Attribute Diffusion: Diffusion Driven Diverse Attribute Editing

Rishubh Parihar*,[1] Prasanna Balaji*,[1] Raghav Magazine[2] Sarthak Vora[3]
Varun Jampani[4] R. Venkatesh Babu[1]

[1]Indian Institute of Science, Bangalore [2]IIT Dharward [3]UCLA [4]Stability AI

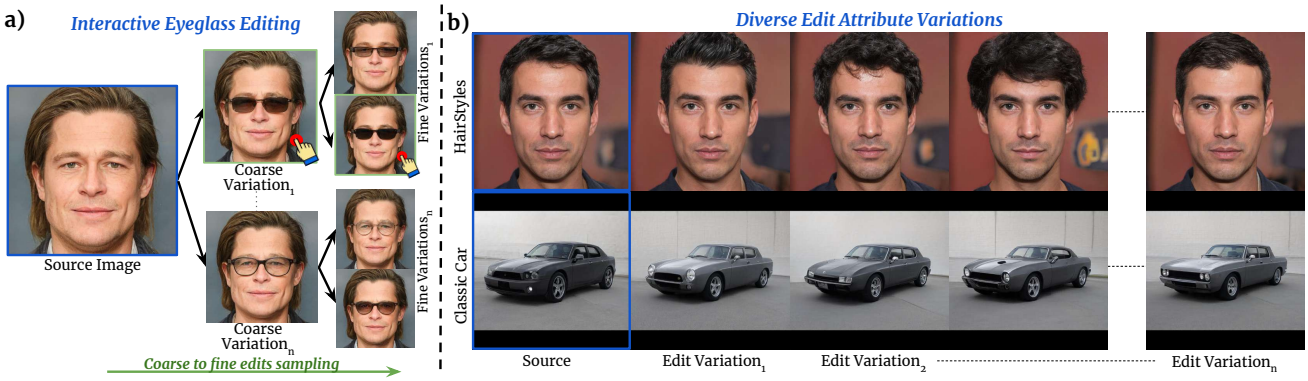

Figure 1. **a) Interactive *coarse-to-fine* attribute exploration**. We propose a principal approach to generate and explore diverse attribute edits. A user can first select from a large set of *coarse* attribute edit variations and then explore more *fine* attribute variations hierarchically. **b) Diverse attribute editing.** The proposed method generates diverse attribute edits by learning a space of plausible attribute variations.

Abstract

Image attribute editing is a widely researched area fueled by the recent advancements in deep generative models. Existing methods treat semantic attributes as binary and do not allow the user to generate multiple variations of the attribute edits. This limits the applications of editing methods in the real world, e.g., exploring multiple eyeglass variations on an e-commerce platform. In this work, we present a technique to generate a collection of diverse attribute edits and a principled way to explore them. Generation and controlled exploration of attribute variations is challenging as it requires fine control over the attribute styles while preserving other attributes and the identity of the subject. Capitalizing on the attribute disentanglement property of the latent spaces of pretrained GANs, we represent the attribute edits in this space. Next, we train a diffusion model to model these latent directions of edits. We propose a coarse-to-fine sampling strategy to explore these variations in a controlled manner. Extensive experiments on various datasets establish the effectiveness and generalization of the proposed approach for the generation and controlled exploration of diverse attribute edits. Code is available at - project page.

1. Introduction

Recent advancements in deep generative models [8, 24, 45, 47] have unlocked multiple image editing and synthesis applications. Of particular interest is fine-grained attribute editing, where a given attribute (e.g., eyeglasses or hairstyles) needs to be edited without altering other attributes or the subject's identity. Existing editing methods either consider attributes as binary [3, 40] and generate a single edit or perform text-based edit to generate a few plausible edits [16, 39, 40]. However, attributes have many appearance, shape, and style variations in the real world. For example, multiple variations across eyeglasses, smiles, and hairstyles exist.

The ability to generate and select from multiple attribute edit variations significantly enriches the user experience. For e.g., a user wants to try out various eyeglasses in a virtual try-on interface before selecting a preferred one. Further, a principled way of exploration is desired where a user can first select among coarse variations such as cooling glasses or reading glasses, then explore more fine variations

of the selected coarse style (ref. Fig. 1a)). We termed such a hierarchical exploration of edits a *coarse-to-fine* exploration and believe this is a natural way of exploring multiple options. In this work, we raise the following questions - ***i)*** *How do we generate multiple variations of a given attribute edit?* ***ii)*** *How do we explore the generated variations in a coarse-to-fine manner?*

One plausible solution is to use text-based editing using a pretrained generative model [16, 38, 40]. However, text represents concepts at a semantic level and limits the description of finer aspects of attributes (such as eyeglass shapes). Instead, we propose to train a generative model to learn the distribution over all attribute variations to sample new edits. Specifically, we model the attribute distribution with a Diffusion Model (DM). Further, to enable a *finer* exploration, we design a hierarchical sampling of DM that enables *coarse-to-fine* sampling of attribute edits. Gathering a real dataset with multiple attribute edits per input image is extremely challenging. Motivated by the success of recent methods leveraging synthetic paired datasets for training models [6, 25, 28, 59], we gathered a synthetic dataset of paired images before and after edit. This dataset can be cheaply obtained using existing editing methods that are designed to perform single edits.

To model this distribution effectively, we need a disentangled and semantically meaningful feature space of attributes. We note that the latent spaces of style-based GANs [23, 24] are semantically rich and provide fine-grained control for attribute editing [3, 15, 40, 48] which makes them suitable for modeling and exploring fine attribute variations. We first establish the existence of subregions in the latent space that control variations of a single attribute and model them with a DM for guided exploration of diverse variations.

We apply DM in the latent space of pretrained style-based GANs to model the distribution over diverse attribute variations. DMs applied in the latent space: **i)** enable a controlled exploration of attribute variations by controlling the denoising trajectory for *coarse-to-fine* sampling; **ii)** covers diverse attribute variations due to excellent mode coverage. Notably, as we apply DM in compressed latent space, it enables *efficient training and inference*. To the best of our knowledge, we are the first to use the DM to model the latent space of pretrained GANs.

A major challenge is the evaluation of the quality of the diverse attribute edits. Specifically, we want to evaluate the *localized diversity* in attribute variations along with *disentanglement* from other attributes. Existing metrics such as FID [17] compute the overall quality of edits globally. Secondly, popular disentanglement metrics such as attribute scores rely on attribute classifiers, which can be biased [33]. To this end, we propose a novel metric *Attribute Diversity Score (ADS)*, that measures both localized diversity in the attribute variations and disentanglement with other attributes. ADS uses the semantic mask of the attribute and quantifies the variations in the desired attribute region vs other attribute regions.

We extensively evaluated our method for diverse attribute editing on multiple datasets. The proposed method can achieve highly diverse attribute edits while preserving the subject's identity. Additionally, the proposed *coarse-to-fine* sampling enables guided exploration of diverse attribute variations. Further, we present detailed ablations and results for the edits on out-of-distribution painting images from Metfaces [22]. Our method generalizes to 3D aware GAN model [8] and performs diverse face attribute editing with 3D consistency. Our main contributions are as follows:

1. A method to generate diverse attribute editing and enabling *coarse-to-fine* exploration of attributes.
2. Diffusion model in the latent space of pretrained style-based GANs and a novel hierarchical sampling method during the reverse diffusion process.
3. Extensive experiments and results for diverse attribute editing on multiple datasets, generalization results on 3D-aware GANs, and out-of-domain images.
4. A novel metric ADS to evaluate both diversity and disentanglement in the generated attribute variations

2. Related work

Image editing with latent manipulation Various image and video editing works have been proposed that leverage semantics in the latent space of GANs [3, 15, 36, 40, 51, 53] and diffusion models [12, 14, 27, 35, 37, 43] to edit images. One direction of works obtains a global edit direction in the latent space for each attribute [12, 15, 43, 48, 50]. Traversal along these global directions edits the corresponding attribute in the generated image. Another cohort of methods obtains a local direction for each latent code. Essentially, a non-linear mapping is learned between the input latent code and the desired edit code, using transformer networks [19, 54], a mapper network [40]. To obtain the edit direction for a given attribute, these works use - attribute classifiers [3, 29], segmentation masks [30], clip-supervision [2, 40, 60] or perform unsupervised decomposition of the latent space [15, 49, 57]. StyleFlow [3] learns a conditional normalizing flow network to learn a deterministic mapping from a source latent to a single edited latent code for each attribute. To enable editing on real images, encoder-decoder frameworks have been proposed that map the real image into the $\mathcal{W}+$ space and use StyleGAN's generator as the frozen decoder after latent editing [1, 5, 44, 52].

Non-binary attribute editing. Some editing works try to model the continuous attribute variations instead of treating attributes as binary. Works like [40, 56] obtain directions for non-binary attributes such as image style or expression

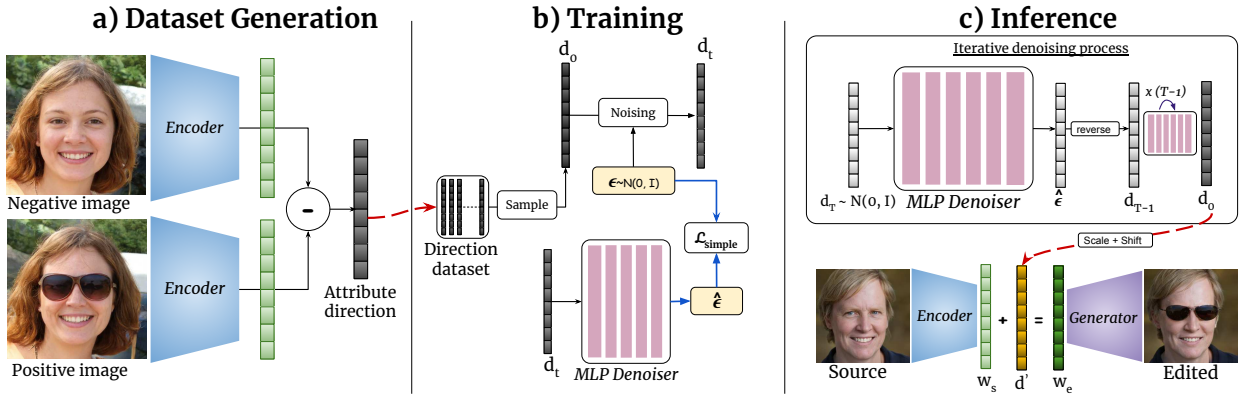

Figure 2. The methodology we present for diverse attribute editing comprises three major stages: a) **Dataset Generation.** We create a dataset of edit directions by embedding negative and positive image pairs into the latent space and computing the difference between these directions. b) **Training.** We train a DDPM model over the dataset of edit directions for the given attribute employing a denoising objective. c) **Inference.** For editing, we add a sampled latent edit direction to the encoded latent of a source image.

change, but they also use a single edit direction for the given attribute/style. Unsupervised methods [15, 57] can learn a finite set of disentangled directions controlling each attribute. StyleSpace [53] learns a set of disentangled vectors in style space controlling each attribute. However, these methods learn only a finite set of edit directions and cannot cover all the possible variations for any given attribute. FLAME [36] proposed a task of attribute style manipulation, where they generate edit variations by navigating in the attribute manifold. Furthermore, StyleFusion [20] showed that a pre-trained StyleGAN could be used to decompose spatial semantic regions. In contrast to these approaches, we learn a distribution over edit directions for a given attribute and sample multiple variations of an edit.

Diffusion Models (DMs) are likelihood-based models that have achieved state-of-the-art performance in sample generation [45] and density estimation [10]. In contrast to GANs, DMs, being likelihood-based models, prevent mode collapse and learn rich multi-modal distributions. DMs modeled as hierarchical denoising autoencoders [18] are trained to iteratively denoise images starting from pure noise. Due to the sequential nature of DMs, applying them in the pixel space for high-resolution images leads to high training costs and slow inference speeds [10]. To this end, latent diffusion models (LDMs) [45] have been proposed to first encode the images into much lower dimensional spatial latent codes and apply DM in the latent space. A range of works have been proposed that leverage the rich text-to-image diffusion models to perform semantic text-based editing [6, 16, 39], however they are unable to generate fine-grained attribute variations. Subsequently, multiple works are proposed that perform latent space diffusion for motion synthesis [9], language generation [32], point clouds generation [58], generating brain imaging [42]. We apply the diffusion model on the highly compressed $\mathcal{W}+$ to model attribute variations.

3. Method

We formulate the task of diverse attribute editing as a distribution learning problem over the attribute variations. Specifically, we train a generative model $\mathcal{G}_\mathcal{A}$ for attribute $\mathcal{A}$, representing all of its variations and enabling a controlled way of exploring them. Instead of learning $\mathcal{G}_\mathcal{A}$ in the image space where multiple attributes can be entangled, we propose to learn $\mathcal{G}_\mathcal{A}$ over a lower dimensional disentangled latent representation $\mathcal{F}(x)$. Such a formulation does not only help in exploring attribute variations in a disentangled manner but is also extremely computationally efficient due to lower dimensional representation. We use $\mathcal{W}/\mathcal{W}+$ latent space of pretrained style-based GAN models [8, 24] as $\mathcal{F}$, due to their exceptional attribute disentanglement properties. Specifically, semantic edit directions exist in the latent space responsible for disentangled editing of a single attribute [15, 48]. We train a DM to implement $\mathcal{G}_\mathcal{A}$ over a dataset of attribute editing directions $\mathcal{D}_\mathcal{A}$ in the $\mathcal{W}$ latent space. We choose DM to model $\mathcal{G}_\mathcal{A}$, as it enables *fine-grained control in the sampling process* due to the hierarchical nature of denoising [34]. Additionally, it has excellent *mode-covering abilities*, which is crucial to model all attribute variations.

Our overall method is shown in Fig. 2. In the following, we first explain dataset creation (Sec. 3.1), followed by model training (Sec. 3.2) and inference (Sec. 3.3). Finally, we present *coarse-to-fine* sampling of DM for a guided exploration of various attribute variations (Sec. 3.4).

3.1. Data Generation

Gathering a real-world dataset of disentangled attribute edits is extremely challenging. Several recent works [6, 25, 28, 41, 59] have shown the efficacy of generating paired datasets from existing models for training specialized models. Motivated by this, to learn the distribution of diverse edits, we gather a synthetic dataset of image pairs with and without

the attribute edits. This synthetic dataset can be easily obtained using popular existing editing methods [4,16,38,40]. Note that these approaches can not be directly used for diverse editing as they *only provide a single edit output per input image*. The gathered dataset consists of image pairs of a positive image $\mathcal{I}^p_{\mathcal{A}}$ (which has the attribute $\mathcal{A}$) and a negative image $\mathcal{I}^n_{\mathcal{A}}$ (which does not have $\mathcal{A}$). Next, we embed these pairs in the latent space using a GAN encoder model $\mathcal{E}$ [44] and obtain an edit direction $d_{\mathcal{A}}$, where $d_{\mathcal{A}} = \mathcal{E}(\mathcal{I}^p_{\mathcal{A}}) - \mathcal{E}(\mathcal{I}^n_{\mathcal{A}})$. This yields a dataset $\mathcal{D}_{\mathcal{A}}$ of diverse edit directions for attribute $\mathcal{A}$. We show the distribution of cosine similarity of the edit directions with the mean attribute edit direction in Fig. 3-Right. Observe that the obtained directions have *high diversity* and do not align with the mean direction. We have provided dataset samples and details about methods used for creation in the supplementary. We note that the inaccuracies in the editing methods could translate into the training dataset. However, our goal is to propose a generalized framework for diverse attribute editing, which will automatically benefit from the advancements in image editing methods designed for generating single edit output.

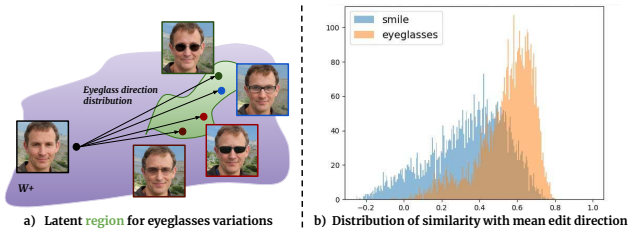

Figure 3. **(Left)** Latent region capturing diverse attribute styles in the $\mathcal{W}/\mathcal{W}+$ latent space of style-based GANs. **(Right)** Histogram of the Cosine Similarity of all edit directions with the mean direction. The spread of values suggests that the editing directions, although for the same attribute, showcase a large variety.

3.2. Training

We train a DDPM model $\mathcal{G}_{\mathcal{A}}$ over the dataset of edit directions $\mathcal{D}_{\mathcal{A}}$ to model the variations of attribute $\mathcal{A}$. Diffusion models enable modeling the rich multimodal distribution of attribute variations present in the $\mathcal{W}/\mathcal{W}+$ latent spaces. Additionally, it enables a hierarchical control over sampling attribute variations in a *coarse-to-fine* manner (Sec. 3.4). During training, we randomly sample a edit direction $\boldsymbol{d_0} \in \mathcal{D}_A$, and corrupt it with a Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$ (shown in Fig. 2).

$$\mathbf{d_t} = \sqrt{\bar{\alpha}_t}\mathbf{d_0} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon} \tag{1}$$

with $\bar{\alpha} = \prod_{i=1}^{T} \alpha_i$, and $0 = \alpha_T < \alpha_{T-1} < ..... < \alpha_0 = 1$, being hyperparameter of diffusion schedule. We implement the denoiser network $\epsilon_\theta(\boldsymbol{d_t}, t)$ as a time-conditioned Multi-Layer Perceptron (MLP) network. To train the denoising network, we use the simple loss [18], between added noise ϵ and $\epsilon_\theta(\boldsymbol{d_t}, t)$:

$$\mathcal{L}_{simple} = \mathbb{E}_{\mathbf{d_0},t,\boldsymbol{\epsilon}}[\|\boldsymbol{\epsilon} - \epsilon_\theta(\mathbf{d_t}, t)\|_2^2] \tag{2}$$

As a normalization step, we subtract the mean direction $\mathbf{m}_{\mathcal{A}}$ from the edit directions and normalize them to unit length before training $\mathcal{G}_{\mathcal{A}}$. The mean $\mathbf{m}_{\mathcal{A}}$ is computed by averaging over all directions $d_i \in \mathcal{D}_{\mathcal{A}}$. Intuitively, it disentangles the attribute's presence (controlled by the mean) from its variations and enables explicit modeling of only attribute variations. We empirically observe that having this preprocessing improves the models' performance.

3.3. Diverse attribute editing

Given a source image $\mathcal{I}_s$ to be edited, we first embed $\mathcal{I}_s$ to its corresponding latent code w_s, where $w_s = \mathcal{E}(\mathcal{I}_s)$. Next we sample a new edit direction $\mathbf{d_0}$ from $\mathcal{G}_{\mathcal{A}}$ by iterative denoising of a noisy sample $\mathbf{d_T} \sim \mathcal{N}(0, I)$. Similar to the truncation trick for sampling latent code ($w' = \bar{w} + \gamma(w - \bar{w})$) in $\mathcal{W}$ space, we obtain edit direction $\mathbf{d}'$ with mean subtracted sampled direction $\mathbf{d_0}$. Finally, we multiply with a scale factor λ before adding it to the source latent.

$$\boldsymbol{w_e} = \boldsymbol{w_s} + \lambda \mathbf{d}' \quad where \quad \mathbf{d}' = \mathbf{m}_{\mathcal{A}} + \gamma \mathbf{d_0} \tag{3}$$

The edited latent code $\boldsymbol{w_e}$ is then passed through the pretrained style-based generator model $\mathcal{G}_{\mathcal{I}}$ to obtain the edited image $\mathcal{I}_e$, where $\mathcal{I}_e = \mathcal{G}_{\mathcal{I}}(\boldsymbol{w_e})$ (Fig. 2). We define γ as the diversity parameter and λ as the scale parameter as they control the diversity and strength of the edits, respectively.

Intuition. The hyperparameter γ controls the diversity in the edits; higher γ will generate edit directions that deviate from the mean edit direction and result in diverse edits. λ controls the strength of the edit, smaller λ values result in very subtle changes in the output, and large λ values generate substantial edits.

3.4. Coarse-to-fine sampling

Directly using Eq. 3 to sample can already provide us with multiple plausible attribute edits; however, it lacks controlled exploration over attribute variations. To this end, we propose a modified sampling from $\mathcal{G}_{\mathcal{A}}$ by 'hijacking' the reverse diffusion process based on a coarse edit. Specifically, we aim to explore the variations hierarchically by first generating a set of coarse attribute variations from which to select. Next, we explore fine-grained variations of the selected coarse variation as shown in Fig. 1. Such a *coarse-to-fine* exploration process is highly intuitive to the user and is common in selecting accessories in the real world. We leverage the hierarchical nature of the image generation process during reverse diffusion, similar to [34], to enable *coarse-to-fine* exploration of attribute styles. The proposed sampling process is shown in Fig. 4, where we start with a sample $\mathbf{d_T} \sim \mathcal{N}(0, I)$. Next, we iteratively de-noise it using the reverse process to obtain an edit direction $\mathbf{d_0}$. To obtain fine-grained variations for the generated edit direction $\mathbf{d_0}$, we denoise the intermediate sample at $\mathbf{t_0}$, $d_{\mathbf{t_0}}$ multiple times to obtain *fine-grained* variations of $\mathbf{d_0}$. Detailed procedure is presented in Algorithm 1. The time split hyperparameter $\mathbf{t_0}$, controls the extent of fineness in the attribute variations (ref. Fig. 9).

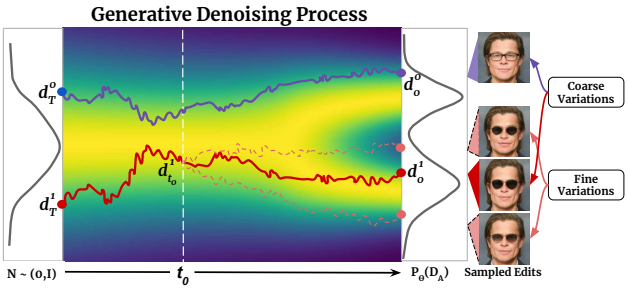

Figure 4. Modified reverse process for hierarchical *coarse-to-fine* sampling of attribute variations. Starting from two noises $\mathbf{d_T^0}$ and $\mathbf{d_T^1}$, we generate two coarse edit directions $\mathbf{d_0^0}$ and $\mathbf{d_0^1}$ by iterative denoising. To obtain the fine variations of $\mathbf{d_0^1}$, we denoise it again from $\mathbf{d_{t_0}^1}$ to obtain another trajectory. The denoised trajectory will have similar coarse structure details and variations in only the fine details as it started from intermediate time step $\mathbf{t_0}$. The split time $\mathbf{t_0}$ controls the *fine-ness* of the exploration.

Algorithm 1: Coarse-to-fine sampling

Data: Diffusion model ϵ_θ, split timestep t_0, Number of fine variations n

Result: Coarse variation d_0 and corresponding fine variations $\{d_0^1, d_0^2, \ldots, d_0^n\}$

$d_T \sim \mathcal{N}(0, I)$;

for t *in* $T \rightarrow 1$ **do**

  $d_{t-1} = d_t - \sqrt{1-\alpha_t} \cdot \epsilon_\theta(d_t, t)$;

end

for i *in* $1 \rightarrow n$ **do**

  $d_{t_0}^i = d_{t_0}$;

  for t *in* $t_0 \rightarrow 1$ **do**

    $d_{t-1}^i = d_t^i - \sqrt{1-\alpha_t} \cdot \epsilon_\theta(d_t^i, t)$;

  end

end

Return: $d_0, \{d_0^1, d_0^2, \ldots, d_0^n\}$;

4. Experiments

We perform extensive experiments and ablations to evaluate our model for diverse attribute editing. First, we discuss the dataset, implementation details, and novel Attribute Diversity Score metric. Next, we present results on diverse attribute editing for faces, out-of-domain face images from Metfaces [22], cars, and church datasets and finally showcase 3D-aware attribute editing on EG3D [8]. Please check the accompanying website for more visual results.

Dataset. We synthetically create image pairs with a single attribute change as explained in Sec. 3.1. Specifically, we used StyleCLIP [40] method to edit $30K$ images from CelebA-HQ [21] dataset for hairstyle attribute. We used the original and edited images as the attribute pairs to obtain edit directions. For eyeglasses and smile attributes, we synthetically created the image pairs by cut-pasting the region of interest similar to [7]. For age, we generated image pairs using the age editing method [4]. For the car dataset, we used StyleCLIP [40] on Stanford cars [26] dataset to obtain the image pairs dataset. Note that our method can use any existing editing method to generate image pairs.

Implementation Details. To learn a distribution over edit directions for each attribute, we train a separate DDPM model over the dataset of edit directions $\mathcal{D}_a$ for each attribute. The denoising MLP network has 10 fully connected layers along with time-conditioning and skip connections. Details about the architecture are provided in the SM. We trained the model for 200 epochs with a batch size of 256. As there is a semantic hierarchy across layers in the StyleGAN2 generator, we select a subset of layers for editing each attribute. In our face experiments, the following layers work best for each attribute (the layer numbers are 0-indexed): layers $5-7$ for eyeglasses, $4-6$ for hairstyles, $5-6$ for the smile and $4-6$ for age attribute. We have used the same set of layers for editing on EG3D latent codes as well. For car experiments, we have used layers $4-7$ and layers $4-8$ for church editing.

4.1. How to evaluate attribute edit variations?

Two major aspects to effectively measure diverse attribute editing are - *diversity* and *disentanglement* in edits.

Attribute diversity reflects the variations in the target attribute edits for a given input image. Existing metrics for diversity - Fréchet Inception Distance (FID) [17] and Inception Score (IS) [46], measure the global diversity and are not suitable for attribute edit variations which are *localized* in a small image region. E.g., for eyeglass editing, we expect variations in eyeglass shapes and color, which are localized in small regions near the eyes.

Attribute disentanglement measures the undesired changes in other attributes while editing the target attribute. Existing metrics use pretrained attribute classifiers to quantify attribute disentanglement [48]. However, attribute classifiers can mimic the dataset bias [33], which is common in face datasets [31] (eyeglasses correlated with age). We propose quantifying the attribute disentanglement without using attribute classifiers by measuring the changes in edit variations in spatial regions of other attributes.

Attribute Diversity Score (ADS) combines attribute diversity and disentanglement in a single metric. For diverse editing, we wish to maximize the changes in the regions associated with target attribute $\mathcal{A}$ while minimizing the diversity in the regions for other attributes. Given a source image $\mathcal{I}^s$, we edit it M times for attribute $\mathcal{A}$ to obtain a set of edited images $\{\mathcal{I}_1^e, ...\mathcal{I}_M^e\}$. Next, we subtract the source image from all the edited images to obtain a set of difference maps $H_i = \mathcal{I}_i^e - \mathcal{I}^s$, which captures pixel-wise change. Further, we compute the per-pixel standard deviation of H_i's to obtain attribute diversity map $P_\mathcal{A}$ to capture the pixel-wise

Table 1. Comparison with attributed editing methods

Attribute	Method	$CS\uparrow$	$FID\downarrow$	$AD_{\mathcal{A}}\uparrow$	$AD_{\mathcal{A}^c}\downarrow$	$ADS\uparrow$
Smile	FLAME	0.946	**46.60**	0.380	0.377	1.007
	N-Flow	0.968	49.04	0.356	0.356	1.000
	LatentCLR	0.955	60.21	0.391	0.390	1.003
	Ours	**0.969**	49.51	0.401	0.324	**1.240**
Eyeglass	FLAME	0.942	90.35	0.404	0.361	1.120
	LatentCLR	**0.975**	75.42	0.351	0.335	1.047
	N-Flow	0.952	**65.24**	0.376	0.366	1.026
	Ours	0.958	66.51	0.401	0.305	**1.317**
Hairstyle	FLAME	0.972	46.27	0.419	0.386	1.086
	N-Flow	0.967	49.55	0.383	0.366	1.046
	LatentCLR	0.969	68.83	0.384	0.381	1.007
	Ours	**0.975**	**45.11**	0.416	0.361	**1.152**

variations in the edits. We aggregate the attribute diversity map for N source images to obtain a mean diversity map $\bar{P}_{\mathcal{A}}$ for attribute $\mathcal{A}$ which signifies the diversity in regions associated with $\mathcal{A}$ as shown in Fig. 5.

$$AD_{\mathcal{A}} = \sum_{(x,y)} M_{\mathcal{A}}(x,y) \cdot \bar{P}_{\mathcal{A}}(x,y) \tag{4}$$

$$AD_{\mathcal{A}^c} = \sum_{(x,y)} (1 - M_{\mathcal{A}}(x,y)) \cdot \bar{P}_{\mathcal{A}}(x,y) \tag{5}$$

where $AD_{\mathcal{A}}$ is the attribute diversity for attribute $\mathcal{A}$ and $M_{\mathcal{A}}$ is the semantic mask for the regions associated with attribute $\mathcal{A}$. We normalize the $AD_{\mathcal{A}}$ and $AD_{\mathcal{A}^c}$, with the number of pixels for region corresponding to $\mathcal{A}$ and $\mathcal{A}^c$. The ADS for attribute A is defined as $ADS = \frac{AD_{\mathcal{A}}}{AD_{\mathcal{A}^c}}$. The obtained ADS quantifies both the attribute variations and disentanglement in a single metric and is reported in Tab. 1.

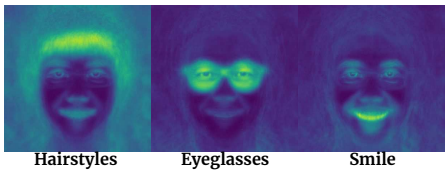

Figure 5. Aggregated attribute diversity map for hairstyle, eyeglasses, and smile editing. Observe that the diversity maps signify variation across multiple edits for each attribute.

4.2. Diverse attribute editing

We present results for hairstyle, smile, eyeglass, and age attribute variations generated by our method in Fig. 6. Additional results of our method are present in SM. Our method generates different hairstyles - bangs, mohawks, curls, and short hairs while retaining other features. Similarly, our method can generate diverse smile and age variations in a disentangled manner with identity preservation. Our proposed method can generate diverse eyeglasses with variations in frame shapes, sizes, and frame colors. Observe that all the edit variations preserve the subject's identity and other attributes. We present diverse attribute edits on real images in Fig. 6-Bottom. Additionally, we present diverse attribute edits on cars and churches in SM and Fig. 8.

Quantitative comparison. We generate five edits for each attribute for a synthetic test set of 1000 images to evaluate the quality and diversity of the edits. We compute FID [17],

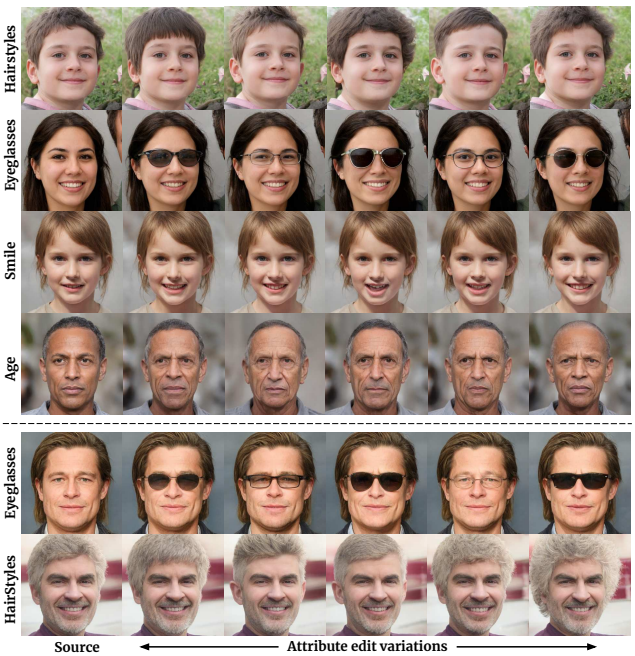

Figure 6. Diverse attribute editing on synthetic (Top) and real (Bottom) face images for several face attributes.

cosine similarity between face embeddings [11] (CS), attribute diversity ($AD_{\mathcal{A}}$), attribute entanglement ($AD_{\mathcal{A}^c}$), and Attribute Diversity Score with $M = 200$ and $N = 100$ as explained in Sec. 4.1. We compare our method against - 1) **N-flow** [13, 55] - a normalizing flow model trained on our dataset of edit directions to learn the distribution of edits, 2) **FLAME** [36], which is a few-shot method and performs diverse edits by generating random linear combinations of diverse attribute edit directions. However, This simplistic approach provides attribute variations susceptible to attribute entanglement and identity distortion indicated by higher $AD_{\mathcal{A}^c}$ score. 3) **LatentCLR** [57], an unsupervised method that learns a set of disentangled directions for editing. We trained a non-linear version of LatentCLR and manually selected the edit directions for each attribute. We used the original codebase for LatentCLR and implemented FLAME ourselves due to the unavailability of the official code (details in SM). Results are shown in Tab. 1. We note that the proposed method performed best in identity preservation and visual quality measured by CS and FID. Notably, it achieves the highest ADS by a large margin with the lowest entanglement ($AD_{\mathcal{A}^c}$) in most cases, suggesting highly disentangled and diverse attribute edits. It's important to highlight that none of these baselines can allow for coarse-to-fine attribute edit exploration, whereas our method leverages the sequential denoising of DMs to obtain hierarchical control in sampling (Fig. 8).

Qualitative comparison. We present qualitative results of comparison of diverse hairstyle generations in Fig. 7. We can observe that FLAME can generate variations, but the

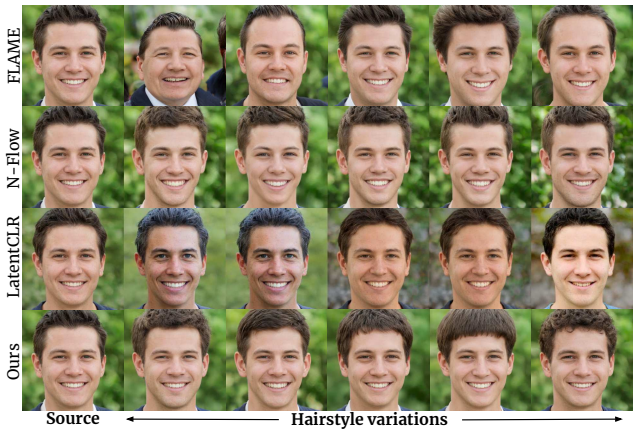

Figure 7. Comparison with existing methods for hairstyle variations. Our method generates diverse attribute variations with superior attribute disentanglement and identity preservation.

identity and other facial attributes of the subject are altered. N-Flow cannot generate diverse enough variations and result in identity change (column 3). LatentCLR significantly entangled the age, skin color, lighting, and subject's identity during editing. The proposed method can generate diverse hairstyle variations like curls, texture, and bangs while preserving the subject's identity and other attributes. We also compare with a text-to-image diffusion-based editing for shape variation [39] in SM.

4.3. Hierarchical Sampling of Attributes

Due to edits being modeled via a diffusion model, our proposed method enables us to explore attribute variations in a *coarse-to-fine* manner (ref. Sec. 3.4). We present results for hierarchical sampling in Fig. 8 for eyeglasses, church styles, and classic car styles. We start with a source image and obtain two different coarse attribute edit directions that define the overall structure of the edit. Next, we sample 3 new edit directions for each coarse style, which follow the same coarse structure (e.g., shapes such as curls, bangs, and short hair) but have subtle fine variations (e.g., texture in case of hairstyles). Fine hairstyle variations are obtained for each of the two coarse variations. Similarly, we obtain two coarse structures for churches and cars and then generate finer variations, such as the headlamp shape. Such a hierarchical sampling facilitates methodical exploration of diverse attribute editing, which is an intuitive way to first choose from coarse styles and then finetune them as per the user's choice. More results are in SM.

Analysis for split timestep $\mathbf{t_0}$. We ablate over $\mathbf{t_0}$ - denoising timestep at which we start generating fine variations (ref. Fig. 4), to quantify its impact on hierarchical sampling. We visualize the generated diverse fine variations in Fig. 9-a). A large value of $\mathbf{t_0}$ results in more broad variations such as eyeglass shapes and shades; on the contrary, splitting at later timesteps ($\mathbf{t_0}$=100) generates subtle variations in eyeglass shapes. The parameter $\mathbf{t_0}$ provides granular control over the fineness in the variations, and a user can select based on preference. This is quantitatively supported in Fig. 9-b), where we plot ADS with edits generated against split timesteps.

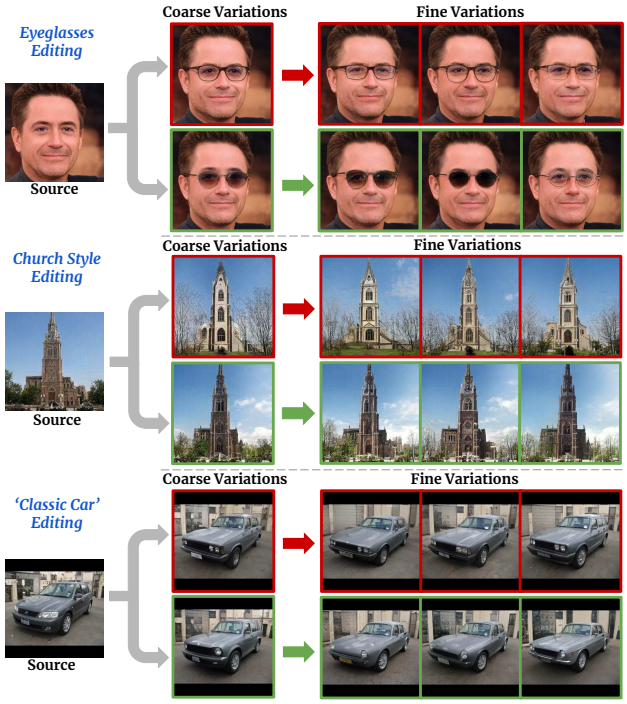

Figure 8. Hierarchical sampling of diverse attribute variations in a *coarse-to-fine* manner. First, we sample two coarse variations of edits and then generate fine variations corresponding to the selected coarse variation, providing a fine-grained control.

4.4. Ablation Study

Diversity parameter γ. We quantitatively analyze the effect of γ in Fig. 10b, where we use FID to measure diversity and Cosine Similarity (CS) to measure the similarity between the identity of the source image and the edited image. As γ increases, the FID score decreases, indicating the generation of more diverse edits at the cost of an inferior CS score, suggesting identity distortion. Through our experiments, we conclude that γ values for eyeglasses, smile, and hairstyle are $12, 12, 14$, respectively.

Strength parameter λ. We analyze the effect of λ for eyeglass edit. We measure the presence of Eyeglass in the edited image with Eyeglass Score (ES), obtained using eyeglass classifier from [23] and identity preservation with CS score in Fig. 10a. To validate the effect of strength parameter λ agnostic to the diversity, we kept the product $\lambda * \gamma$ constant while ablating over λ. We can observe an increase in λ results in more prominent eyeglasses until a threshold beyond which the person's identity is modified. We have identified that $\lambda = 0.75$ to $\lambda = 1.25$ work well for most in-

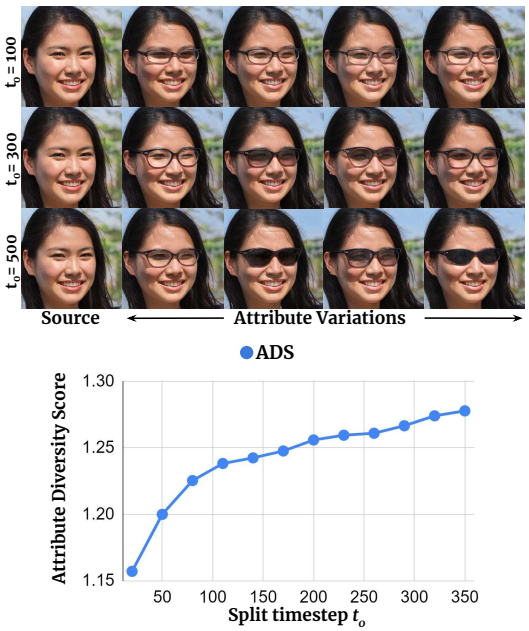

Figure 9. (**Top**) We present generated variations with split timestep $\mathbf{t_0}$. Splitting early ($\mathbf{t_0} = 500$) results in large variations in terms of shades and shapes, whereas splitting late ($\mathbf{t_0} = 100$) results in fine-variation of similar eyeglasses. (**Bottom**) Quantitatively, we observe that early splitting results in high attribute diversity (high ADS), and late splitting generates only fine variations.

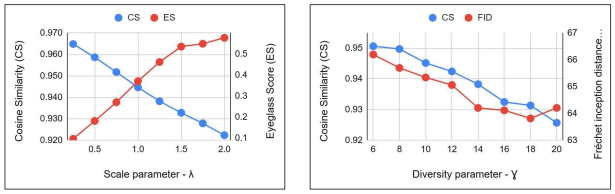

Figure 10. We ablate over different values of scale parameter λ and diversity parameter γ to analyze their impact on the eyeglass attribute. a) An increase in λ results in an increase in attribute score for eyeglasses by the identity is distorted, indicated by lower CS scores. b) Increase in γ results in lower FID, indicating high diversity at the cost of lower CS.

puts on all experimented attributes. We present qualitative results for the ablations in Fig.3 & 4 in SM.

4.5. Data and Latent Space Generalization

Data Generalization. We present results on out-of-domain painting images from Metfaces [22] in Fig. 11(Top). For Metfaces, we generate multiple attribute edit directions from our diffusion models trained with real image pairs as explained in Sec 3.2. We can observe that the generated directions generalize well to the out-of-domain painting images and generate diverse attribute edits. Notably, the styles of the generated edits blend naturally with the painting styles despite domain shift, without looking like the real domain on which the model was trained. Additionally, we provide results for cars and churches attribute variations

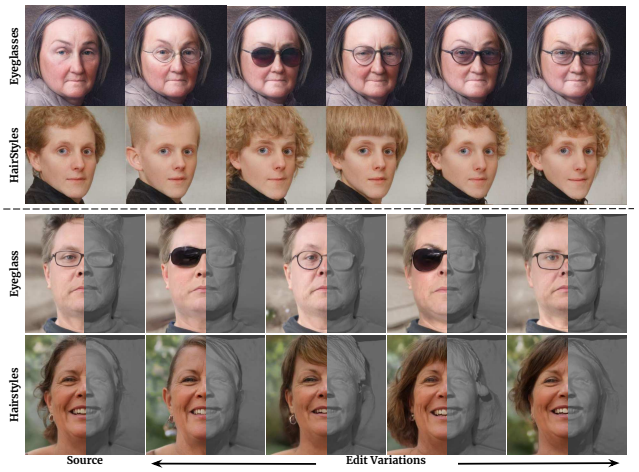

Figure 11. (Top) Results for diverse attribute editing on out-of-domain painting images from Metfaces. (Bottom) Diverse eyeglasses and hairstyle editing on 3D aware GAN-EG3D.

generated hierarchically in Fig. 8. Our method generates high-quality attribute variations explored in a coarse-to-fine manner. Additional results are provided in the SM.

3D GAN Generalization. Our method generalizes beyond 2D StyleGANs to EG3D [8], a 3D-aware generative model. We train DM on edit directions from EG3D's latent space with $10K$ image pairs, and randomly sample directions for editing following Sec. 3.3. In Fig. 11(Bottom), it can be observed that our method can generate diverse attribute edits while maintaining 3D consistency and the subject's identity. We can observe the shape changes associated with eyeglass edits in the geometry of the edited outputs. Additional visualization of EG3D results is provided in the SM.

5. Conclusion

This work explores a challenging problem of diverse attribute editing by using pretrained style-based GANs. Existing methods for attribute editing are limited to generating unidirectional attribute edits. To generate multiple attribute edits, the proposed method trains a diffusion model on edit directions in the latent space. Further, a novel coarse-to-fine sampling strategy is proposed to guide the exploration of attribute variations in an intuitive hierarchical manner. The proposed method works well for diverse editing of several attributes and generalizes to editing in 3D-aware GANs. The primary limitation is that the method inherits the inaccuracies in the existing editing methods and the GAN encoder models used to generate the dataset. Further, as the proposed method works in the latent space of StyleGANs, it is limited to generating attribute variations of data distributions in which StyleGAN models perform well.

Acknowledgments. We thank Tejan Karmali for providing helpful comments and reviewing the draft. Rishubh Parihar is supported by PMRF fellowship and KIAC, IISc.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019.

[2] Rameen Abdal, Peihao Zhu, John Femiani, Niloy Mitra, and Peter Wonka. Clip2stylegan: Unsupervised extraction of stylegan edit directions. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022.

[3] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (TOG)*, 40(3):1–21, 2021.

[4] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Only a matter of style: Age transformation using a style-based regression model. *ACM Transactions on Graphics (TOG)*, 40(4):1–12, 2021.

[5] Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit H Bermano. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. *arXiv preprint arXiv:2111.15666*, 2021.

[6] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023.

[7] Lucy Chai, Jonas Wulff, and Phillip Isola. Using latent space regression to analyze and leverage compositionality in gans. *arXiv preprint arXiv:2103.10426*, 2021.

[8] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022.

[9] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, Jingyi Yu, and Gang Yu. Executing your commands via motion diffusion in latent space. *arXiv preprint arXiv:2212.04048*, 2022.

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

[11] Tim Esler. Github - face recognition using pytorch. *https://github.com/timesler/facenet-pytorch*, 2021.

[12] Rohit Gandikota, Joanna Materzyńska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: Lora adaptors for precise control in diffusion models. In *European Conference on Computer Vision*, pages 172–188. Springer, 2025.

[13] Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. FFJORD: free-form continuous dynamics for scalable reversible generative models. *CoRR*, abs/1810.01367, 2018.

[14] Jiayi Guo, Xingqian Xu, Yifan Pu, Zanlin Ni, Chaofei Wang, Manushree Vasu, Shiji Song, Gao Huang, and Humphrey Shi. Smooth diffusion: Crafting smooth latent spaces in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7548–7558, 2024.

[15] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in Neural Information Processing Systems*, 33:9841–9850, 2020.

[16] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

[17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[19] Xueqi Hu, Qiusheng Huang, Zhengyi Shi, Siyuan Li, Changxin Gao, Li Sun, and Qingli Li. Style transformer for image inversion and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11337–11346, 2022.

[20] Omer Kafri, Or Patashnik, Yuval Alaluf, and Daniel Cohen-Or. Stylefusion: A generative model for disentangling spatial segments. *arXiv preprint arXiv:2107.07437*, 2021.

[21] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[22] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, 33:12104–12114, 2020.

[23] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[24] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.

[25] Gwanghyun Kim and Se Young Chun. Datid-3d: Diversity-preserved domain adaptation using text-to-image diffusion for 3d generative model, 2023.

[26] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013.

[27] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. In *The Eleventh International Conference on Learning Representations*.

[28] Daiqing Li, Huan Ling, Seung Wook Kim, Karsten Kreis, Adela Barriuso, Sanja Fidler, and Antonio Torralba. Bigdatasetgan: Synthesizing imagenet with pixel-wise annotations, 2022.

[29] Hanbang Liang, Xianxu Hou, and Linlin Shen. Ssflow: Style-guided neural spline flows for face image manipulation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 79–87, 2021.

[30] Huan Ling, Karsten Kreis, Daiqing Li, Seung Wook Kim, Antonio Torralba, and Sanja Fidler. Editgan: High-precision semantic image editing. *Advances in Neural Information Processing Systems*, 34, 2021.

[31] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[32] Justin Lovelace, Varsha Kishore, Chao Wan, Eliot Shekhtman, and Kilian Weinberger. Latent diffusion for language generation. *arXiv preprint arXiv:2212.09462*, 2022.

[33] Daniel McDuff, Shuang Ma, Yale Song, and Ashish Kapoor. Characterizing bias in classifiers using generative models. *Advances in neural information processing systems*, 32, 2019.

[34] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

[35] Rishubh Parihar, Abhijnya Bhat, Abhipsa Basu, Saswat Mallick, Jogendra Nath Kundu, and R Venkatesh Babu. Balancing act: Distribution-guided debiasing in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6668–6678, 2024.

[36] Rishubh Parihar, Ankit Dhiman, Tejan Karmali, and R. Venkatesh Babu. Everything is there in latent space: Attribute editing and attribute style manipulation by stylegan latent space exploration. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1828–1836, 2022.

[37] Rishubh Parihar, VS Sachidanand, Sabariswaran Mani, Tejan Karmali, and R Venkatesh Babu. Precisecontrol: Enhancing text-to-image diffusion models with fine-grained attribute control. In *European Conference on Computer Vision*, pages 469–487. Springer, 2025.

[38] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023.

[39] Or Patashnik, Daniel Garibi, Idan Azuri, Hadar Averbuch-Elor, and Daniel Cohen-Or. Localizing object-level shape variations with text-to-image diffusion models. *arXiv preprint arXiv:2303.11306*, 2023.

[40] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021.

[41] William Peebles, Jun-Yan Zhu, Richard Zhang, Antonio Torralba, Alexei Efros, and Eli Shechtman. Gan-supervised dense visual alignment. In *CVPR*, 2022.

[42] Walter HL Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F Da Costa, Virginia Fernandez, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. Brain imaging generation with latent diffusion models. In *Deep Generative Models: Second MICCAI Workshop, DGM4MICCAI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings*, pages 117–126. Springer, 2022.

[43] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10619–10629, 2022.

[44] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2287–2296, 2021.

[45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[46] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

[47] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022.

[48] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9243–9252, 2020.

[49] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1532–1540, 2021.

[50] Nurit Spingarn-Eliezer, Ron Banner, and Tomer Michaeli. Gan "steerability" without optimization. *arXiv preprint arXiv:2012.05328*, 2020.

[51] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6142–6151, 2020.

[52] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021.

[53] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12863–12872, 2021.

[54] Yanbo Xu, Yueqin Yin, Liming Jiang, Qianyi Wu, Chengyao Zheng, Chen Change Loy, Bo Dai, and Wayne Wu. Transeditor: Transformer-based dual-space gan for highly controllable facial editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7683–7692, 2022.

[55] Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge J. Belongie, and Bharath Hariharan. Pointflow: 3d

point cloud generation with continuous normalizing flows. *CoRR*, abs/1906.12320, 2019.

[56] Huiting Yang, Liangyu Chai, Qiang Wen, Shuang Zhao, Zixun Sun, and Shengfeng He. Discovering interpretable latent space directions of gans beyond binary attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12177–12185, 2021.

[57] Oğuz Kaan Yüksel, Enis Simsar, Ezgi Gülperi Er, and Pinar Yanardag. Latentclr: A contrastive learning approach for unsupervised discovery of interpretable directions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14263–14272, 2021.

[58] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. Lion: Latent point diffusion models for 3d shape generation. *arXiv preprint arXiv:2210.06978*, 2022.

[59] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10145–10155, 2021.

[60] Wanfeng Zheng, Qiang Li, Xiaoyan Guo, Pengfei Wan, and Zhongyuan Wang. Bridging clip and stylegan through latent alignment for image editing. *arXiv preprint arXiv:2210.04506*, 2022.