical results observed in our study.

Figure 2: We show inversion results on OOD images from different datasets. The figure shows that the diffusion-based technique outperforms the SOTA GAN-based OOD inversion techniques.

Traversing the Latent Space To Perform Different Image Manipulations

Diffusion models, such as Imagen (Saharia et al. 2022), DALL-E2 (et al 2022), and Stable Diffusion (Rombach et al. 2021b,a), have shown prowess in editing tasks, especially text-based editing. Diffusion-CLIP (Kim, Kwon, and Ye 2022) offers an advanced method using clip loss for text-conditioned image editing, while other image-to-image translation approaches (Saharia et al. 2021) condition the generation on various input types. *This raises the question of why the diffusion space is so amenable to editing.* We believe that the answer lies in the diffusion space's geometric properties that help traverse different points in the latent space. We observe that an exploration of these properties yields valuable insights. We have already highlighted the capability of different types of OOD images to be represented in a diffusion space. However, a common feature of such high-dimensional representation spaces is the "Curse of Dimensionality" (Zhao, Zhu, and Zhang 2019; Vershynin. 2015; Giné et al. 2006; Scott 2015), which results in increasingly sparse spaces with higher dimensions. Therefore, we must avoid the curse of dimensionality for diffusion spaces to ensure that distances between embeddings in our latent space are bounded. We present two lemmas in this section, guaranteeing that the distance between any two latents is always bounded.

Lemma 2.1: Sampling from a d dimensional Isotropic Gaussian can be approximated closely by sampling from a Uniform Prior on the $d-1$ dimensional hypersphere with radius $\sqrt{d}$. It is known that, that in high dimensions, the distribution of a normalized random vector sampled from the isotropic Gaussian $\frac{\mathbf{X}}{\|X\|}, \mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is Uniform on the unit hypersphere. Using concentration inequalities, we then derive the $\sqrt{d}$ bound.

Lemma 2.2: Given any two datasets of samples each, their Wasserstein distance is bounded in a spherical latent space and is of the order of as where r is the radius of the hypersphere This lemma proposes that the distance between two data points in the spherical space is always bounded implying that the distance behaves well as the dimension increases, allowing us to evade the curse of dimensionality. Consequently, we observe that the geometry of the latent space allows for two important properties: (1) a well-defined underlying geometry of the latent space and (2) all data points are located at a finite distance from each other. These properties suggest that it is possible to traverse from any given inverted image to any other inverted image using a geodesic while remaining on the manifold. Therefore, during manipulation, the quality of images can be maintained easily. In contrast, traversing in the $\mathcal{W}$ space of StyleGAN often results in low-quality images as we deviate from the manifold (Choi et al. 2022). *The key takeaway is that the bounded distance between embeddings in our well-parameterized latent space allows for efficient image manipulation and avoids issues that arise in other high-dimensional spaces.*

Utilizing the geometry of the latent space for traversal To understand how any transformation in an image is reflected in the latent space, we must examine the phenomenon of traversal. By starting with a given starting latent z_S and following a path γ, we can end up at a modified latent z_T. Our path may be along a geodesic (line segment) or non-linear if it does not follow any geodesic.

Geodesic interpolation We first formalize the geodesic interpolation in Algorithm 1 between a source image I_S and a target image I_T. The inversion and reconstruction operations are denoted as $Invert()$ and $Recon()$, respectively. The dot product is represented by $<>$, and Uni refers to the Uniform Distribution. The resulting interpolated image is denoted by I_{int}. We use this process to evaluate geometric interpolation between large sets of source images and their corresponding synthetic manipulations as targets. By interpolating through the geodesic that connects I_S and I_T, we observe that the generated images further away from z_S show a progressive change in the manipulation concerned, moving closer towards I_T. In other words, the resulting images display a gradation in their level of manipulation as we move along the geodesic.

Algorithm 1: Geodesic Interpolation

Data: $I_S, I_T \in \mathbb{R}^{H \times W}$
Result: I_{int}
$z_S \leftarrow Invert(I_S) \quad z_T \leftarrow Invert(I_T)$
$\phi \leftarrow cos^{-1}(< z_S, z_T >)$
$\alpha \sim Uni([0, 1])$
$z_{int} \leftarrow \frac{sin((1-\alpha)\phi)}{sin(\phi)} z_S + \frac{sin(\alpha\phi)}{sin(\phi)} z_T$
$I_{int} \leftarrow Recon(z_{int})$

Geodesics for the same manipulation are parallel across images To assess if a manipulation direction for one image pair generalizes to others, we must examine its applicability across multiple instances. If the direction shows a high degree of generality, it would suggest that the latent space is well-structured and different attributes of OOD images are disentangled in the latent space. This enables estimating a global manipulation direction using only a few examples and applying it to a more extensive set of images. We measure the degree of parallelism between the geodesics obtained by calculating their average angle. A value close to zero indicates a higher degree of parallelism and greater generality. We show the degree of parallelism for three kinds of manipulations: "white box effect in images", Gaussian blur, and gray-scaling images across a set of diffusion models using angles computed between directions in Table 3. In this case, an angle ≈ 0 implies that the directions are parallel, while a similar experiment using SPHInX (Subramanyam et al. 2022) in the StyleGAN latent space has a much larger angle. Therefore, we observe that diffusion models show a high degree of generality for manipulations in practice. The parallel directions for a particular manipulation in the latent space point toward the disentanglement of specific properties in the diffusion space for OOD images.

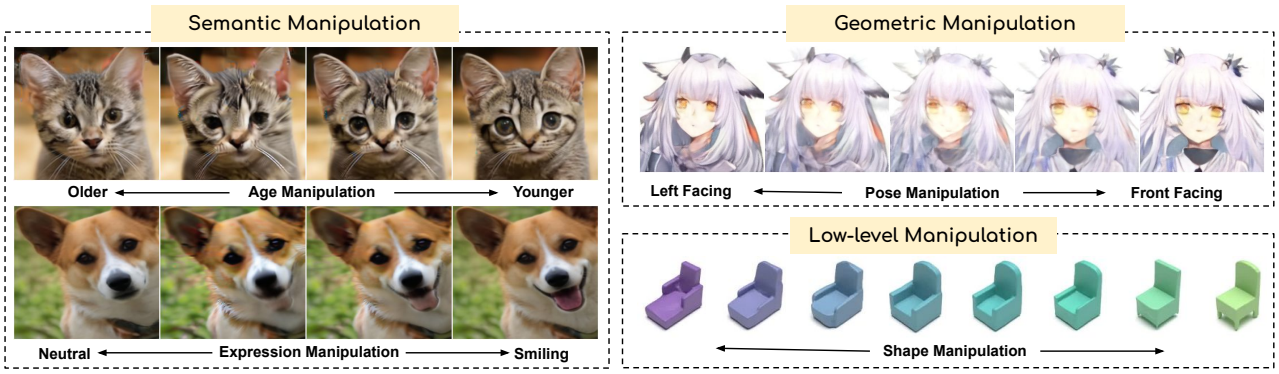

Figure 3: We show that linear traversal along a geodesic in diffusion space yields high-quality results for different manipulations.

Models	Gaussian Blur	Grayscale	White box
SPHInX (FFHQ)	1.12	0.89	1.05
DAE (FFHQ)	0.09	**0.07**	0.08
D2C (FFHQ)	0.11	0.12	**0.07**
LDM (FFHQ)	**0.08**	0.11	0.09

Table 3: We analyze the parallelism of manipulation directions, reporting results as average angles in radians. In diffusion space, directions for specific OOD image manipulations are nearly parallel (close to 0), while the leading GAN-inversion technique for OOD images doesn't show this consistency.

Linear traversals along a geodesic Prior works such as DiffusionCLIP (Kim, Kwon, and Ye 2022) have already shown that given an initial source latent z_S, it is possible to manipulate it to a target latent z_T using gradient descent on a loss objective $\mathcal{L}_{obj}$. We investigate how the properties of a diffusion space translate into the manipulation of various kinds that interest the community. The manipulations can be broadly categorized into three groups: low-level, semantic, and geometric manipulations. Low-level image manipulations include tasks like adjusting lighting and color, while semantic manipulations involve modifying attributes with higher semantic meanings. On the other hand, geometric manipulations require complex transformations like pose transfer. We perform each of these types of manipulations by optimizing appropriate loss functions. For example, for object pose modification, the loss objective combines an identity loss to maintain image consistency and an alignment loss to achieve the desired pose. After the target image is obtained, we perform interpolation following Algorithm 1 between z_S and z_T to get an approximation of the geodesic that corresponds to this manipulation. If the intermediate images obtained show a steady gradation in their properties, we term the direction as *linear* i.e., representable by a geodesic. We found that a broad class of linear directions can be obtained for many differing tasks. We leverage the parallel nature of manipulation directions to identify average global direction by utilizing multiple pairs of ground-truth images where a specific property, is altered. The global direction is applied to unseen images to effect the same manipulation. This strategy works effectively when ground-truth pairs for a certain manipulation can be accessed, and linear traversal is possible in the latent space. We provide visual results in Figure 3 for different types of manipulations along appropriate geodesics.

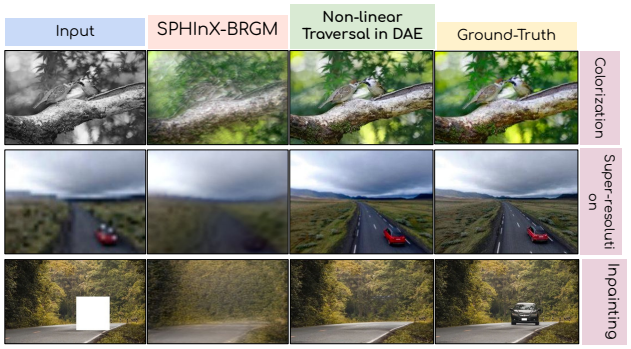

Figure 4: We show that non-linear traversal in diffusion space yields high-quality results for image colorization, super-resolution, and object removal using inpainting.

Learning non-linear traversals in the diffusion space To learn non-linear traversals in the diffusion space, we use a neural network f_θ, learning non-linear directions. We choose to solve three popular computer vision tasks: (1) Inpainting (INP), (2) Colorization (CL), and (3) Super Resolution (SR), using the diffusion space for OOD images. We extract a 32×32 section from the original image and use the modified image as input for inpainting. In the case of colorization, we use grayscale images as input. On the other hand, super-resolution involves increasing the resolution from 8×8 to 256×256, representing a $32\times$ upsampling. We first use bicubic interpolation to generate a blurred 256×256 image that serves as the input. For each task, input and ground-truth images are inverted to various diffusion spaces, and a standard MLP network, comprising five layers with two 256-size bottleneck layers and ReLU activations, is trained to learn a mapping between these in-

	CL		SR		INP	
Models	**PSNR (↑)**	**SSIM (↑)**	**PSNR (↑)**	**SSIM (↑)**	**PSNR (↑)**	**SSIM (↑)**
SPHInX-PULSE (FFHQ)	26.85	0.8726	25.11	0.8604	23.22	0.8148
SPHInX-BRGM (FFHQ)	27.34	0.8823	27.13	0.8837	26.16	0.8701
DAE (FFHQ)	**32.12**	**0.968**	**32.08**	**0.964**	32.06	0.963
D2C (FFHQ)	32.02	0.961	32.03	0.963	**32.12**	**0.970**
LDM (FFHQ)	31.87	0.957	31.72	0.952	32.04	0.964

Table 4: In the table, we assess non-linear traversal performance for colorization, super-resolution, and inpainting. The results indicate easier traversal and higher quality outputs in diffusion space compared to GANs' latent space. All the results are calculated using the LHQ-256 (Skorokhodov, Sotnikov, and Elhoseiny 2021) dataset.

verted latent pairs. We use the Adam optimizer (Kingma and Ba 2014) with a learning rate of 10^{-3}. Significant improvement is achieved through spherical regularization (Menon et al. 2020), which projects the network output back onto the sphere of our latent space. This approach, feasible due to the known manifold geometry of any diffusion space, ensures consistent latent space optimization, thereby preserving quality and accelerating convergence. We employ a standard $L1$ loss between the predicted and ground-truth latents, which suffices without additional image space losses. The models are trained and tested on the LHQ-256 (Skorokhodov, Sotnikov, and Elhoseiny 2021) dataset. As measured by the PSNR and SSIM metrics, evaluation results are summarized in Table 4.

Comparing non-linear traversal in StyleGAN's latent space For OOD images across tasks, direct baselines in pretrained GAN latent spaces are scarce due to inversion challenges. Yet, models like PULSE (Menon et al. 2020) and BRGM (Marinescu, Moyer, and Golland 2020) exist for faces. We built baselines using SPHInX (Subramanyam et al. 2022) for OOD inversion in StyleGAN's space, followed by adapting PULSE and BRGM. This sheds light on non-linear traversal in StyleGAN's space for OOD images. After modifying PULSE and BRGM for each task, we used SPHInX for inversion and changed the loss function appropriately for tasks like colorization and inpainting. Evaluation of the LHQ-256 dataset (Table 4) underscores the ease of traversing diffusion models over GAN-latent spaces. Visual comparisons are in Figure 4, showcasing the inpainting model's versatility, including object removal.

Figure 5: Using just 10 instances from ImageNet's Fruits and Cars classes, this figure demonstrates FS-DAE's ability to produce diverse, high-quality samples for both.

Few-Shot Generation

GANs are data-hungry and inefficient in training from scratch on new data, moreover their latent spaces do not allow selective sampling of OOD data from a few samples. We note that the task of generating new samples using a small fixed set of samples (few-shot generation) can be posed as a geometric problem in a latent space. Given N inverted latents, it is possible to consider any point on the $\binom{N}{2}$ lines as a valid new generation. As a preliminary study, we only consider the linear version in this work, i.e., we sample new points from the $\binom{N}{2}$ geodesics in the latent space of DAE and D2C. Our method is benchmarked against the state-of-the-art Latent Learner (Mondal et al. 2023) and another few-shot StyleGAN approach (Ojha et al. 2021). We evaluated our technique on a setup consistent with (Mondal et al. 2023; Ojha et al. 2021), using images from datasets like Babies and sunglasses. Our technique's efficacy is evident from FID scores in Table 5. *Notably, unlike the comparable works needing fine-tuning for each set of few-shot samples, our method (FS-DAE & FS-D2C) surpasses the SOTA without fine-tuning.* We also show visual results in Figure 5 for two classes, i.e., 10 images of cars and fruits, respectively, taken from ImageNet. Our approach generates high-quality images with significant variations, thus opening new research directions for few-shot image generation.

Method	**Babies**	**Sunglasses**	**Sketches**	**Bitmoji**
Latent Learner	63.31	35.64	**35.59**	64.50
Ojha et.al	74.39	42.13	45.67	69.54
FS-DAE (Ours)	**63.12**	**35.51**	36.21	**64.43**
FS-D2C (Ours)	63.24	35.59	37.90	64.76

Table 5: We compare FID scores for the different approaches for a few-shot generation. Our naive geometric approach generates near state-of-the-art results for this task.

Conclusion

We investigate the latent spaces of advanced diffusion models like D2C, DAE, and LDM, focusing on OOD image inversion and manipulation. These models outperform state-of-the-art GANs, as shown by our rigorous experiments and theoretical analysis. They are robust and particularly advantageous in low-resource settings, where training new models is impractical. Our research promotes further exploration of this paradigm.

References

Abdal, R.; Qin, Y.; and Wonka, P. 2019. Image2StyleGAN: How to Embed Images Into the StyleGAN Latent Space?

Abdal, R.; Zhu, P.; Mitra, N. J.; and Wonka, P. 2021. StyleFlow: Attribute-conditioned Exploration of StyleGAN-Generated Images using Conditional Continuous Normalizing Flows. *ACM Transactions on Graphics*, 40(3): 1–21.

Bansal, A.; Borgnia, E.; Chu, H.-M.; Li, J. S.; Kazemi, H.; Huang, F.; Goldblum, M.; Geiping, J.; and Goldstein, T. 2022. Cold Diffusion: Inverting Arbitrary Image Transforms Without Noise.

Bau, D.; Zhu, J.-Y.; Strobelt, H.; Zhou, B.; Tenenbaum, J. B.; Freeman, W. T.; and Torralba, A. 2018. GAN Dissection: Visualizing and Understanding Generative Adversarial Networks. *ArXiv*, abs/1811.10597.

Ben-David, S.; Blitzer, J.; Crammer, K.; and Pereira, F. 2006. Analysis of Representations for Domain Adaptation. In Schölkopf, B.; Platt, J.; and Hoffman, T., eds., *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

Brock, A.; Donahue, J.; and Simonyan, K. 2019. Large Scale GAN Training for High Fidelity Natural Image Synthesis. *ArXiv*, abs/1809.11096.

Choi, J.; Lee, J.; Yoon, C.; Park, J. H.; Hwang, G.; and Kang, M. 2022. Do Not Escape From the Manifold: Discovering the Local Coordinates on the Latent Space of GANs. arXiv:2106.06959.

De Bortoli, V. 2022. Convergence of denoising diffusion models under the manifold hypothesis.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Dhariwal, P.; and Nichol, A. 2021. Diffusion Models Beat GANs on Image Synthesis.

et al, A. R. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents.

Giné , E.; Koltchinskii, V.; Li, W.; and Zinn, J. 2006. Preface. In *High Dimensional Probability*, v–vi. Institute of Mathematical Statistics.

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

Ha, D.; and Eck, D. 2017. A Neural Representation of Sketch Drawings. *CoRR*, abs/1704.03477.

Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A.; Kingma, D. P.; Poole, B.; Norouzi, M.; Fleet, D. J.; and Salimans, T. 2022. Imagen Video: High Definition Video Generation with Diffusion Models.

Härkönen, E.; Hertzmann, A.; Lehtinen, J.; and Paris, S. 2020. GANSpace: Discovering Interpretable GAN Controls. In *Proc. NeurIPS*.

Karras, T.; Aittala, M.; Laine, S.; Härkönen, E.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2021. Alias-Free Generative Adversarial Networks. In *NeurIPS*.

Karras, T.; Laine, S.; and Aila, T. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4396–4405.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2019. Analyzing and Improving the Image Quality of StyleGAN. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8107–8116.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and Improving the Image Quality of StyleGAN. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8107–8116.

Kim, G.; Kwon, T.; and Ye, J. C. 2022. DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2426–2435.

Kingma, D. P.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization. *CoRR*, abs/1412.6980.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Marinescu, R. V.; Moyer, D.; and Golland, P. 2020. Bayesian Image Reconstruction using Deep Generative Models. *ArXiv*, abs/2012.04567.

Matthey, L.; Higgins, I.; Hassabis, D.; and Lerchner, A. 2017. dSprites: Disentanglement testing Sprites dataset. https://github.com/deepmind/dsprites-dataset/.

Menon, S.; Damian, A.; Hu, S.; Ravi, N.; and Rudin, C. 2020. PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models.

Mondal, A. K.; Tiwary, P.; Singla, P.; and AP, P. 2023. Few-shot Cross-domain Image Generation via Inference-time Latent-code Learning. In *The Eleventh International Conference on Learning Representations*.

Ojha, U.; Li, Y.; Lu, C.; Efros, A. A.; Lee, Y. J.; Shechtman, E.; and Zhang, R. 2021. Few-shot Image Generation via Cross-domain Correspondence. In *CVPR*.

Pope, P.; Zhu, C.; Abdelkader, A.; Goldblum, M.; and Goldstein, T. 2021. The Intrinsic Dimension of Images and Its Impact on Learning.

Preechakul, K.; Chatthee, N.; Wizadwongsa, S.; and Suwajanakorn, S. 2022. Diffusion Autoencoders: Toward a Meaningful and Decodable Representation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Radford, A.; Metz, L.; and Chintala, S. 2015. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *CoRR*, abs/1511.06434.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021a. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021b. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752.

Saharia, C.; Chan, W.; Chang, H.; Lee, C. A.; Ho, J.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2021. Palette: Image-to-Image Diffusion Models. *ACM SIGGRAPH 2022 Conference Proceedings*.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.; Ghasemipour, S. K. S.; Ayan, B. K.; Mahdavi, S. S.; Lopes, R. G.; Salimans, T.; Ho, J.; Fleet, D. J.; and Norouzi, M. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding.

Sauer, A.; Schwarz, K.; and Geiger, A. 2022. StyleGAN-XL: Scaling StyleGAN to Large Diverse Datasets.

Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; and Komatsuzaki, A. 2021. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. *ArXiv*, abs/2111.02114.

Scott, D. W. 2015. Multivariate density estimation: theory, practice, and visualization. *John Wiley & Sons*.

Shen, Y.; Gu, J.; Tang, X.; and Zhou, B. 2019. Interpreting the Latent Space of GANs for Semantic Face Editing. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9240–9249.

Shen, Y.; and Zhou, B. 2020. Closed-Form Factorization of Latent Semantics in GANs. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1532–1540.

Sinha*, A.; Song*, J.; Meng, C.; and Ermon, S. 2021. D2C: Diffusion-Denoising Models for Few-shot Conditional Generation. In *Neural Information Processing Systems*.

Skorokhodov, I.; Sotnikov, G.; and Elhoseiny, M. 2021. Aligning Latent and Image Spaces to Connect the Unconnectable. *arXiv preprint arXiv:2104.06954*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021a. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021b. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.

Subramanyam, R.; Narayanaswamy, V.; Naufel, M.; Spanias, A.; and Thiagarajan, J. J. 2022. Improved StyleGAN-v2 based Inversion for Out-of-Distribution Images. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvari, C.; Niu, G.; and Sabato, S., eds., *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 20625–20639. PMLR.

Tov, O.; Alaluf, Y.; Nitzan, Y.; Patashnik, O.; and Cohen-Or, D. 2021. Designing an Encoder for StyleGAN Image Manipulation.

van den Oord, A.; Vinyals, O.; and Kavukcuoglu, K. 2017. Neural Discrete Representation Learning. In *NIPS*.

Vershynin., R. 2015. Estimation in high dimensions: a geometric perspective.". *Sampling theory, a renaissance*.

Voynov, A.; and Babenko, A. 2020. Unsupervised discovery of interpretable directions in the gan latent space. In *International Conference on Machine Learning*, 9786–9796. PMLR.

Wang, H.; Ge, S.; Lipton, Z.; and Xing, E. P. 2019. Learning Robust Global Representations by Penalizing Local Predictive Power. In *Advances in Neural Information Processing Systems*, 10506–10518.

Wang, Z.; Bovik, A.; Sheikh, H.; and Simoncelli, E. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612.

Xiao, Z.; Kreis, K.; and Vahdat, A. 2021. Tackling the Generative Learning Trilemma with Denoising Diffusion GANs. *CoRR*, abs/2112.07804.

Zhao, D.; Zhu, J.; and Zhang, B. 2019. Latent Variables on Spheres for Autoencoders in High Dimensions.