Hierarchical Diffusion Autoencoders and Disentangled Image Manipulation

Zeyu Lu[1,2] Chengyue Wu[3] Xinyuan Chen[2] Yaohui Wang[2] Lei Bai[2] Yu Qiao[2] Xihui Liu[3]

[1] Shanghai Jiao Tong University [2] Shanghai Artificial Intelligence Laboratory
[3] The University of Hong Kong

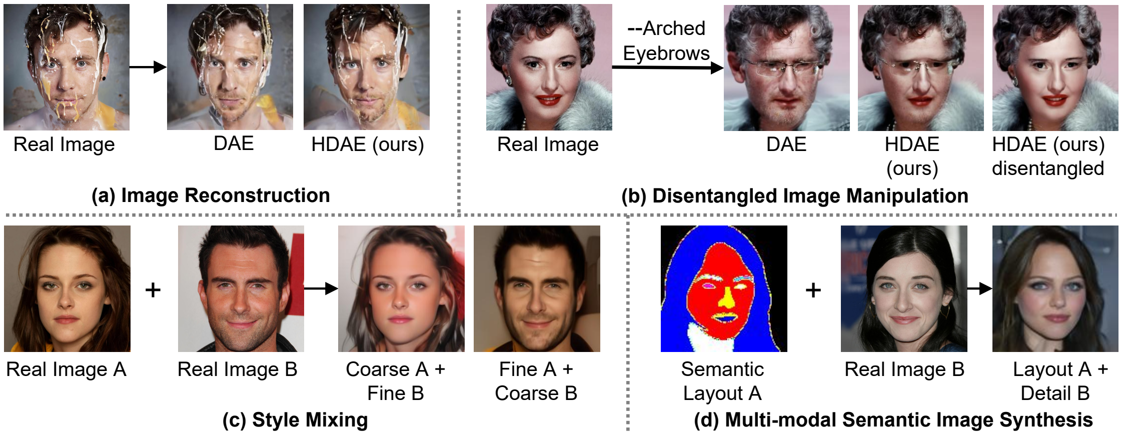

Figure 1. **Applications of Hierarchical Diffusion Autoencoders (HDAE).** (a) Near-perfect image reconstruction. (b) Disentangled image manipulation. Our approach disentangles "arched eyebrows" with other related attributes such as "female" and "eyeglasses". (c) Style mixing with different levels of features from different images. (d) Multi-modal semantic image synthesis with a layout image and a real image providing information on style and details.

Abstract

Diffusion models have attained impressive visual quality for image synthesis. However, how to probe and manipulate the latent space of diffusion models has not been extensively explored. Prior work diffusion autoencoders encode the semantic representations with a single latent code, neglecting the low-level details and leading to entangled representations. To mitigate those limitations, we propose Hierarchical Diffusion Autoencoders (HDAE) that exploits the coarse-to-fine feature hierarchy for the latent space of diffusion models. Our HDAE converges 2+ times faster and encodes richer and more comprehensive coarse-to-fine representations of images. The hierarchical latent space inherently disentangles different semantic levels of features. Furthermore, we propose a truncated feature based approach for disentangled image manipulation. We demonstrate the effectiveness of our proposed HDAE with extensive experiments and applications on image reconstruction, style mixing, controllable interpolation, image editing, and multi-modal semantic image synthesis. The code will be released upon acceptance.

1. Introduction

Diffusion models [14, 47] have demonstrated impressive image generation quality and achieved remarkable success in various applications, such as text-to-image generation [30, 35], image editing [13, 32, 41], and inpainting [26, 40].

A semantically meaningful, editable, and decodable latent space is of particular importance in interpreting generative models as well as applications such as image editing. There have been various works on designing and manipulating the latent space of GANs [12, 44, 45]. However, the latent space of diffusion models has been underexplored. Preechakul *et al*. proposed Diffusion Autoencoders (DAE) [32], which leverages a learnable encoder to discover high-level semantic representations and leverages a diffusion model to encode the stochastic variations and decode images from the semantic latent code and the stochastic latent code.

However, the semantic latent code of diffusion autoencoders is simply represented by a single feature vector predicted by the last layer features of the semantic encoder, ignoring the rich low-level and mid-level features. Such a latent space is insufficient to encode the rich information from images. In practice, we observe that the fine-grained representations (*e.g.*, background and low-level details) are omitted by the semantic encoder and only encoded by the stochastic encoder, leading to difficulties in manipulating those fine attributes. Furthermore, different levels of representations are entangled within the single holistic semantic latent code, making it difficult to find a single direction in the latent space to manipulate a specific attribute without affecting other attributes.

To mitigate those problems, we design the **Hierarchical Diffusion Autoencoders (HDAE)** that exploits the coarse-to-fine feature hierarchy of the semantic encoder and the diffusion-based decoder for comprehensive and disentangled representations. Our design of the hierarchical latent space is motivated by the observation that feature maps at different scales correspond to different abstraction levels of features. The high-resolution feature maps contain low-level features (*e.g.*, color, texture, and details) and the low-resolution feature maps contain high-level features (*e.g.*, structure, layout, abstract attributes). In particular, we extract different levels of features from the semantic encoder and use them to predict the semantic latent code for the corresponding feature levels of the diffusion-based decoder. We extensively investigate different design choices of HDAE, as demonstrated in Fig.2. Our HDAE converges three times faster than DAE [32]. Image reconstruction and image editing experiments demonstrate that the latent representations of HDAE are richer and more comprehensive than DAE. Moreover, the different semantic levels of features are inherently disentangled in the hierarchical latent space, and we probe the hierarchical latent space with style mixing and controllable image interpolation experiments.

To further improve the disentanglement of attributes for image manipulation, we propose to conduct manipulation with truncated features, based on the observation that the majority, low-value feature channels are a critical cause of entanglement. Experiments demonstrate that the truncated features facilitate disentangled attribute manipulation of face images, *e.g.*, we can disentangle "old" from "wearing eyeglasses" so as to edit a face image towards older without adding eyeglasses to the face.

In summary, we propose Hierarchical Diffusion Autoencoders which exploits the coarse-to-fine features to obtain a comprehensive and disentangled latent space for diffusion models. We further propose a novel approach for disentangled attribute manipulation with truncated features. Experiments are conducted on FFHQ, CelebA-HQ, and LSUN Cat datasets. We demonstrate that our model can capture rich and disentangled semantic representations with extensive experiments on image reconstruction, style mixing, controllable image interpolation, disentangled image editing, and multi-modal semantic image synthesis, as shown in Fig.1.

2. Related Work

Diffusion models. Diffusion models have shown great capability in image synthesis [**?**, 6, 10, 13, 14, 23, 25–27, 29, 30, 32, 35, 39, 41–43, 47, 49, 50, 52]. Song *et al.* [49] proposed score-based generative models as a way of modeling a data distribution using its gradients. Ho *et al.* [14] proposed denoising diffusion probabilistic models (DDPMs) which achieved high sample quality based on the score-based generative models [49] and the diffusion models [47]. Inspired by the progress, many works improved the sampling speed [6, 48], sampling quality [29, 50], and conditional synthesis [10]. Diffusion models have also shown wide applications in text-to-image generation [30, 35, 39, 43], image translation [27, 42, 52], image editing [13, 23, 32, 41], image inpainting [26, 40], video generation [**?**, 7, 15, 46, 54, 55], audio generation [8, 59] and text generation [5, 22].

Latent space of generative models. Researchers have made attempts to interpret and manipulate the latent space of generative models. The StyleGAN [18] generator maps the random noise vector to a semantically meaningful latent space, inspiring various follow-up works exploring the controllability and interpretability of the latent space of GANs [16, 18, 28, 34, 38, 51, 56]. Various works explored learning based [37, 53, 57], optimization based [1, 2, 16, 56], or hybrid [4, 62] approaches for GAN inversion, aiming to encode an image to the latent space of GANs. GAN inversion enables diverse image editing methods by manipulating the latent code. Shen *et al.* [44] proposed InterfaceGAN which adopts an SVM to find the semantic directions for attribute manipulation. Härkönen *et al.* [12] proposed GANspace which performed PCA on early feature layers. Some methods [24, 60, 63] found distinguishable directions based on mutual information. Recently some works [3, 20, 21, 31] explored image editing guided by CLIP [33] model. In particular, Patashnik *et al.* [31] proposed StyleClip which used the pre-trained CLIP [33] as the loss supervision to match the manipulated results with the text condition. Our work explores the hierarchical latent space of diffusion models.

Latent space of diffusion models. Despite the various work studying the latent space of GANs, the latent space of diffusion models lack semantic meaning and cannot be easily applied for semantic manipulation. Latent diffusion [39] applies diffusion model in the latent space of images instead of the original image space. DiffusionCLIP [20] conducts image editing by DDIM inversion and model finetuning guided by CLIP. However, the latent space of diffusion models cannot be directly manipulated for image editing. Diffusion autoencoders [32] adopts a semantic encoder to obtain a meaningful and decodable latent space for diffusion models. However, such latent space lacks find details and causes attribute entanglement. We propose the Hierarchical Diffusion Autoencoders which provides a comprehensive and hierarchical latent space for diffusion models.

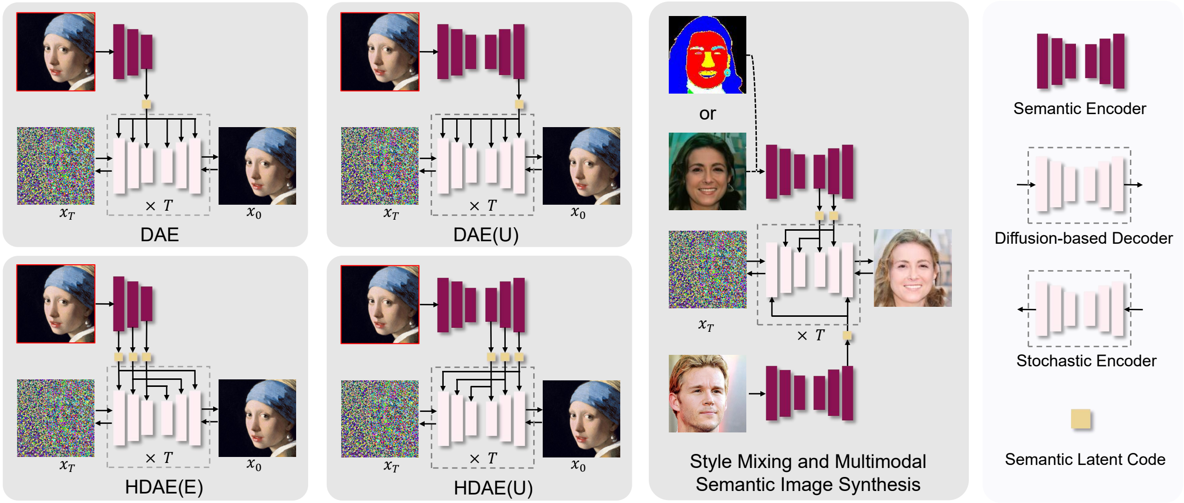

Figure 2. **Overview of different network structures.** In general, diffusion autoencoders [32] apply the semantic encoder to encode the semantic latent code, the stochastic encoder (*i.e*., the DDIM forward process) to encode the stochastic latent code, and the diffusion-based decoder (*i.e*., the DDIM reverse process) to generate images based on the latent codes. DAE and DAE(U) are non-hierarchical diffusion autoencoders. HDAE(E) and HDAE(U) are different variants of our proposed hierarchical diffusion autoencoders. In addition, we show style mixing and multimodal image synthesis with HDAE, where the low-level and high-level latent codes are from different images.

3. Methodology

3.1. Preliminaries

Diffusion probabilistic models. Denoising diffusion probabilistic models (DPMs) [14] is a class of generative models. The forward process defines a Markov chain gradually adding Gaussian noise to an image x_0, generating a sequence of images $x_0, x_1, \cdots, x_T$. The reverse process iteratively removes noise from the noisy image x_t by sampling from $p(x_{t-1}|x_t)$. The noise $\epsilon_\theta(x_t, t)$ is predicted by a U-Net which takes the noisy image x_t and timestep t as input. The model is trained with the L_2 loss between the predicted noise and the actual noise $\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2$

Song *et al*. [48] proposed Denoising Diffusion Implicit Model (DDIM) with a deterministic forward process. By matching the marginal distribution of DDPM, it shares the same training objective and solution with DDPM. We can run the DDIM generation process backward deterministically to obtain the noise map x_T, which represents the stochastic latent codes of the image x_0.

Diffusion autoencoders. In pursuit of a meaningful latent space, Preechakul *et al*. [32] proposed Diffusion Autoencoders. They apply a convolutional neural network as the semantic encoder to encode images into a semantic vector $z_s = \text{Enc}_\phi(x_0)$ and apply the DDIM forward process as a stochastic encoder that encodes the image $x_0 \in \mathbb{R}^{H\times W\times 3}$ to a stochastic variant $x_T \in \mathbb{R}^{H\times W\times 3}$. The DDIM reverse process acts as the decoder which models $p_\theta(x_{t-1}|x_t, z_s)$ and iteratively generates x_0 given the semantic latent code z_s and the stochastic latent code x_T. The DDIM adopts a U-Net structure with shared parameters for each timestep, and the U-Net is conditioned by the semantic code z_s and timestep t by adaptive group normalization layers (AdaGN). The semantic latent vector $z_s \in \mathbb{R}^{512}$ captures the meaningful and decodable representations that can be used for image reconstruction and manipulation.

3.2. Hierarchical Diffusion Autoencoders

The semantic latent code in diffusion autoencoders is a 512-dimensional feature vector. Such representations bring two limitations. Firstly, the single feature vector from the final layer of the semantic encoder omits the low-level and mid-level features, making it not sufficient to encode the comprehensive semantic information in the images. Empirically, we observe that the images reconstruction and manipulation results with diffusion autoencoders suffer from insufficient details. Secondly, the holistic feature representation ignores the intrinsic fine-grained-to-abstract and low-level-to-high-level hierarchy of visual features.

Design space of latent representations and network architectures. To address those issues, we propose hierarchical diffusion autoencoders which explore the hierarchical semantic latent space of diffusion autoencoders. In particular, we keep the architecture of the diffusion-based decoder and explore the design space of the semantic encoder and the latent space, as shown in Fig. 2.

- *DAE,* i.e., *diffusion autoencoders.* The Naïve diffusion autoencoders [32] leverage a naïve CNN-based semantic encoder, where the semantic code is extracted

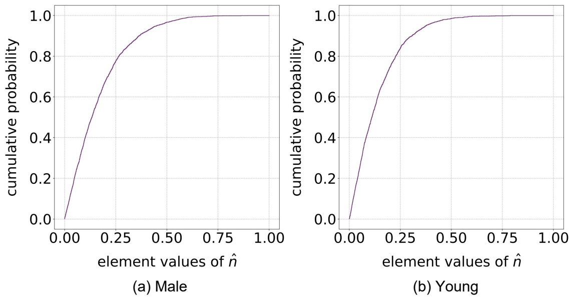

Figure 3. **The empirical cumulative distribution function of the element values in the normalized classifier weights $\hat{n}$.** Most elements of $\hat{n}$ are of low values, and only a few are of high values.

from the last layer of the semantic encoder, followed by global average pooling and fully-connected layers.

– *DAE(U),* i.e., *diffusion autoencoders with U-Net semantic encoder.* We replace the naïve encoder in DAE with a U-Net encoder, denoted by DAE(U). With the skip connections and downsampling-upsampling design, the last layer of the U-Net might be able to capture both low-level and high-level features. The spatial feature map from the last layer of the U-Net is mapped into a 512-dimensional feature vector z_s.

– *HDAE(E),* i.e., *hierarchical diffusion autoencoders with naïve semantic encoder.* To exploit the feature hierarchy of the diffusion autoencoders, we extract different semantic levels of feature maps from the semantic encoder to predict the hierarchical semantic latent codes $z_s^1, z_s^2, \cdots, z_s^L$. The different levels of semantic features are fed into the corresponding levels of the diffusion-based decoder.

– *HDAE(U),* i.e., *hierarchical diffusion autoencoders with U-Net semantic encoder.* HDAE(U) also adopts the hierarchical latent space design but leverages a semantic encoder with the U-Net structure.

Experiments demonstrate that the hierarchical structures of HDAE(E) and HDAE(U) provide richer semantic representations and more efficient training, and that HDAE(U) is the best-performed architecture design.

Advantages and applications of HDAE. Firstly, while the latent space of DAE lacks low-level details, the hierarchical latent space of HDAE encodes comprehensive fine-grained-to-abstract and low-level-to-high-level features, leading to more accurate and detail-preserving image reconstruction and manipulation results. Secondly, the feature hierarchy naturally enables applications such as style mixing (illustrated in Fig. 2), multimodal semantic image synthesis, and controllable image interpolation, as demonstrated in Sec. 4.

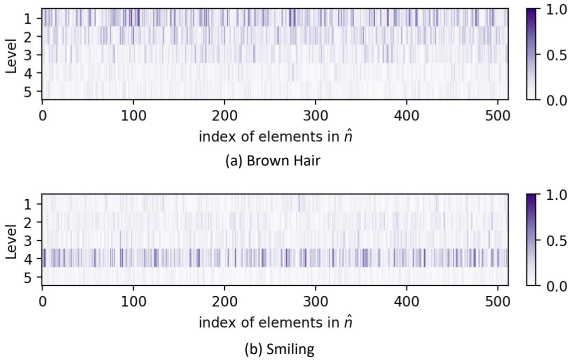

Figure 4. **The values of the 5×512-dimensional $\hat{n}$, visualized by levels.** (1) Most values are of low values and truncating those values lead to better disentanglement. (2) Feature hierarchy is shown.

3.3. Disentangled Image Manipulation with Truncated Features

With a linear classifier trained with the semantic vectors, diffusion autoencoders [32] can be applied for image manipulation. We can edit an image by moving the semantic vector towards the direction n, which is obtained from the weights of the linear classifier for the target attribute.

A critical issue for image editing is the entanglement of features in the latent space. For example, in face editing, “old” is often entangled with “wearing glasses”, and “arched eyebrows” is often entangled with “female”. By analyzing the distribution of the classifier direction n, we propose an approach to disentangle attributes by adjusting n.

In our HDAE models with L hierarchical layers, we concatenate the 512-dimensional semantic codes from each of the L layers into a single vector and derive the classifier direction $n \in \mathbb{R}^{512 \times L}$. We derive the normalized classifier weights $\hat{n}$ as follows:

$$\hat{n}_i = \frac{|n_i| - \min_i(|n_i|)}{\max_i(|n_i|) - \min_i(|n_i|)} \quad (1)$$

As shown by the empirical cumulative distribution function and the visualization of the values of $\hat{n}$ in Fig. 3 and Fig. 4, most values are relatively low, and only a few values are high. We hypothesize that the few high-value elements indicate the dominant direction of an attribute classifier, while the majority, low-value elements are noisy and may lead to attribute entanglement. In particular, we denote the set of the top-k largest values of $\hat{n}$ as top-k$(\hat{n})$, and truncate the n accordingly as follows.

$$n'_i = \begin{cases} n_i, \text{ if } \hat{n}_i \in \text{top-k}(\hat{n}) \\ 0, \text{ else} \end{cases} \quad (2)$$

Our experiments in Fig. 10 validate that the truncated n' leads to better-disentangled image manipulation.

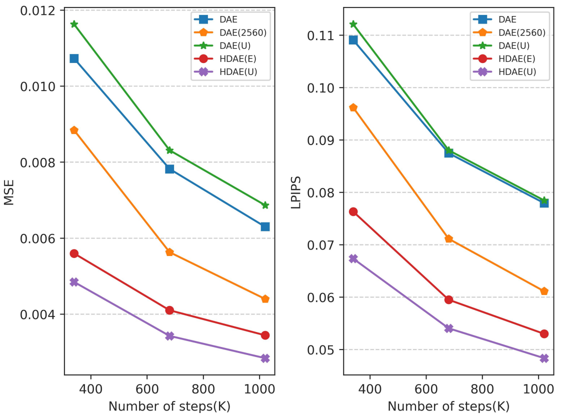

Figure 5. **Ablation study of different architecture designs for image reconstruction on FFHQ dataset.** MSE and LPIPS are evaluated on the test set. Both variants of HDAE outperform DAE baselines. HDAE(U) achieves the best image reconstruction performance. We adopt HDAE(U) as our best model in experiments.

4. Experiments

4.1. Experimental settings

Following Diffusion Autoencoders [32], we train the hierarchical diffusion autoencoders on FFHQ [18] dataset and train the attribute classifiers on CelebA-HQ [17] dataset.

The ablation study in Sec. 4.2 is trained on 128 × 128 images and other results are derived from models trained on 256 × 256 images. More implementation details and experimental settings can be found in the appendix.

4.2. Ablation Study

We conduct ablation studies to test the effectiveness of the design choices of diffusion autoencoders and hierarchical diffusion autoencoders demonstrated in Fig. 2. We split the FFHQ dataset into 65,000 images for training and 5,000 for testing. For a fair comparison, we add another baseline, denoted as DAE(2560), where the semantic code dimension of DAE is expanded from 512 to 2560, sharing the same size as the semantic code dimension (512 × 5) of our HDAE models. We report the image reconstruction quality on the test set, evaluated by pixel-wise MSE and the perceptual quality, in Fig. 5. We draw the following conclusions:

- *The hierarchical diffusion autoencoders perform much significantly better than non-hierarchical ones.* As illustrated in Fig. 5, both variants of hierarchical diffusion autoencoders, HDAE(E) and HDAE(U), perform significantly better and converge much faster than DAE, DAE(2560), and DAE(U). The image reconstruction performances of HDAE(E) and HDAE(U) at 340K steps are better than the performances of DAE and DAE(U) at 1,020K steps.

Model	Setting	SSIM↑	LPIPS↓	MSE↓
StyleGAN2($\mathcal{W}$) [19]	-	0.677	0.168	0.016
StyleGAN2($\mathcal{W}+$) [19]	-	0.827	0.114	0.006
VQ-GAN [11]	-	0.782	0.109	3.61e-3
VQ-VAE2 [36]	-	0.947	0.012	4.87e-4
HFGI [53]	-	0.877	0.127	0.0617
DDIM [48]	T=100, 128^2	0.917	0.063	0.002
DAE [32]	T=100, 128^2, random x_T	0.677	0.073	0.007
HDAE(U) (ours)	T=100, 128^2, random x_T	0.793	0.038	2.96e-3
DAE [32]	T=100, 128^2, encoded x_T	0.991	0.011	6.07e-5
HDAE(U) (ours)	T=100, 128^2, encoded x_T	**0.993**	**0.009**	**5.01e-5**

Table 1. **Image reconstruction evaluation of models trained on FFHQ and tested on CelebA-HQ.** HDAE(U) outperforms DAE with random stochastic code x_T, indicating that HDAE(U) encodes richer details than DAE. HDAE(U) with encoded x_T achieves state-of-the-art, near-perfect reconstruction.

Task	HDAE(U)	DAE	HDAE(U)+TF	DAE+TF	Similar
Image Reconstruction	82.5%	11.7%	-	-	5.8%
Image Manipulation	65.7%	3.6%	-	-	30.7%
Disentangled Image Manipulation	22.8%	1%	76.2%	0%	-

Table 2. **Human perceptual evaluation on image reconstruction, image manipulation and disentangled image manipulation.** Since DAE and HDAE(U) achieve near-perfect reconstruction for naked human eyes, we conduct the human perceptual evaluation on image reconstruction with random stochastic code x_T. "+TF" denotes image manipulation with truncated features.

- *The U-Net structure for the semantic encoder has a negative effect on DAE, but benefits HDAE.* As demonstrated in Fig. 5, DAE(U) performs worse than DAE, but HDAE(U) performs better than HDAE(E).

- *HDAE(U) is the best-performing model.* Therefore, we adopt HDAE(U) in the following experiments.

4.3. Image Reconstruction

Setup. Image reconstruction quality reflects how well the latent representations encode the image information, especially the details. We evaluate the image reconstruction quality of different approaches: GAN inversion methods (StyleGAN2 [19] with $\mathcal{W}$ space, pretrained StyleGAN2 with $\mathcal{W}+$ space, VQ-GAN [11], HFGI [53]), VAE-based method (VQ-VAE2 [36]), and diffusion-based methods (DDIM [48], and DAE [32], and our HDAE(U)). All these models are trained on FFHQ [18] and tested on 30,000 images from CelebA-HQ [17]. To validate the potential of our model beyond the face domain, we show image reconstruction results on cat datasets by training the models on LSUN [58] Cat and testing on the AFHQ Cat [9]. In addition, we show our image reconstruction results on LSUN Bedroom and LSUN Horse datasets. DDIM, DAE, and HDAE(U) are trained on images of size 128 × 128 and we use $T = 100$ for inference. Structural Similarity SSIM, Pixel-wise MSE, and perceptual quality metric LPIPS [61] are adopted to evaluate the image reconstruction quality. To illustrate what information is encoded by the stochastic encoder, we show the

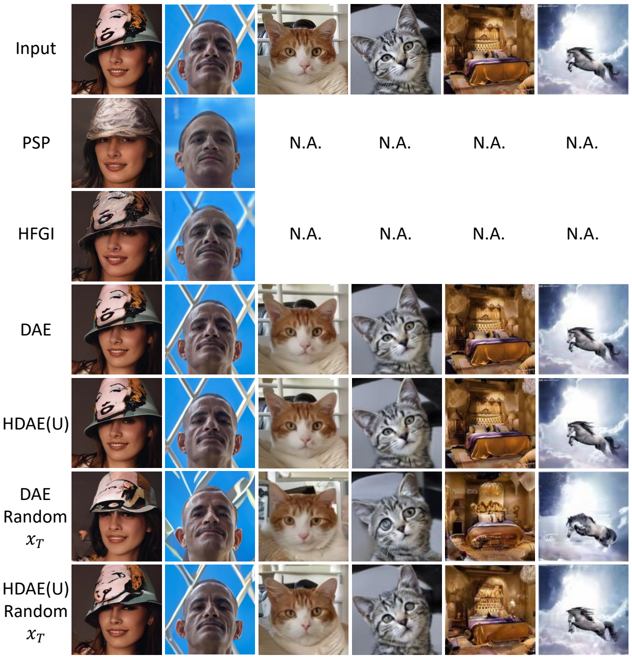

Figure 6. **Quantitative results of image reconstruction.** We evaluate the image reconstruction experiments on the face, cat, bedroom, and horse images. HDAE(U) with random stochastic code x_T (7th row) preserves more details in backgrounds, appearance, expressions, and identity information than DAE with random stochastic code x_T (6th row). HDAE(U) and DAE with encoded stochastic code x_T (4th and 5th row) attain near-perfect reconstruction. "N.A." denotes not applicable for this model.

images reconstructed from DAE and HDAE(U) with their corresponding x_T as well as random x_T for comparison.

Quantitative results. The results Tab. 1 demonstrate that: (1) HDAE(U) achieves near-perfect image reconstruction performance, outperforming previous GAN inversion methods, VAE-based approaches, and diffusion-based approaches. (2) Comparing image reconstruction results with the x_T encoded by a stochastic encoder and random x_T, some detail-related information is encoded by the stochastic encoder. (3) The image reconstruction performance degrades more for DAE than HDAE(U) when replacing x_T encoded by the stochastic encoder with a random x_T, indicating that in HDAE(U) more information is encoded in the semantic encoder and less information is encoded by the stochastic encoder.

Qualitative results. The qualitative results of image reconstruction are shown in Fig. 6. Previous GAN inversion approaches PSP [37] and HFGI [53] cannot preserve background and details of the original images, while DAE and HDAE(U) with encoded x_T reconstruct images nearly identical to the input images. Comparing *DAE with random* x_T with *HDAE(U) with random* x_T, we find that HDAE(U) with random x_T preserves background and details better. DAE with random x_T fails to reconstruct the details, such as the eyes of the cat, the interior decoration of the bedroom, and the legs of the horse. This observation indicates that the semantic encoder of HDAE(U) encodes more comprehensive features, allowing detail-preserving applications such as image manipulation. More examples are in the appendix.

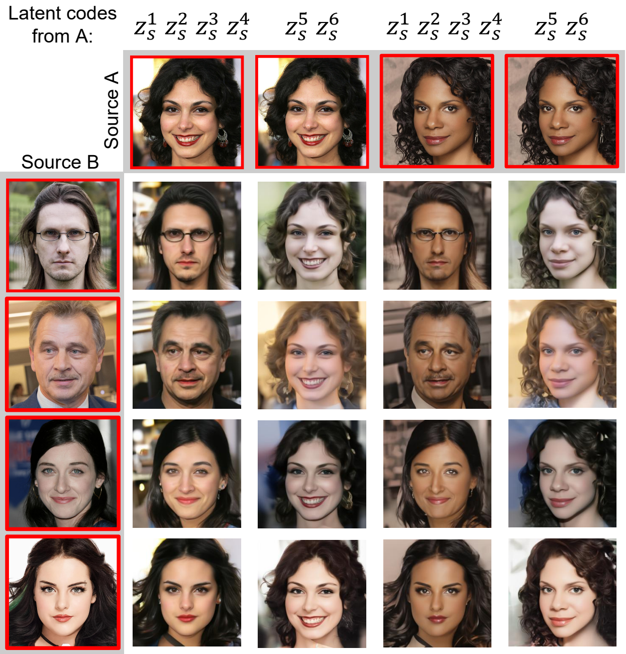

Figure 7. **Style mixing results with hierarchical latent space.** z_s^1,z_s^2,z_s^3,z_s^4 represent the low-level latent codes and z_s^5,z_s^6 represent the high-level latent codes. Given the real images in red boxes, we can mix the high-level latent codes from source A(B) and low-level latent codes from source B(A).

Human perceptual evaluation. We conduct a human perceptual evaluation, where users are asked to vote for the image reconstruction quality of HDAE(U) with random x_T and DAE with random x_T in Tab. 2. We collect 450 votes from 15 participants, and the results are shown in Tab. 2. Users clearly prefer the image reconstruction results by our HDAE(U) over DAE. More details are in the appendix.

4.4. Interpreting the Hierarchical Latent Space

The plot in Fig. 4 indicates a strong correlation between feature levels and attributes. For instance, "brown hair" is correlated with the low-level feature layers 1 and 2, while "smiling" is correlated with the high-level feature layer 4. It reveals the hierarchical latent space where different layers represent different abstraction levels of the representations. We visualize the latent space hierarchy by style mixing and controllable image interpolation experiments.

Style mixing. To interpret and visualize the hierarchical latent space, we mix the high-level latent codes z_s^5, z_s^6 from image A(B) and the low-level latent codes z_s^1, z_s^2, z_s^3, z_s^4 from image B(A), and use the mixed latent codes to generate a

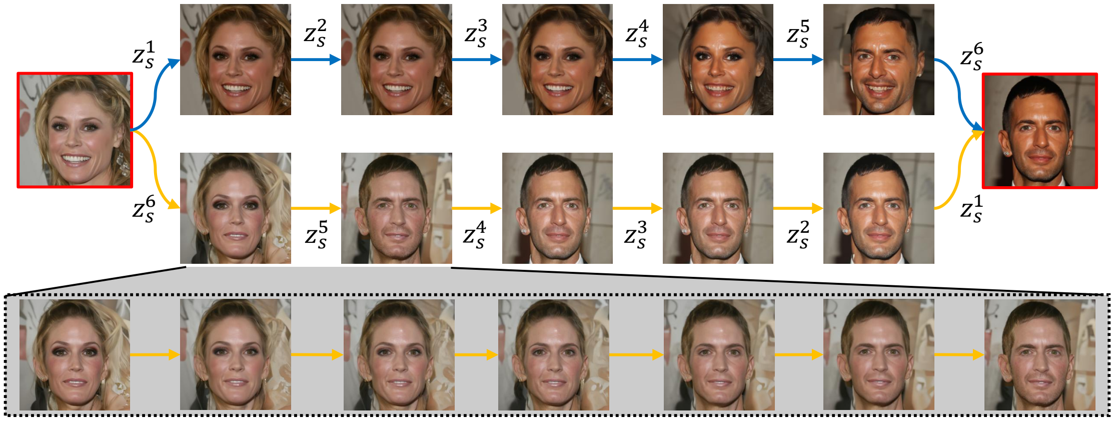

Figure 8. **Controllable image latent space interpolation along different paths.** Given the real images in red boxes, the first path shows the image interpolation changing from low-level latent codes to high-level latent codes. The second path shows the reverse interpolation process from high-level to low-level. The third path shows the smooth interpolation between two images by smoothly interpolating z_s^5.

new image, as shown in Fig. 2. Fig. 7 shows a clear hierarchy of the latent space. $z_s^1, z_s^2, z_s^3, z_s^4$ control the spatial details such as background, color, and lighting, and z_s^5, z_s^6 control the high-level semantics and image structure such as pose, gender, face shape, and eyeglasses.

Controllable image interpolation. Image interpolation based on semantic codes is a common way to visualize and verify the properties of the latent space. With our hierarchical latent space, we can control different paths of image interpolation, as shown in Fig. 8. Given the leftmost and rightmost real images in red boxes, in the first row, we interpolate from left to right by changing low-level features first and then high-level features. In the second row, the interpolation is conducted in a reverse way, from high-level features to low-level features. In the third row, we illustrate the continuous changes between the two images in the second row. Results indicate that z_s^1, z_s^2, z_s^3 control lighting and color, z_s^4 controls background, z_s^5 controls gender and z_s^6 controls pose.

4.5. Image Manipulation

Detail-preserving image manipulation with HDAE. With the linear attribute classifiers for the latent codes, we can edit real images by manipulating the semantic latent codes with the classifier direction. Fig. 9 shows the image manipulation results on HFGI [53], StyleClip [31], DAE [32], and our HDAE(U). DAE and HDAE(U) are trained on FFHQ, and the linear attribute classifiers are trained on the CelebA-HQ. As demonstrated in Fig. 9, our HDAE(U) preserves the details (*e.g*., background, face identity, and the forehead pendant) of the input image better than other approaches. The image manipulation results demonstrate that the representations learned by our HDAE are rich and semantically meaningful. More examples can be found in the appendix.

Disentangled image manipulation with truncated features. As introduced in Sec. 4.5, we leverage the truncated features for disentangled image manipulation. We compare the qualitative results of image manipulation with truncated features ($k = 24$) and without truncated features ($k = 3072$) on DAE and HDAE(U). We also compare with GAN-based methods HFGI [53] and StyleClip [31]. As shown in Fig. 10, for the naïve manipulation without truncation, "old" is entangled with "eyeglasses", "arched eyebrows" is entangled with "female", and "female" is entangled with "makeup". Our HDAE(U) with truncated features successfully disentangles those attributes and gives the best results compared with other methods. The truncated features with $k = 24$ effectively disentangle the attributes for manipulation, and the attributes become more entangled as k increases (from $k = 24$ to $k = 3072$). An ablation study of k and more examples are in the appendix.

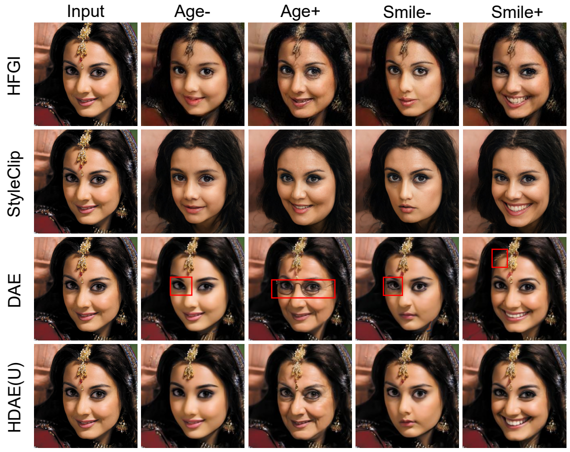

Figure 9. **Comparisons on real image manipulation between HFGI [53], StyleClip [31], DAE [32] and HDAE(U).** HDAE(U) preserves the details (*e.g*., face identity, background, and the forehead pendant) in the original image better than other approaches.

Human perceptual evaluation. We conduct user perceptual evaluations on the image manipulation and disentangled

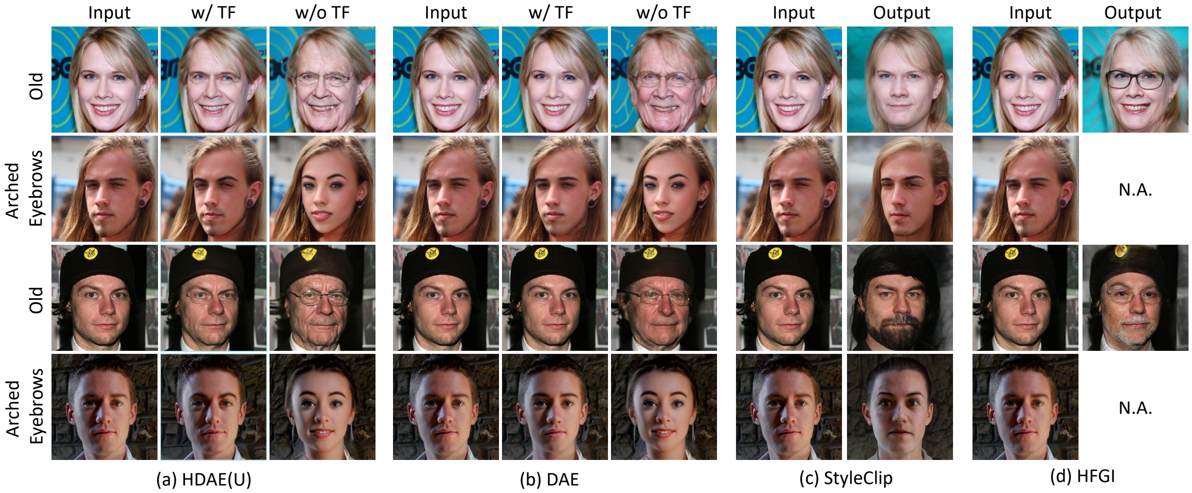

Figure 10. **Disentangled attribute manipulation results.** HDAE(U) is trained on FFHQ [18] with image resolutions 256 × 256. We preserve the top k largest values for the truncated features. "w/ TF" denotes using truncated features and $k = 24$ gives the best manipulation results in terms of attribute disentanglement (see appendix for ablation on k). "w/o TF" is the comparison experiment without truncated features, where we can observe severe attribute entanglements: old - wearing eyeglasses, arched eyebrows - female, female - makeup. "N.A." denotes not applicable for this model.

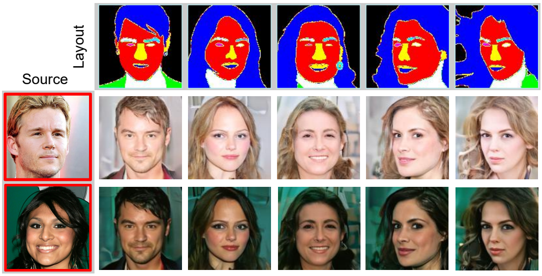

Figure 11. **Multi-modal semantic image synthesis.** Our HDAE conducts style mixing with the layout information from the semantic layout map and details from the source image.

image manipulation results of HDAE(U), DAE, HDAE(U) with truncated features and DAE with truncated features. For image manipulation, we collect 1,575 votes from 15 participants. For disentangled image manipulation, we collect 2,100 votes from 15 participants. Results in Tab. 2 indicate that HDAE(U) performs better than DAE, and that HDAE(U) with truncated features performs better than HDAE(U), DAE, and DAE with truncated features. Details are in the appendix.

Discussion. The hierarchical latent space and truncation-based approach are orthogonal approaches to improve image manipulation from different perspectives. The feature hierarchy provides a comprehensive and semantically meaningful latent space. The truncation-based approach empowers disentangled image manipulation. Therefore, HDAE(U) with truncated features shows the best detail-preserving and disentangled image manipulation results.

4.6. Other Applications

Multi-modal semantic image synthesis. We train an extra layout encoder that maps the semantic label map into the latent space of HDAE. HDAE can synthesize images based on the high-level features from a semantic label map and the low-level features from a real image. We can control the layout with the label map, and control the style and details with the image, as shown in Fig. 11.

Unconditional image synthesis. By training a latent DDIM model to predict the latent codes, our model can be leveraged for unconditional image synthesis. Results and more details can be found in the appendix.

5. Conclusion

We present Hierarchical Diffusion Autoencoders (HDAE) which leverages the feature hierarchy to build a hierarchical latent space for diffusion models. The latent representations are rich and comprehensive, with a coarse-to-fine hierarchy. We further propose a novel disentangled attribute manipulation approach with truncated features. Extensive experiments and applications on image reconstruction, style mixing, controlled image interpolation, disentangled image editing, and multimodal semantic image synthesis are conducted to validate the effectiveness of our approach.

Acknowledgments: This research is funded by Shanghai Artificial Intelligence Laboratory. This work is partially supported by the National Key R&D Program of China (NO.2022ZD0160100). This work is partially supported by the National Natural Science Foundation of China under Grants No. 62102150. This work is supported in part by HKU Startup Fund, HKU Seed Fund for Basic Research, HKU Seed Fund for Translational and Applied Research, and HKU-SCF FinTech Academy R&D Funding Scheme.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *ICCV*, 2019. 2

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *CVPR*, 2020. 2

[3] Rameen Abdal, Peihao Zhu, John Femiani, Niloy J. Mitra, and Peter Wonka. Clip2stylegan: Unsupervised extraction of stylegan edit directions. In *SIGGRAPH*. ACM, 2022. 2

[4] Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit Bermano. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In *CVPR*, 2022. 2

[5] Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *NeurIPS*, 2021. 2

[6] Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. In *ICLR*, 2022. 2

[7] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023. 2

[8] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J. Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. *arxiv:2009.00713*, 2020. 2

[9] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, 2020. 5

[10] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 2

[11] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 5

[12] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable GAN controls. In *NeurIPS*, 2020. 1, 2

[13] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arxiv:2208.01626*, 2022. 1, 2

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 1, 2, 3

[15] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 2

[16] Kyoungkook Kang, Seongtae Kim, and Sunghyun Cho. GAN inversion for out-of-range images with geometric transformations. In *ICCV*, 2021. 2

[17] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *NeurIPS*, 2018. 5

[18] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2, 5, 8

[19] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 5

[20] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *CVPR*, 2022. 2

[21] Umut Kocasari, Alara Dirik, Mert Tiftikci, and Pinar Yanardag. Stylemc: Multi-channel based fast text-guided image generation and manipulation. In *WACV*, 2022. 2

[22] Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. Diffusion-lm improves controllable text generation. *arxiv:2205.14217*, 2022. 2

[23] Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance. *arxiv:2112.05744*, 2021. 2

[24] Yu-Ding Lu, Hsin-Ying Lee, Hung-Yu Tseng, and Ming-Hsuan Yang. Unsupervised discovery of disentangled manifolds in gans. *arxiv:2011.11842*, 2020. 2

[25] Zeyu Lu, Di Huang, Lei Bai, Jingjing Qu, Chengyue Wu, Xihui Liu, and Wanli Ouyang. Seeing is not always believing: Benchmarking human and model perception of ai-generated images. *NeurIPS*, 2023. 2

[26] Andreas Lugmayr, Martin Danelljan, Andrés Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, 2022. 1, 2

[27] 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. In *ICLR*, 2022. 2

[28] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. PULSE: self-supervised photo upsampling via latent space exploration of generative models. In *CVPR*, 2020. 2

[29] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021. 2

[30] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, 2022. 1, 2

[31] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *ICCV*, 2021. 2, 7

[32] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *CVPR*, 2022. 1, 2, 3, 4, 5, 7

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2

[34] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016. 2

[35] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arxiv:2204.06125*, 2022. 1, 2

[36] Ali Razavi, Aäron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *NeurIPS*, 2019. 5

[37] 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 *CVPR*, 2021. 2, 6

[38] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *TOG*, 2021. 2

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2

[40] Andrés Romero, Angela Castillo, Jose Abril-Nova, Radu Timofte, Ritwik Das, Sanchit Hira, Zhihong Pan, Min Zhang, Baopu Li, Dongliang He, Tianwei Lin, Fu Li, Chengyue Wu, Xianming Liu, Xinying Wang, Yi Yu, Jie Yang, Rengang Li, Yaqian Zhao, Zhenhua Guo, Baoyu Fan, Xiaochuan Li, Runze Zhang, Zeyu Lu, Junqin Huang, Gang Wu, Junjun Jiang, Jiayin Cai, Changlin Li, Xin Tao, Yu-Wing Tai, Xiaoqiang Zhou, and Huaibo Huang. NTIRE 2022 image inpainting challenge: Report. In *CVPRW*, 2022. 1, 2

[41] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arxiv:2112.05744*, 2022. 1, 2

[42] Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *SIGGRAPH*. ACM, 2022. 2

[43] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *arxiv:2205.11487*, 2022. 2

[44] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *TPAMI*, 2022. 1, 2

[45] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *CVPR*, 2021. 1

[46] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *ICLR*, 2023. 2

[47] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 1, 2

[48] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 2, 3, 5

[49] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *NeurIPS*, 2019. 2

[50] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. In *NeurIPS*, 2020. 2

[51] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *TOG*, 2021. 2

[52] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. *arxiv:2205.12952*, 2022. 2

[53] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-fidelity GAN inversion for image attribute editing. In *CVPR*, 2022. 2, 5, 6, 7

[54] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023. 2

[55] Yaohui Wang, Xin Ma, Xinyuan Chen, Antitza Dantcheva, Bo Dai, and Yu Qiao. Leo: Generative latent image animator for human video synthesis. *arXiv preprint arXiv:2305.03989*, 2023. 2

[56] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *CVPR*, 2021. 2

[57] Yinghao Xu, Yujun Shen, Jiapeng Zhu, Ceyuan Yang, and Bolei Zhou. Generative hierarchical features from synthesizing images. In *CVPR*, 2021. 2

[58] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. LSUN: construction of a large-scale image dataset using deep learning with humans in the loop. *arxiv:1506.03365*, 2015. 5

[59] Jiashuo Yu, Yaohui Wang, Xinyuan Chen, Xiao Sun, and Yu Qiao. Long-term rhythmic video soundtracker. In *ICML*, 2023. 2

[60] Oguz Kaan Yüksel, Enis Simsar, Ezgi Gülperi Er, and Pinar Yanardag. Latentclr: A contrastive learning approach for unsupervised discovery of interpretable directions. In *ICCV*, 2021. 2

[61] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5

[62] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A. Efros. Generative visual manipulation on the natural image manifold. In *ECCV*, 2016. 2

[63] Peiye Zhuang, Oluwasanmi Koyejo, and Alexander G. Schwing. Enjoy your editing: Controllable gans for image editing via latent space navigation. In *ICLR*, 2021. 2