Image Editing Based on h-space of Diffusion Model: Unsupervised Separation of Shared and Unique Representations

Ziqian Liu[1], Pietro Gori[*], and Yunlong He[*]

[*]Supervisor: pietro.gori@telecom-paris.fr, yunlong.he@telecom-paris.fr

1. Introduction

In recent years, Denoising Diffusion Models (DDM) [11], with their advantages in detailed texture reconstruction and sample diversity, have gradually replaced Variational Autoencoders (VAE) [19, 32] and Generative Adversarial Networks (GAN) [7, 16], becoming mainstream methods in the fields of image generation and editing. The model represented by Stable Diffusion [33] have sparked a surge of research in text-driven and image-driven based image editing. In the field of image editing, most works aim to semantically modify specific regions of an image while keeping other regions unchanged. For example, although these works [3, 6, 46] eventually achieving significant results, they lack interpretability regarding how input conditions influence the intermediate denoising process and remain ambiguous as to the precise control of the generation process. Therefore, exploring the latent space of DDM in depth continues to be an ongoing challenge.

This difficulty is rooted in the fact that Denoising Diffusion Models (DDMs) do not have a low-dimensional and semantically explicit Latent Space like GANs. However, recently Kwon et al. [21] demonstrated that the deepest bottleneck feature space (h-space) of the diffusion model U-Net presents a high degree of linearity, making it suitable as a semantic latent space for image manipulation. This discovery has sparked a surge of research into disentangling semantic directions based on the h-space, enabling the unsupervised discovery of global [4, 30, 47, 9, 44, 23] and local [9, 20, 44] editing directions that independently control various semantic attributes. Although these works have made notable progress in terms of interpretability and controllability, their editing granularity is still mostly limited to the global or local changes of a single attribute, and it is still difficult to realize the accurate isolation and replacement of information in specific regions of an image based on unsupervised realization.

As revealed by related literature, an image can be considered as a composition of multiple conceptual factors [37, 24]. Building upon the ideas from our previous work [1, 26, 25, 31], we summarize the multiple factors used to represent an image into two categories: "Common" and "Salient". We assume that the target sample has incorporated some new patterns unrelated to the background sample while retaining the background content. So statistically, the Common factors shared between the target and background samples should follow the same data distribution, whereas the distinguishing features of the Target are captured exclusively by its Salient factors. It is worth noting that our approach fundamentally focuses on a binary division of "Common vs. Salient", rather than attempting to disentangle all possible independent dimensions in the latent space. Furthermore, compared to previous works [1, 26], we achieve high-quality latent factor separation without relying on external classifiers, significantly enhancing generalization capabilities. Overall, this representation is highly interpretable in image editing applications, enabling the unsupervised separation of Common content and Salient content within an image.

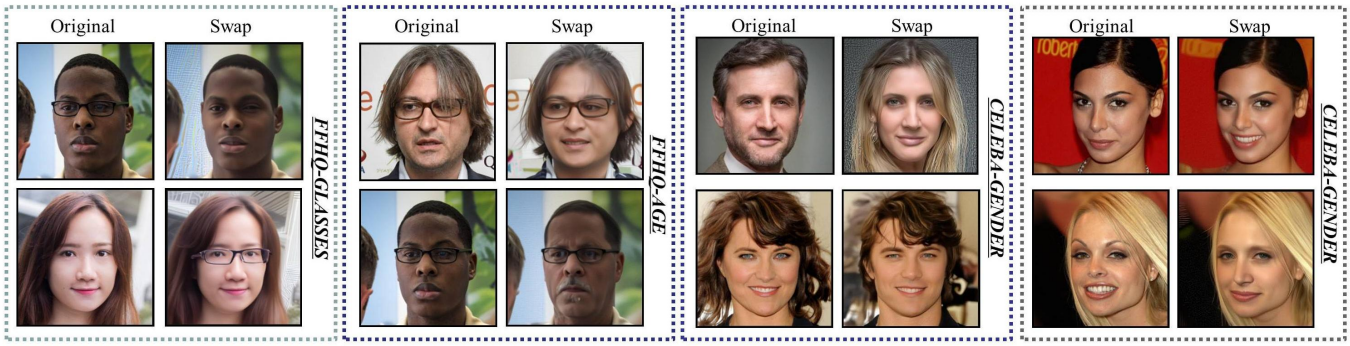

Figure 1: **Qualitative Results.** We propose an unsupervised method to extract both shared and domain-specific features from the latent space of diffusion models. Our approach achieves excellent performance in both global and local editing tasks, without relying on any semantic supervision such as text prompts, semantic masks, or other user-provided guidance. Furthermore, our method does not require any training or fine-tuning of the diffusion model, nor does it involve any labeled data.

In this work, inspired by the network design idea in literature [40], we propose a mapping network for separating Common and Salient factor based on the verified linear structure of the h-space. This architecture employs a lightweight and simple branch network to unsupervised decompose the bottleneck features of the diffusion U-Net into Common and Salient latent factors. During training, the network can automatically separate these two types of factors and independently control the corresponding regions. The qualitative results are illustrated in Fig.1.

To the best of our knowledge, this is the first work to systematically separate Common and Salient factor at the bottleneck feature level of diffusion models for image representation. Our main contributions are as follows:

1. We present a plug-and-play editing approach that enables content separation within images without modifying the original parameters of the Diffusion U-Net. As can be seen in Fig.2.

2. We propose a lightweight, unsupervised separation network based on h-space, which efficiently separates Common and Salient factors.

3. Extensive experiments demonstrate the superior performance of our method in terms of interpretability and controllability in image editing.

4. We release the code at GitHub.

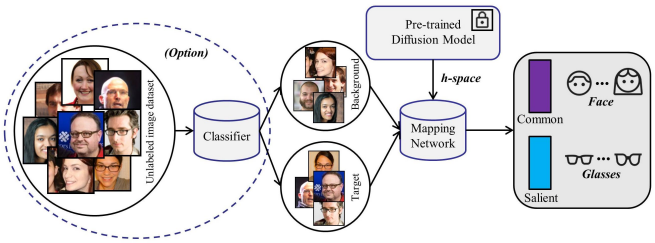

Figure 2: **Overview of the proposed application of Mapping Network**. Our proposed method is built upon the h-space[21] of a pretrained Diffusion Model and learns the common and salient information shared and unique across two classes of images in an unsupervised approach. Note: If the dataset has already been divided into two classes, then there is no need to use Classfier

2. Related Works

Latent Spaces in VAEs and GANs. Variational Autoencoders (VAEs) [19, 32] and Generative Adversarial Networks (GANs) [7, 16] have well-defined latent spaces, which naturally offer advantages in semantic disentanglement, controllable image editing, and interpretability (e.g., for VAE [2], GAN [39]). Numerous works have attempted to disentangle "common" and "distinctive" representations within such models, applying them to tasks such as domain adaptation [5, 12] and image-to-image translation [22, 14]. The key distinction between these works and ours lies in the objective. Our goal is to identify the latent representations of background categories (e.g., NoGlasses) and target categories (e.g., With-

Glasses), with the aim of generating new images, rather than merely translating them into another domain. Another important difference is that we aim to encode all distinctive variations of the target domain relative to the background domain, rather than isolating a single particular distinctive attribute (as in [27]). Moreover, in works such as [13], which extract and transfer distinctive attributes of the target domain to the background domain using GAN-based models—even with the aid of spatial conditions like masks to preserve context—results often suffer from issues like blurry details and structural distortions in the target regions. Overall, VAEs and GANs tend to be less capable than DDMs in preserving fine-grained detail fidelity.

Latent Space of DDMs. Denoising Diffusion Models (DDMs) [11] have recently become a major focus in image editing research due to their superior ability to reconstruct high-resolution textures and fine details. Asyrp et al. [21] were the first to verify that the deepest bottleneck layer of the diffusion U-Net (h-space) exhibits strong linearity and can be interpreted as a semantic latent space. Their approach adds CLIP-guided semantic directions Δh_t to the bottleneck features at each denoising timestep $(x_T, ..., x_1)$ to achieve controllable editing. However, in domains where semantics cannot be accurately described by natural language (e.g., BraTS), the generalization and efficiency of such supervised semantic directions are significantly limited. Subsequently, Jeong et al. [15] proposed fusing the bottleneck features of a content image into those of the original image using Slerp during reverse sampling. The resulting image retains most of the original content while integrating the semantic characteristics of the content image. This result further confirms the richness of semantic information encoded in h-space and its validity as a latent space.

Due to the previous belief that DDMs lacked a latent space, no prior work has focused on feature disentanglement at the h-space bottleneck level. Nevertheless, several studies have introduced unsupervised frameworks—such as Jacobian analysis [30, 9, 20], PCA [9], and contrastive learning [4]—to mine global or local semantic directions in h-space. For instance, the work in [30] computes local Jacobian directions for each image's bottleneck features, aligns the most similar directions across images, and averages them to obtain both local directions for individual samples and globally shared directions across samples. While these approaches have achieved semantic disentanglement and effective editing based on h-space, they generally require post-hoc manual interpretation. More importantly, their controllable granularity remains low, limiting the ability to learn or capture fine-grained object details. Another line of research exploring local semantic directions in h-space [20, 44] has shown promise in capturing fine details, but relies on user-provided masks, which restricts flexibility and scalability.

Semantic Image Editing Based on DDMs. In object-driven image editing, accurately compositing foreground objects into a background image typically requires multiple spatial conditions, such as semantic layouts [45, 29] and coarse bounding boxes [36, 35, 41, 28, 8]. However, these guiding inputs must be manually created and provided by the user, which can be a significant barrier in practice. A novel approach is DreamID [43], which achieves face (Salient region) swapping while preserving contextual pixels, without relying on spatial conditions like ControlNet or masks. However, in most real-world cases, it's difficult to find two real face images of the same identity, making it challenging to construct triplet ID groups as required in DreamID. In contrast, our method does not rely on external spatial conditions or identity pairing, making it more generalizable. It can replace target regions while preserving the original composition and fine details of the image.

On the other hand, some works perform local editing based solely on textual descriptions, significantly simplifying user input. Representative examples include Prompt-to-Prompt [10] and Plug-and-Play [38]. However, a closer inspection of their results reveals that they often fail to preserve contextual pixels around the edited object. This is because fine-grained visual attributes are difficult to express precisely with concise natural language, often leading to attribute entanglement. When such prompts are applied in the cross-attention layers,

attention leakage occurs. Subsequent research [42] mitigated this by designing loss functions and retraining Stable Diffusion to ensure that each noun in the description generates disentangled attention maps. While this resulted in better attention distribution, the authors also pointed out that the 16×16 resolution of the cross-attention map limits fine-grained control. Moreover, even with ideal cross-attention maps, it does not guarantee high-quality editing of fine image details.

3. Background

3.1 Denoising Diffusion Model

DDIM [34] and DDPM [11] are latent variable models that both learn data distributions by progressively estimating the noise ϵ in noisy images x_t during the reverse denoising process. In the forward diffusion process of DDPM($q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\,\mathbf{x}_{t-1}, \beta_t\mathbf{I})$) transforms the real data distribution $q(x_0)$ stepwise into an approximate standard normal distribution $\mathcal{N}(0, \mathbf{I})$, where $\{\beta_t\}_{t=1}^T$ is a predefined variance schedule and $\{\mathbf{x}_t\}_{t=1}^T$ are the noisy versions of x_0 at each step t. Conversely, the sampling formula for the reverse process is given in Equation 1.

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{1-\beta_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\,\epsilon_\theta(\mathbf{x}_t)\right) + \sigma_t \mathbf{z}_t \tag{1}$$

where $\epsilon_\theta^t(\mathbf{x}_t)$ is a denoising network primarily used to predict the noise ϵ. Overall, due to its reliance on a Markov chain, DDPM suffers from the drawback of requiring a large number of sampling steps, making the final image generation process slow.

Building on DDPM, DDIM redefines the forward diffusion as a non-Markovian process: $q_\sigma(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}\left(\sqrt{\alpha_{t-1}}\mathbf{x}_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\frac{\mathbf{x}_t - \sqrt{\alpha_t}\mathbf{x}_0}{\sqrt{1-\alpha_t}}, \sigma_t^2\mathbf{I}\right)$, and achieves significantly faster image generation by applying an implicit deterministic update during the reverse sampling process. The final sampling formula for the reverse process is shown in Equation 2.

$$\mathbf{x}_{t-1} = \sqrt{\alpha_{t-1}}\left(\frac{\mathbf{x}_t - \sqrt{1-\alpha_t}\,\epsilon_t^\theta(\mathbf{x}_t)}{\sqrt{\alpha_t}}\right) + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\epsilon_t^\theta(\mathbf{x}_t) + \sigma_t\mathbf{z}_t \tag{2}$$

When σ_t=0, the sampling process becomes deterministic, corresponding to DDIM sampling.

3.2 h-space

In Asyrp [21], after an exhaustive enumeration of semantic latent space choices, the authors found that the 8th layer (Not affected by Skip-connection) of the U-net is most suitable for semantic latent space(h-space). The detailed framework is illustrated in Figure 2.

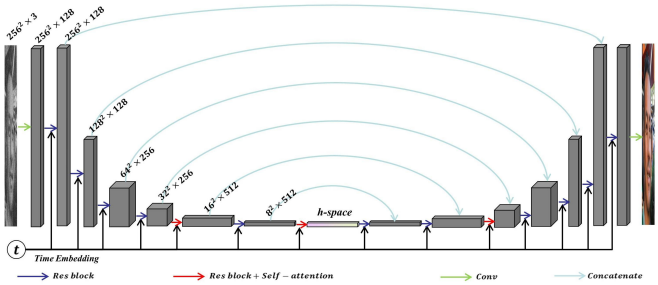

Figure 3: **h-space(8th layers) in U-Net based diffusion model architecture.**

The authors proposed a new controllable reverse process based on based on an abbreviated version of DDIM, with the corresponding formulation shown in Equation 3.

$$\mathbf{x}_{t-1} = \sqrt{\alpha_{t-1}}\underbrace{\mathbf{P}_t(\epsilon_t^\theta(\mathbf{x}_t))}_{\text{predicted } \mathbf{x}_0} + \underbrace{\mathbf{D}_t(\epsilon_t^\theta(\mathbf{x}_t))}_{\text{direction pointing to } \mathbf{x}_t} + \underbrace{\sigma_t\mathbf{z}_t}_{\text{random noise}} \tag{3}$$

As shown in Equation 3, it first applies deterministic DDIM forward diffusion to reverse the real image x_0 into a noise image x_T. Then, the noisy image x_T undergoes a reverse process, as described in Equation 4, to achieve image editing. Additionally. When performing local editing, our denoise process is quite simple (we do not adopt the quality enhancement steps proposed in Asyrp).

$$\tilde{\mathbf{x}}_{t-1} = \begin{cases} \sqrt{\alpha_{t-1}}\,\mathbf{P}_t(\epsilon_\theta(\tilde{\mathbf{x}}_t \mid \tilde{\mathbf{h}}_t)) + \mathbf{D}_t, & \text{if } T \geq t \geq t_{\text{edit}} \\ \sqrt{\alpha_{t-1}}\,\mathbf{P}_t(\epsilon_\theta(\tilde{\mathbf{x}}_t \mid \mathbf{h}_t)) + \mathbf{D}_t + \sigma_t^2\mathbf{z}_t, & \text{if } t_{\text{edit}} > t \end{cases} \tag{4}$$

Where t_edit denotes the editing interval, and $\tilde{h}_t$ represents the modified bottleneck feature at the deepest layer.

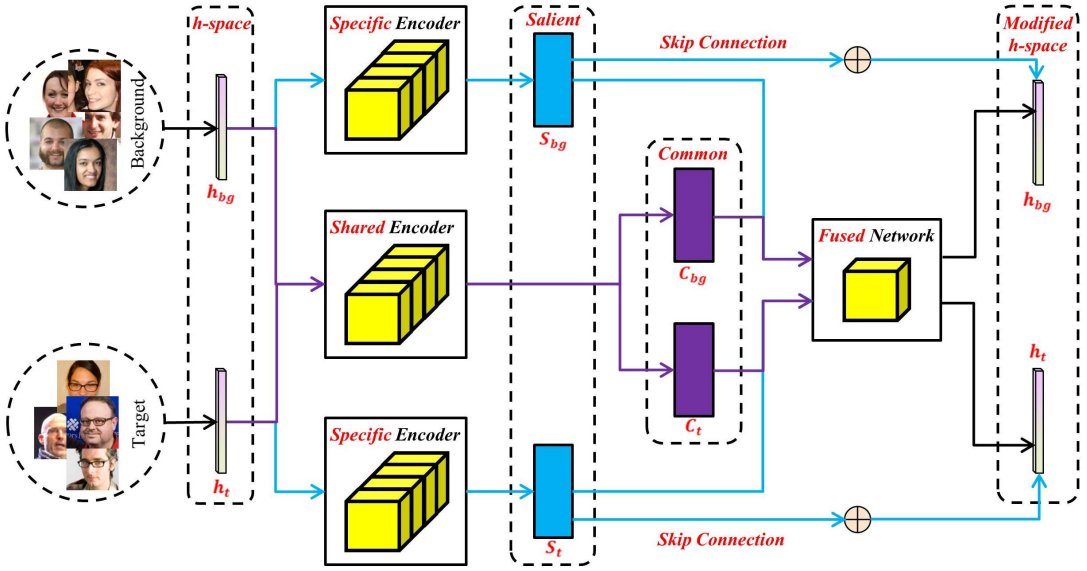

Figure 4: **Detailed process of the Mapping network acting on h-space.** For example, when we have a dataset that is divided into two categories based on a specific attribute, the reconstruction of Target image uses $\tilde{h}_t$=Fused(Common, Salient). In contrast, reconstructing Background image uses $\tilde{h}_t$=Fused(Common, 0).

4. Proposed Method

4.1 Model

Building upon the h-space, we propose a Mapping Network to decompose the bottleneck features at each editing timestep $(X_T, ..., X_t)$ into Common factors and Salient factors. We introduce three encoders with unshared parameters: a Shared Encoder that extracts Common features across different class of images, and two Specific Encoders, one dedicated to extracting the Salient factors from Target samples and the other from Background samples. Finally, a Fusion Network is employed to integrate the decomposed Common and Salient factors, replacing the original bottleneck features in h-space. Through this iterative reconstruction process, the mapping network is able to gradually extract the unique Salient factors associated with the Target category. The overall architecture is shown in Figure 3.

We assume that X=x_i and Y=y_j represent independent and identically distributed background images $P_{(X)}$ and target images $P_{(Y)}$ from the dataset, respectively. Both this two class of images are assumed to be generated from a pair of latent variables ($\mathbf{C} \in \mathbb{R}^L, \mathbf{S} \in \mathbb{R}^M$). Where C represents the generating factors shared between X and Y, while S describes only the Salient factors unique to Y.

In order to simultaneously ensure appearance consistency and factor disentanglement within the controllable diffusion model of Common and Salient factors, we design a composite objective function consisting of both image-space losses and latent-space losses. The overall objective is defined as:

$$\begin{aligned}\mathcal{L}_{\text{total}} &= \mathcal{L}_{\text{latent}} + \mathcal{L}_{\text{img}} \\ &= \lambda\mathcal{L}_{\text{Silence}}(S_{bg}) + \lambda\mathcal{L}_{\text{Consistency}}(h, f_\theta(C,S)) \\ &+ \lambda\mathcal{L}_{\text{KL}}(C_{bg}, C_t) + \lambda\mathcal{L}_{\text{Orthogonal}}(C,S) + \lambda\mathcal{L}_{\text{Class}}(C,S) \\ &+ \lambda\mathcal{L}_{\text{ID}}(x_{rec}, x_{real}) + \lambda\mathcal{L}_{\text{Pixel}}(x_{rec}, x_{real}) \\ &+ \lambda\mathcal{L}_{\text{LPIPS}}(x_{rec}, x_{real}) \end{aligned} \tag{5}$$

In the following, we will introduce each loss in detail

4.2 Latent Space Loss

Silence Regularization. To prevent background information from leaking into S, we enforce $x_i \sim P(x \mid \mathbf{z}_i, \mathbf{s}_i = 0)$ at each edit time step in the h-space when we reconstruct the Background class image. During the reconstruction of background images, only the Common factor is activated, which facilitates modeling C and S as a factorize distribution. The corresponding loss is:

$$\mathcal{L}_{\text{sbg}} = \|\mathbf{s}_{\text{bg}}\|_2^2 \tag{6}$$

Consistency Loss. In h-space, we require the final output of the Mapping Network, namely the modified h-space features, to match the ground-truth h-space features of real images. This constraint ensures that the combined latent factors C and S can fully reconstruct the original latent representation, preventing any drift or degradation in either component. The corresponding loss is:

$$\mathcal{L}_{\text{lat}} = \|f_\theta(\mathbf{c}_{\text{bg}}, s_{bg}) - \mathbf{w}_{\text{bg}}\|_2^2 + \|f_\theta(\mathbf{c}_t, \mathbf{s}_t) - \mathbf{w}_t\|_2^2 \quad (7)$$

Bi-directional KL Divergence. We apply a bidirectional KL divergence on the pooled Common logits over the temporal dimension, enforcing the statistical alignment of the C distributions between the two domains. This prevents Salient features from leaking into C, while allowing both domains to share the average data distribution. The corresponding loss is:

$$\mathcal{L}_{\text{DAO}} = \frac{1}{T}\sum_{i=1}^{T}\left[\text{KL}(p_{\text{bg}}^{(i)} \,\|\, p_t^{(i)}) + \text{KL}(p_t^{(i)} \,\|\, p_{\text{bg}}^{(i)})\right] \quad (8)$$

Orthogonality Constraint. To prevent confusion and redundancy between background and salient features, we explicitly enforce independence between the latent variables $C = (z_1, ..., z_L)$ and $S = (s_1, ..., s_M)$ in practice. This strengthens the separation between the Common factor and the Salient factor. The corresponding loss is:

$$\mathcal{L}_{\text{Orthogonal}} = \frac{1}{TBHW}\sum_{t,b,h,w}\left(\mathbf{c}_{tbhw}^\top \bar{\mathbf{s}}_{tbhw}\right)^2 \quad (9)$$

Classification Loss. To ensure that $P_{(X|c,s=0)}$ can reconstruct the Background class and $P_{(Y|c,s=s')}$ can reconstruct the Target class, we attach a classification head after each encoder to predict the class based on the extracted features. We assign label 0 to the C factors from both class X and class Y, and label 1 to the S factors from class Y, using binary cross-entropy loss ($\mathcal{B}$). The corresponding loss is:

$$\begin{aligned}\mathcal{L}_{\text{Class}} = \big[&\mathcal{B}(c_{\text{bg}}, 0) + \mathcal{B}(s_{\text{bg}}, 0) \\ &+ \mathcal{B}(c_t, 0) + \mathcal{B}(s_t, 1)\big]\end{aligned} \quad (10)$$

4.3 Image Space Loss

Here, we introduce Identity Loss, Pixel-wise Loss, and LPIPS Loss to jointly ensure high fidelity of the generated results in both visual and semantic aspects. The corresponding loss is:

$$\mathcal{L}_{\text{ID}}(x, \hat{x}) = 1 - \cos\left(f_{\text{id}}(x), f_{\text{id}}(\hat{x})\right) \quad (11)$$

$$\mathcal{L}_{\text{Pixel}}(x, \hat{x}) = \frac{1}{N}\|\hat{x} - x\|_2^2 \quad (12)$$

$$\mathcal{L}_{\text{LPIPS}} = \sum_l w_l \|\phi_l(x) - \phi_l(\hat{x})\|_2^2 \quad (13)$$

5. Experiments

In this section, we evaluate the capability of our method to extract salient and common factors on two face datasets (FFHQ [17] and CelebA-HQ [18]) from both qualitative and quantitative perspectives. Additionally, we report the SR (Swap Recognition) accuracy for both the Background and Target categories based on a pretrained classifier, using the swapped images. The SR is computed according to the following formula.

$$SR = \frac{1}{|D|}\sum_{x\in D}\mathbb{1}\big(C(M(x, y)) = y\big) \quad (14)$$

And Representative examples from the datasets are shown in Fig. 5.

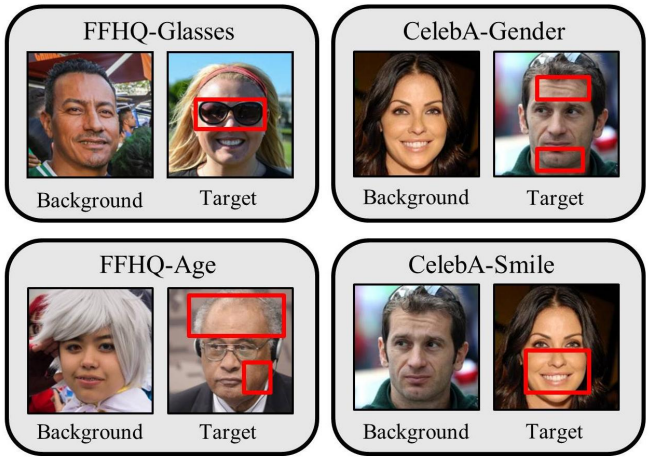

Figure 5: **Example of experimental datasets.** We selected the "Age" and "Glasses" attributes from the FFHQ dataset, and the "Smile" and "Gender" attributes from the CelebA-HQ dataset for subsequent experiments. Among these, Glasses and Smile correspond to local editing, while Age and Gender correspond to global editing.

5.1 FFHQ Dataset

In this dataset, we focus on the attributes "Glasses" and "Age". Taking the "Glasses" attribute as an example, we divide the dataset into two categories based on the presence of eyeglasses. One category, referred to as the Target class ($y = 1$), contains face images with eyeglasses, while the other, the Background class ($y = 0$), consists of face images without eyeglasses. However, despite filtering, the dataset remains excessively large. Therefore, we select 10,000 images with eyeglasses and 10,000 without as training data, and 3,500 images with eyeglasses and 3,500 without as testing data.

It is important to note that, unlike Cristiano et al. [31], we do not strictly control for the presence of other accessory-related attributes when constructing the subsets. For example, in their work, the Background class was curated to exclude not only the "Eyeglasses" attribute but also any other accessory-related attributes. In contrast, our categorization is based solely on the eyeglasses label, which may inadvertently include images with other accessories such as hats. This undoubtedly increases the difficulty of our task.

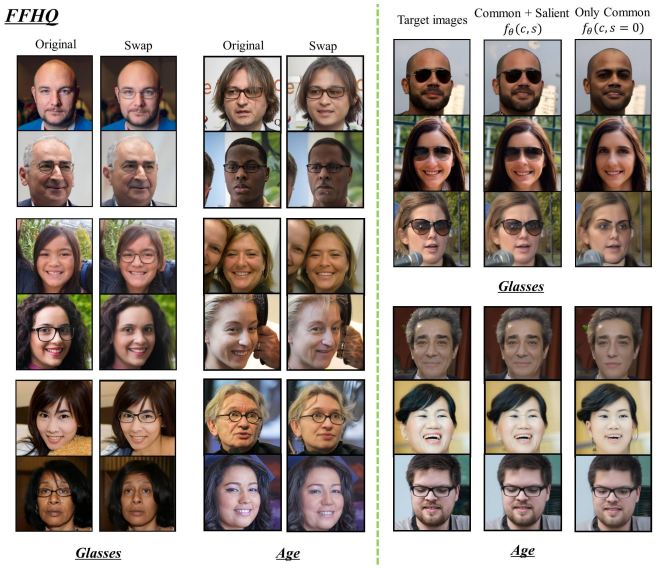

Figure 6: **Qualitative results of editing glasses and age attributes in the FFHQ dataset.** On the left side of the figure, we swap the salient representation s between two image categories while keeping their common representation c fixed. This leads to a successful exchange of the target attributes, indicating that the salient features have been effectively captured. On the right side, we use the mapping network to extract c and s in the h-space and reconstruct the images. The reconstructed results closely resemble the original inputs, demonstrating the fidelity of the representation. Furthermore, by setting $s = 0$, we effectively remove the target-specific attributes—such as glasses in FFHQ-GLASSES and aging-related features (e.g., wrinkles and facial hair) in FFHQ-AGE—while preserving the underlying identity and structure.

The quantitative metrics for the "Glasses" attribute in the FFHQ dataset are shown in the table below.

Table 1: **The image-level and classification-level quantitative metrics between the background (x) domain and the target (y) domain.*(FFHQ-Glasses)***

	↓LPIPS		↓MSE		↑Identity		↓FID		↑SR	
	x	y	x	y	x	y	x	y	x	y
Ours	0.21	0.23	0.01	0.01	0.74	0.72	31.42	42.14	0.36	0.54

The quantitative metrics for the "Age" attribute in the FFHQ dataset are presented in the table below.

Table 2: **The image-level and classification-level quantitative metrics between the background (x) domain and the target (y) domain.** ***(FFHQ-Age)***

	↓LPIPS		↓MSE		↑Identity		↓FID		↑SR	
	x	y	x	y	x	y	x	y	x	y
Ours	0.21	0.22	0.01	0.01	0.79	0.69	50.85	41.54	0.3	0.45

5.2 CelebA-HQ Dataset

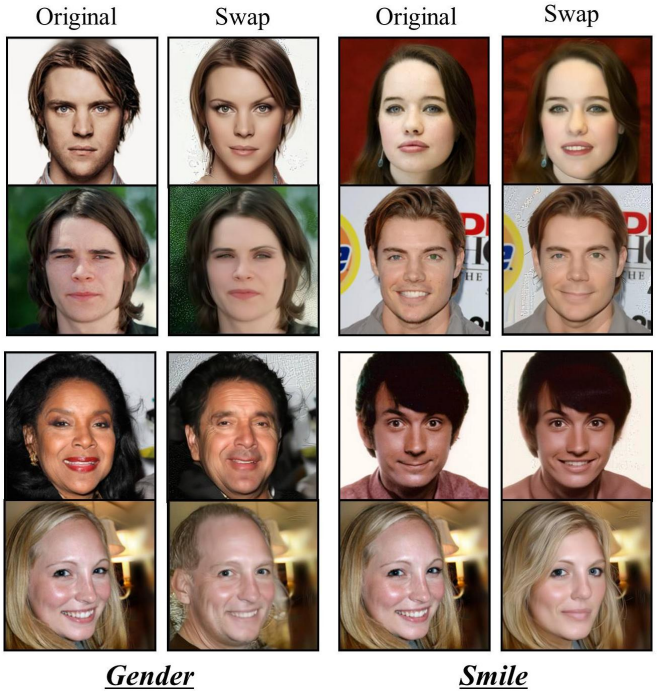

Figure 7: **Qualitative results of editing Smile and Gender attributes in the CelebA dataset.**

The quantitative metrics for the "Gender" attribute in the CelebA dataset are presented in the table below.

Table 3: **The image-level and classification-level quantitative metrics between the background (x) domain and the target (y) domain.** ***(CelebA-Gender)***

	↓LPIPS		↓MSE		↑Identity		↓FID		↑SR	
	x	y	x	y	x	y	x	y	x	y
Ours	0.21	0.21	0.01	0.01	0.86	0.74	45.86	31.23	0.22	0.48

In addition, the quantitative metrics for the "Smile" attribute in the CelebA dataset are presented in the table below.

Table 4: **The image-level and classification-level quantitative metrics between the background (x) domain and the target (y) domain.** ***(CelebA-Smile)***

	↓LPIPS		↓MSE		↑Identity		↓FID		↑SR	
	x	y	x	y	x	y	x	y	x	y
Ours	0.18	0.17	0.01	0.01	0.84	0.85	24.69	23.83	0.19	0.25

5.3 Ablation

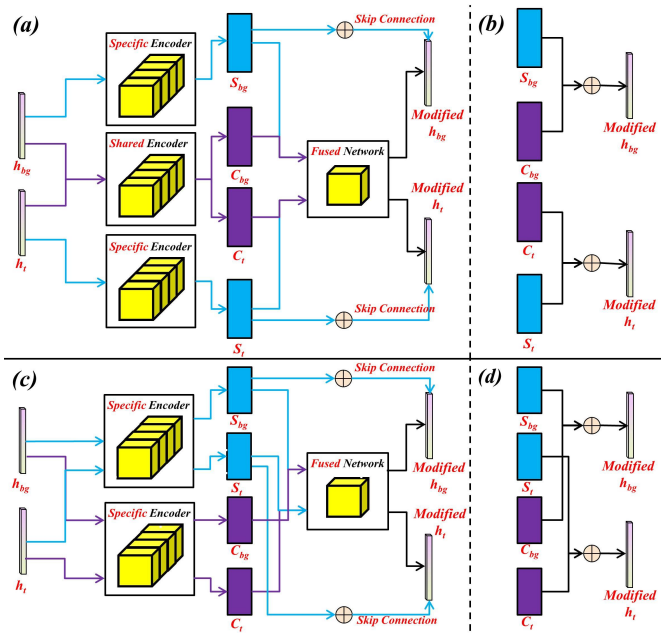

Figure 8: **Four different mapping network architectures.** In (b) and (d) it is Modified_h=C+S. and in (a) and (c) it is Modified_h=C+Fused(C,S).

Architecture and Depth of the Mapping Network. Here, we propose four different frameworks for extracting the C and S factors from the bottleneck features in the h-space, as illustrated in Fig.8. For time-step editing, we use $t_{edit} = 400$ and $t_{edit} = 0$ as the default editing steps. Under a consistent setting of learning rate = 0.0001 and denoising steps = 50, we conduct ablation studies on the number of stacked layers in the Mapping Network across the four different architectures. This allows us to investigate how the editing time step and the network complexity of the Mapping Network influence the disentanglement and representation ability of the C/S factors. The corresponding evaluation metrics in four different attribute datasets are reported in the below.

5.3.1 FFHQ-Glasses

Table 5: t_edit=[400, T]

Arc \ Layers	8					12				
	LPIPS	MSE	ID	FID	SR	LPIPS	MSE	ID	FID	SR
(a)	0.233	0.005	0.733	39.743	0.42	0.252	0.006	0.677	42.001	0.49
(b)	0.209	0.004	0.768	34.976	0.32	0.234	0.005	0.721	39.256	0.38
(c)	0.184	0.003	0.828	33.163	0.31	0.214	0.005	0.763	34.158	0.34
(d)	0.190	0.004	0.807	32.074	0.25	0.233	0.005	0.736	40.067	0.30

Table 6: t_edit=[0, T]

Arc \ Layers	8					12				
	LPIPS	MSE	ID	FID	SR	LPIPS	MSE	ID	FID	SR
(a)	0.233	0.005	0.727	36.779	0.45	0.273	0.006	0.614	43.098	0.62
(b)	0.225	0.005	0.740	35.093	0.32	0.236	0.005	0.694	33.731	0.42
(c)	0.169	0.003	0.850	29.378	0.3	0.196	0.004	0.780	32.661	0.42
(d)	0.188	0.004	0.813	30.73	0.3	0.215	0.004	0.760	33.450	0.31

By comparing the results of the quantitative metrics mentioned above, we observe that architecture (a) achieves a higher editing success rate than architecture (c). This indicates that, for Subsequet editing tasks, it is more effective to use a Shared Encoder to extract the common features from images of different classes, and to design two separate Specific Encoders to capture the class-specific features for each category. This supports the conclusion that a Shared/Specific architecture is more conducive to the disentanglement and extraction of C/S factors.

Furthermore, by comparing architectures (a) and (b), or (c) and (d), we find that the introduction of the "Fused" network significantly improves the editing success rate at the cost of the image quality of the swapped images. This also suggests that the "Fused" network contributes positively to the disentanglement of C/S factors.

5.3.2 FFHQ-Age

Here, we focus specifically on comparing the effects of different Editing Timesteps and network depths (number of stacked layers) on image editing performance under framework (a) across various datasets.

Table 7: t_edit=[400, T]

Arc \ Layers	8					12				
	LPIPS	MSE	ID	FID	SR	LPIPS	MSE	ID	FID	SR
(a)	0.208	0.004	0.761	46.49	0.28	0.244	0.006	0.689	48.63	0.34

Table 8: t_edit=[0, T]

Arc \ Layers	8					12				
	LPIPS	MSE	ID	FID	SR	LPIPS	MSE	ID	FID	SR
(a)	0.214	0.005	0.740	46.19	0.38	0.255	0.006	0.630	50.07	0.49

For the *Age* attribute, we observe that although the image quality metrics under $t_{\text{edit}} = [400, T]$ are generally better than those under $t_{\text{edit}} = [0, T]$, the SR scores in the $t_{\text{edit}} = [400, T]$ setting remain significantly low. This may be attributed to an insufficient number of editing steps, resulting in an incomplete swapping of the s factor, and thus the intended attribute modification is not adequately reflected. On the other hand, the $t_{\text{edit}} = [0, T]$ setting follows the same previously observed pattern—trading off image quality in exchange for higher editing success rates. Therefore, a trade-off between visual fidelity and editing effectiveness must be carefully considered. As a result, we select the configuration with $t_{\text{edit}} = [0, T]$ and a Mapping Network with 8 layers as the optimal editing setup.

5.3.3 CelebA-Gender

Table 9: t_edit=[400, T]

Arc \ Layers	8					12				
	LPIPS	MSE	ID	FID	SR	LPIPS	MSE	ID	FID	SR
(a)	0.213	0.004	0.806	50.16	0.31	0.262	0.006	0.711	59.42	0.45

Table 10: t_edit=[0, T]

Arc \ Layers	8					12				
	LPIPS	MSE	ID	FID	SR	LPIPS	MSE	ID	FID	SR
(a)	0.213	0.004	0.796	38.55	0.35	0.266	0.006	0.658	53.11	0.49

5.3.4 CelebA-Smile

Table 11: t_edit=[400, T]

Arc \ Layers	8					12				
	LPIPS	MSE	ID	FID	SR	LPIPS	MSE	ID	FID	SR
(a)	0.160	0.003	0.879	23.58	0.16	0.175	0.003	0.843	24.26	0.22

Table 12: t_edit=[0, T]

Arc \ Layers	8					12				
	LPIPS	MSE	ID	FID	SR	LPIPS	MSE	ID	FID	SR
(a)	0.155	0.002	0.898	40.26	0.15	0.192	0.004	0.805	41.23	0.29

Based on the above analysis, we find that the configuration with $t_{\text{edit}} = [0, T]$ and a Mapping Network depth of 8 layers yields the best performance in terms of both image editing and reconstruction. It achieves effective editing while maintaining a reasonable level of image quality.

6. Discussion and Perspectives

In this work, we successfully designed a novel idea of mapping network that unsupervisedly disentangles all common and salient factors of a target dataset relative to a background dataset. Our experiments demonstrated that the proposed model effectively separates and captures salient attributes (e.g., glasses in FFHQ) as well as common attributes (e.g., gender in FFHQ[17]). The model exhibits excellent performance on the facial FFHQ and CelebA-HQ dataset. We believe that factor disentanglement in the h-space[21] of diffusion models is an important research direction with great potential for interdisciplinary applications. Moreover, diffusion models naturally generate high quality images with superior textures and resolution compared to VAEs[19, 32] and GANs[7, 16].

Below are three important observations from our experiments:

1. As shown in (a) of Fig.9, our method relies on h-space to separate common and salient factors. Therefore, if the bottleneck features in h-space cannot faithfully reconstruct the original image (in terms of structure and detail), the effectiveness of our approach will be compromised.

2. As illustrated in (b) of Fig.9, we found that the number of denoising steps in h-space is crucial, as it directly affects texture and detail reconstruction. Insufficient steps may lead to the loss of fine features. We found that setting the inverse denoising steps to around **Inv_step = 100** is sufficient for high-fidelity reconstructions that closely resemble the original images.

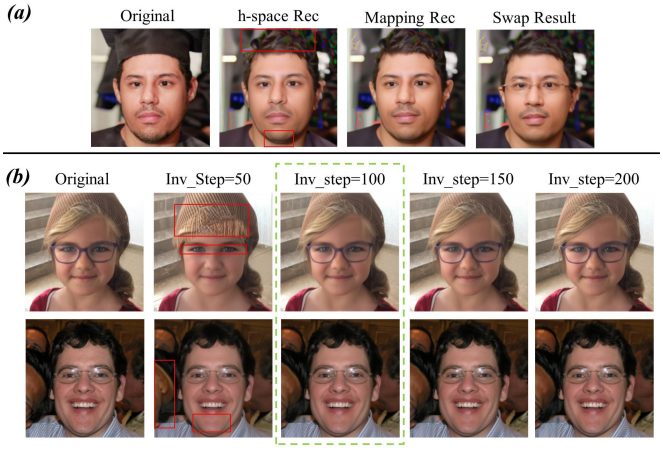

Figure 9: **Some observations and hyperparameter selection regarding the h-space.** As shown in Figure (a), when the denoising step is set to Inv_step = 50, not only are image details lost, but even some important attributes are missing. Since our Mapping Network performs reconstruction and swap operations based on the h-space, if the bottleneck features in h-space lose detail or attribute information, our reconstruction and swap results will also be degraded accordingly. Figure (b) presents the reconstruction results under different denoising steps. We found that setting **Inv_step = 100** is the most appropriate, as it allows the reconstructed image to almost perfectly match the original.

3. At the beginning of training, our strategy was to preserve global information across time steps by applying global supervision on the prediction errors of the same sample across all time steps. However, we found that this approach did not yield good results. This may be because it averages signals from different time steps, leading the network to minimize the overall loss by encoding information like glasses into more stable global channels (such as Common salients), which causes Salient factors to barely encode glasses anymore and results in a failure of the swapping mechanism. Eventually, similar to Asyrp, we adopted the strategy of updating the mapping network at each time step, which led to improvements in both qualitative and quantitative metrics.

As future work, we plan to extend our framework to handle multi-class images, where salient factors from different classes can be disentangled and exchanged across samples.

References

[1] Florence Carton, Robin Louiset, and Pietro Gori. "Double infogan for contrastive analysis". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2024, pp. 172–180.

[2] Ricky TQ Chen et al. "Isolating sources of disentanglement in variational autoencoders". In: *Advances in neural information processing systems* 31 (2018).

[3] Siyi Chen et al. "Exploring low-dimensional subspaces in diffusion models for controllable image editing". In: *arXiv preprint arXiv:2409.02374* (2024).

[4] Yusuf Dalva and Pinar Yanardag. "Noiseclr: A contrastive learning approach for unsupervised discovery of interpretable directions in diffusion models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 24209–24218.

[5] Behnam Gholami et al. "Unsupervised multi-target domain adaptation: An information theoretic approach". In: *IEEE Transactions on Image Processing* 29 (2020), pp. 3993–4002.

[6] Peyman Gholami and Robert Xiao. "Streamlining image editing with layered diffusion brushes". In: *arXiv preprint arXiv:2405.00313* (2024).

[7] Ian J Goodfellow et al. "Generative adversarial nets". In: *Advances in neural information processing systems* 27 (2014).

[8] Jing Gu et al. "Swapanything: Enabling arbitrary object swapping in personalized image editing". In: *European Conference on Computer Vision*. Springer. 2024, pp. 402–418.

[9] René Haas et al. "Discovering interpretable directions in the semantic latent space of diffusion models". In: *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE. 2024, pp. 1–9.

[10] Amir Hertz et al. "Prompt-to-prompt image editing with cross attention control". In: *arXiv preprint arXiv:2208.01626* (2022).

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models". In: *Advances in neural information processing systems* 33 (2020), pp. 6840–6851.

[12] Judy Hoffman et al. "Cycada: Cycle-consistent adversarial domain adaptation". In: *International conference on machine learning*. Pmlr. 2018, pp. 1989–1998.

[13] Bingwen Hu et al. "Unsupervised eyeglasses removal in the wild". In: *IEEE transactions on cybernetics* 51.9 (2020), pp. 4373–4385.

[14] Xun Huang et al. "Multimodal unsupervised image-to-image translation". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 172–189.

[15] Jaeseok Jeong, Mingi Kwon, and Youngjung Uh. "Training-free content injection using h-space in diffusion models". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2024, pp. 5151–5161.

[16] 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*. 2019, pp. 4401–4410.

[17] 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*. 2019, pp. 4401–4410.

[18] Tero Karras et al. "Progressive growing of gans for improved quality, stability, and variation". In: *arXiv preprint arXiv:1710.10196* (2017).

[19] Diederik P Kingma, Max Welling, et al. *Auto-encoding variational bayes*. 2013.

[20] Theodoros Kouzelis et al. "Enabling Local Editing in Diffusion Models by Joint and Individual Component Analysis". In: *arXiv preprint arXiv:2408.16845* (2024).

[21] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. "Diffusion models already have a semantic latent space". In: *arXiv preprint arXiv:2210.10960* (2022).

[22] Hsin-Ying Lee et al. "Diverse image-to-image translation via disentangled representations". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 35–51.

[23] Hang Li et al. "Self-discovering interpretable diffusion latent directions for responsible text-to-image generation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 12006–12016.

[24] Nan Liu et al. "Compositional visual generation with composable diffusion models". In: *European Conference on Computer Vision*. Springer. 2022, pp. 423–439.

[25] Robin Louiset et al. "Separating common from salient patterns with Contrastive Representation Learning". In: *arXiv preprint arXiv:2402.11928* (2024).

[26] Robin Louiset et al. "SepVAE: a contrastive VAE to separate pathological patterns from healthy ones". In: *arXiv preprint arXiv:2307.06206* (2023).

[27] Liqian Ma et al. "Exemplar guided unsupervised image-to-image translation with semantic consistency". In: *arXiv preprint arXiv:1805.11145* (2018).

[28] Yingmao Miao et al. "Shining yourself: High-fidelity ornaments virtual try-on with diffusion model". In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025, pp. 359–368.

[29] Seok-Hwan Oh et al. "Key-point Guided Deformable Image Manipulation Using Diffusion Model". In: *arXiv preprint arXiv:2401.08178* (2024).

[30] Yong-Hyun Park et al. "Unsupervised discovery of semantic latent directions in diffusion models". In: *arXiv preprint arXiv:2302.12469* (2023).

[31] Cristiano Patrício et al. "Unsupervised Contrastive Analysis for Salient Pattern Detection using Conditional Diffusion Models". In: *arXiv preprint arXiv:2406.00772* (2024).

[32] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. "Stochastic backpropagation and approximate inference in deep generative models". In: *International conference on machine learning*. PMLR. 2014, pp. 1278–1286.

[33] Robin Rombach et al. "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 10684–10695.

[34] Jiaming Song, Chenlin Meng, and Stefano Ermon. "Denoising diffusion implicit models". In: *arXiv preprint arXiv:2010.02502* (2020).

[35] Yizhi Song et al. "Imprint: Generative object compositing by learning identity-preserving representation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 8048–8058.

[36] Yizhi Song et al. "Objectstitch: Object compositing with diffusion model". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 18310–18319.

[37] Jocelin Su et al. "Compositional image decomposition with diffusion models". In: *arXiv preprint arXiv:2406.19298* (2024).

[38] Narek Tumanyan et al. "Plug-and-play diffusion features for text-driven image-to-image translation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 1921–1930.

[39] Andrey Voynov and Artem Babenko. "Unsupervised discovery of interpretable directions in the gan latent space". In: *International conference on machine learning*. PMLR. 2020, pp. 9786–9796.

[40] Hu Wang et al. "Multi-modal learning with missing modality via shared-specific feature modelling". In: *Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition. 2023, pp. 15878–15887.

[41] Binxin Yang et al. "Paint by example: Exemplar-based image editing with diffusion models". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 18381–18391.

[42] Fei Yang et al. "Dynamic prompt learning: Addressing cross-attention leakage for text-based image editing". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 26291–26303.

[43] Fulong Ye et al. "DreamID: High-Fidelity and Fast diffusion-based Face Swapping via Triplet ID Group Learning". In: *arXiv preprint arXiv:2504.14509* (2025).

[44] Zhenbo Yu et al. "TtfDiffusion: Training-free and text-free image editing in diffusion models with structural and semantic disentanglement". In: *Neurocomputing* 619 (2025), p. 129159.

[45] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2023, pp. 3836–3847.

[46] Yang Zhang et al. "Enhancing semantic fidelity in text-to-image synthesis: Attention regulation in diffusion models". In: *European Conference on Computer Vision*. Springer. 2024, pp. 70–86.

[47] Zijian Zhang et al. "Unsupervised discovery of interpretable directions in h-space of pre-trained diffusion models". In: *arXiv preprint arXiv:2310.09912* (2023).