Enabling Local Editing in Diffusion Models by Joint and Individual Component Analysis

Theodoros Kouzelis[1,2]
theodoros.kouzelis@athenarc.gr
Manos Plitsis[3,4]
manos.plitsis@athenarc.gr
Mihalis A. Nikolaou[5]
m.nicolaou@cyi.ac.cy
Yannis Panagakis[2,4]
yannisp@di.uoa.gr

[1] National Technical University of Athens, Athens,Greece
[2] Archimedes AI, Athena RC, Athens, Greece
[3] Institute for Language and Speech, Athena RC Processing, Athens,Greece
[4] Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Athens,Greece
[5] Computation-based Science and Technology Research Center, The Cyprus Institute, Nicosia, Cyprus

Abstract

Recent advances in Diffusion Models (DMs) have led to significant progress in visual synthesis and editing tasks, establishing them as a strong competitor to Generative Adversarial Networks (GANs). However, the latent space of DMs is not as well understood as that of GANs. Recent research has focused on unsupervised semantic discovery in the latent space of DMs by leveraging the bottleneck layer of the denoising network, which has been shown to exhibit properties of a semantic latent space. However, these approaches are limited to discovering global attributes. In this paper we address the challenge of local image manipulation in DMs and introduce an unsupervised method to factorize the latent semantics learned by the denoising network of pre-trained DMs. Given an arbitrary image and defined regions of interest, we utilize the Jacobian of the denoising network to establish a relation between the regions of interest and their corresponding subspaces in the latent space. Furthermore, we disentangle the *joint* and *individual* components of these subspaces to identify latent directions that enable *local* image manipulation. Once discovered, these directions can be applied to different images to produce semantically consistent edits, making our method suitable for practical applications. Experimental results on various datasets demonstrate that our method can produce semantic edits that are more localized and have better fidelity compared to the state-of-the-art. https://zelaki.github.io/localdiff/

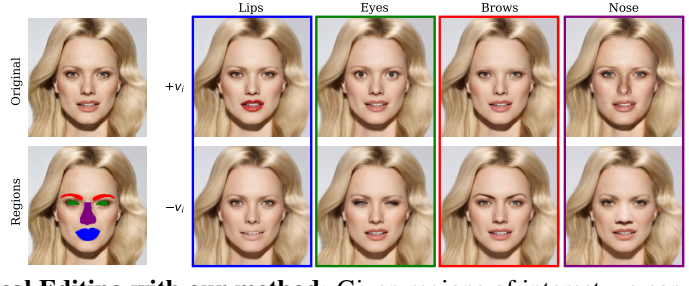

Figure 1: **Local Editing with our method**: Given regions of interest we can identify latent directions that result in diverse semantic edits without affecting the rest of the image. Linear interpolation within the identified semantic directions leads to gradual changes in the generated image like opening and closing the eyes.

1 Introduction

Diffusion models [11] have emerged as the new state-of-the-art paradigm of deep generative models. They have surpassed the long-standing dominance of Generative Adversarial Networks (GANs) [7] in image synthesis [8] and they have also shown strong potential in a variety of computer vision tasks, such as text-guided image synthesis [20, 25, 26, 28], image editing [9, 10, 13, 36] and inverse problems [19, 38].

However, while GANs generate images by sampling from a semantically informed latent space [7], that can then be used to guide the generative process and obtain editing capabilities [9, 30], in DMs such a semantic latent space is harder to identify, mostly due to the iterative nature of the diffusion process. Some approaches [1, 6, 14, 16] edit the latent variables (i.e., intermediate noisy images) guiding the generative process to a desired output, but require training a classifier [1, 6, 16] or fine-tuning the whole model for each new attribute [14]. Recently, Kwon et al. [15] discovered that the bottleneck layer of the denoising U-net [27] (coined $\mathcal{H}$-space) possesses the properties of a semantic latent space. Building upon this finding, recent works have attempted to discover "interpretable directions" in the generator's latent space [8, 22]. Once discovered, these latent representations of high-level concepts can be utilized to bring about predictable changes to the images. However, the latent directions discovered by previous works [8, 15, 22] result in global image manipulation without the ability of fine-grained regional control.

In this paper, we develop, to the best of our knowledge, the first unsupervised approach for local image editing in unconditional DMs. Given a pre-trained DM, a real image, and a set of regions of interest, such as the eyes or the lips in a face image (see Figure 1), our goal is to discover interpretable directions that specifically alter the selected image regions. Firstly, we utilize the row space of the Jacobian constrained to each specified region, which is obtained by its Singular Value Decomposition (SVD). The row space is spanned by directions that manipulate the attributes in each region of interest. However,this approach lacks an explicit constraint ensuring localized edits, and hence other regions are inadvertently affected. To alleviate this and achieve local manipulation, we further propose decomposing the Jacobian associated with each region of interest into two distinct components: a *joint* and an *individual* component. The row space of the joint component comprises latent directions that induce global changes across the entire image. In contrast, the row space of the individual component, which is orthogonal to the joint, is spanned by latent directions that specifically target a designated region of interest without influencing other regions. To obtain this decomposition we utilize the so-called *Joint and Individual Variation Explained (JIVE)* [18] method which is an iterative algorithm that estimates the joint and individual components in an arbitrary number of matrices. We further observe that the directions discovered from local regions of one image are readily applicable to other images, producing the same semantic manipulation, thus alleviating the need to recompute the decomposition for every sample.

Our main contributions can be summarized as follows:

- We propose the first method that identifies semantic directions in the latent space of unconditional DMs that are localized to specific image regions, thus enabling local editing.
- Local editing is achieved in an *unsupervised manner* by decomposing the set of Jacobians that correspond to different image regions in joint and individual components capturing global and local variation respectively.
- We demonstrate that the semantic directions discovered by our method generalize from one image to others making them ideal for plug-and-play applications.

- We show both qualitatively and quantitatively the superiority of our approach for local image editing against existing alternatives, even supervised ones.

2 Related Work

Diffusion Models Diffusion Models [11, 31] continue to push forward the state-of-the-art for image synthesis through architectural advances such as Latent Diffusion [26] and speeding up the generation process [32, 35]. Song et al. [34] have integrated DMs and score-based models [33] under an SDE formulation, improving our understanding of DMs as a reverse diffusion process. Classifier guidance [6] and its variants [1, 20, 29] control the generation process by guiding it toward a specific class. In [23] an additional encoder is introduced to capture semantic variation and control the generation process. However such approaches trade controllability with additional inference and training costs respectively. Instead, Kwon et al. [15] showed that the bottleneck layer of of-the-self DMs can be utilized to guide the generative process, exhibiting properties of a semantic latent space.

Interpretable Latent Directions in Generative Models Following the success of deep generative models in generating realistic and diverse images there has been a surge of interest in understanding the structure of their latent space. Many works [2, 8, 9, 15, 22, 39, 40] aim to identify latent subspaces that capture meaningful semantic variation in the generated images. Most notably for GANs, [9] find such subspaces by applying Principal Component Analysis (PCA) to the intermediate generator's representations. In [2, 8] semantic latent subspaces are found by the SVD of the Jacobian matrix. Most related to our work, [39, 40] relate a latent subspace with a specific image region by leveraging the gradient of the GAN generator, while [21] operate directly on the feature maps and jointly discover factors representing spatial parts and their appearances.

In DMs, following the work of Kwon et al. [15] recent works aim to discover interpretable directions in the latent space. In [8], they leverage the Jacobian of the generator to identify a semantic subspace in $\mathcal{H}$ without any supervision and [22] utilize the linearity of $\mathcal{H}$ to pull-back the metric tensor from $\mathcal{H}$ to the image space, establishing a semantic subspace. However, the latent directions detected by these methods tend to control global image attributes whereas we disentangle the directions responsible for global and local edits.

3 Preliminary

3.1 Diffusion Models and $\mathcal{H}$-Space

Diffusion Models are a class of generative models where generation is modeled as a denoising process. A forward diffusion process adds increasing amounts of Gaussian noise to an image $\mathbf{x_0}$ in T steps, and a learned reverse process gradually removes the noise. The forward process is defined as:

$$\mathbf{x}_t = \sqrt{a_t}\mathbf{x_0} + \sqrt{1-a_t}\mathbf{e}, \quad \mathbf{e} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{1}$$

where a_t defines the noise schedule. DDIM [32] redefines (1) as a non-Markovian process and the approximate reverse process becomes:

$$\mathbf{x}_{t-1} = \sqrt{a_{t-1}} \underbrace{\left(\frac{\mathbf{x}_t - \sqrt{1-a_t}\,\mathbf{e}_t^\theta(\mathbf{x}_t)}{\sqrt{a_t}} \right)}_{\mathbf{P}_t} + \underbrace{\sqrt{1-a_t-\sigma_t^2}}_{\mathbf{D}_t}\,\mathbf{e}_t^\theta(\mathbf{x}_t) + \sigma_t \mathbf{z}_t \tag{2}$$

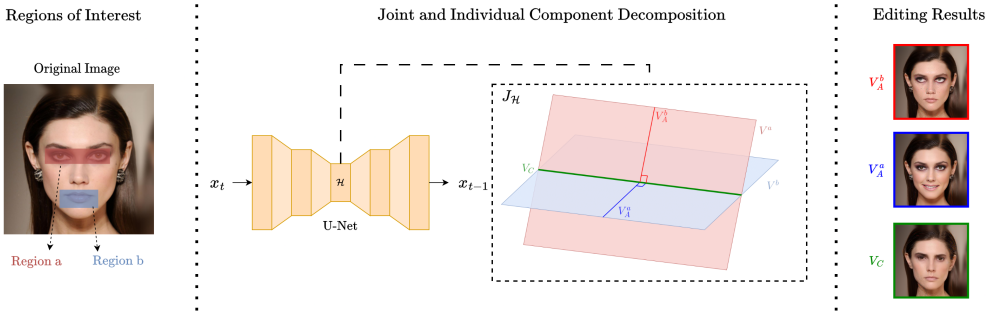

Figure 2: **An overview of our method.** Left: The regions of interest are selected. In this example, region a and region b correspond to the eyes and the mouth respectively. Center: The row space of the Jacobian of each region $\mathbf{V}^a$ and $\mathbf{V}^b$ is decomposed to the joint subspace $\mathbf{V}_C$ and the individual subspaces $\mathbf{V}_A^a$, $\mathbf{V}_A^b$. Right: Editing in $\mathcal{H}$ with directions from the joint subspace results in global edits, whereas editing with directions from the individual subspaces results in localized edits.

where $\mathbf{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\sigma_t = \eta\sqrt{(1-a_{t-1})/(1-a_t)}\sqrt{1-a_t/a_{t-1}}$. When $\eta = 0$, the process becomes deterministic and guarantees nearly perfect inversion. Kwon et al. [13] observed that the bottleneck feature maps of the denoising U-Net exhibit the properties of a semantic latent space. Given a pre-trained denoising network $\mathbf{e}^\theta(\cdot)$ they show that a semantic latent direction $\mathbf{v} \in \mathcal{H}$ that modifies the latent code $\mathbf{h_t}$ for every timestep of the denoising process can cause a desirable semantic change in the output image. Thus the denoising process of Eq. 2 becomes:

$$\mathbf{x}_{t-1} = \sqrt{a_{t-1}}\,\mathbf{P}_t\left(\mathbf{e}_t^\theta(\mathbf{x}_t|\mathbf{h}_t + \alpha\mathbf{v})\right) + \mathbf{D}_t\left(\mathbf{e}_t^\theta(\mathbf{x}_t|\mathbf{h}_t + \alpha\mathbf{v})\right) \tag{3}$$

where $\mathbf{e}_t^\theta(\mathbf{x}_t|\mathbf{h}_t + \alpha\mathbf{v})$ denotes adding $\alpha\mathbf{v}$ to the feature maps $\mathbf{h}_t$, α indicates the editing strength and $\mathbf{v}$ is assumed to be a unit vector i.e. $\mathbf{v}^T\mathbf{v} = 1$. In this work, once we have discovered a latent direction $\mathbf{v}$ as presented in Section 4 we use the editing process in Eq. 3 to edit a region of interest.

4 Methodology

In this section, we describe our method in detail. In Section 4.1 we describe how the SVD of the Jacobian can identify a subspace spanned by directions that control the principal modes of variation in a region of interest. In Section 4.2, given M regions of interest, we proceed to decompose the Jacobian of each region to a *joint* and *individual* component to achieve localized edits.

4.1 Jacobian Decomposition

Let $\{\mathbf{e}_t^\theta(\mathbf{x}_t|\mathbf{h}_t)\}_m \in \mathbb{R}^{d_m}$ be the output of the denoising network in a specified region $\mathbf{m}$ where d_m is the number of pixels in the region and the bottleneck $\mathbf{h}_t \in \mathbb{R}^{d_h}$ is of dimension d_h. Then the derivative of the output region w.r.t. $\mathbf{h}_t$ at timestep t is given by the Jacobian matrix $\{\mathbf{J}_h^t\}_m = \frac{\partial\{\mathbf{e}_t^\theta(\mathbf{x}_t|\mathbf{h}_t)\}_m}{\partial\mathbf{h}_t} \in \mathbb{R}^{d_m \times d_h}$. This is a matrix whose rows are the derivatives of each pixel value within region $\mathbf{m}$ w.r.t. $\mathbf{h}_t$ and d_m is the number of pixels in the region. For presentation

purposes, we will refer to the Jacobian of an output region $\mathbf{m}$ as $\mathbf{J}^{(m)}$ for the rest of the paper. Given an arbitrary vector $\mathbf{v} \in \mathbb{R}^{d_h}$, the directional derivative:

$$\lim_{\varepsilon \to 0} \frac{\{\mathbf{e}_t^\theta(\mathbf{x}_t|\mathbf{h}_t + \varepsilon \mathbf{v})\}_m - \{\mathbf{e}_t^\theta(\mathbf{x}_t|\mathbf{h}_t)\}_m}{\varepsilon} = \mathbf{J}^{(m)}\mathbf{v} \tag{4}$$

measures the instantaneous change in $\{\mathbf{e}_t^\theta\}_m$ resulting from a perturbation of $\mathbf{h}_t$ along the direction of $\mathbf{v}$. The unit-norm perturbation of $\mathbf{h}_t$ that maximizes the magnitude of this change is $\mathbf{v}_1 := \underset{\mathbf{v}}{\operatorname{argmax}} \|\mathbf{J}^{(m)}\mathbf{v}\|$. This is the first right singular vector of $\mathbf{J}^{(m)}$. Hence, a perturbation of $\mathbf{h}_t$ along $\mathbf{v}_1$ maximizes the magnitude of the instantaneous change in the output noisy image at timestep t. By maximizing $\|\mathbf{J}^{(m)}\mathbf{v}\|$ while remaining orthogonal to $\mathbf{v}_1$, one can derive the second right singular vector $\mathbf{v}_2$. By continuing this process we obtain r directions in $\mathcal{H}$-space that maximize the variability of the noisy image at time t. Thus the right singular vectors $\mathbf{V}^{(m)}$ from the SVD of the Jacobian i.e. $\mathbf{J}^{(m)} = \mathbf{U}^{(m)}\mathbf{S}^{(m)}\mathbf{V}^{(m)^T}$ span a subspace i.e. the row space of the Jacobian, that captures the principal modes of variation in the region of interest.

In practice it is highly inefficient to estimate the Jacobian of the denoising U-net directly, thus we rely on the *subspace iteration* method [8] to approximate the SVD of the $\mathbf{J}^{(m)}$ without ever storing it to memory. For a detailed description of the algorithm, please refer to [8] and the Appendix.

4.2 Joint and Individual Components in the Latent Space of DMs

The method described above lacks an explicit constraint that ensures localized edits. Note that a latent vector that maximizes the variability in a specified region can inadvertently affect other regions. However, for local editing, we would like to manipulate a specified region while not affecting the rest of the image. Our idea is to *disentagle* the *joint* and *individual* components of the Jacobian of each region. In this manner editing within the row space of the joint component results in global edits whereas directions from the row space of the individual component result in local edits.

Formally, given a real image $\mathbf{I}$, a set of N image regions that segment the image into parts i.e. $M = [\, \mathbf{m}_i \mid i \in (1, \cdots, N),\ \cup_{i=1}^N \{\mathbf{m}_i\} = \mathbf{I}\,]$ and the Jacobian of each region $\{\mathbf{J}^{(i)}\}_{i=1}^N$ we seek to decompose each Jacobian to a joint and an individual component $\mathbf{J}^{(i)} \approx \mathbf{C}^{(i)} + \mathbf{A}^{(i)}$ that adheres to the following properties:

- The row spaces of the matrices capturing joint variation, i.e., joint matrices $\mathbf{C}^{(i)}$, are defined as sharing a common subspace denoted as $\text{Row}(\mathbf{C}) = \text{Row}(\mathbf{C}^{(i)}),\ \forall i \in (1, \cdots, N)$
- Components $\mathbf{A}^{(i)}$ are deemed individual since they are imposed to be orthogonal to the joint component, i.e. $\text{Row}(\mathbf{C}) \perp \text{Row}(\mathbf{A}^{(i)}),\ \forall i \in (1, \cdots, N)$
- The intersection of the row subspaces of the individual components is the zero vector space, $\cap_{i=1}^N \text{Row}(\mathbf{A}_i) = \mathbf{0}$

Let $\mathbf{J} = \left[\mathbf{J}^{(1)^T}, \cdots, \mathbf{J}^{(N)^T}\right]^T \in \mathbb{R}^{q \times d_h}$ be the concatenation of the Jacobians of each region along their rows, where $q = d_m^{(1)} + \cdots + d_m^{(N)}$. The joint and individual components are obtained by solving the following constrained optimization problem:

$$\min_{C, \{A^{(i)}\}_{i=1}^N} \left\| \mathbf{J} - \mathbf{C} - \left[\mathbf{A}^{(1)^T}, \cdots, \mathbf{A}^{(n)^T}\right]^T \right\|_{\text{F}}^2 \tag{5}$$

$$\text{s.t.} \quad \text{rank}(\mathbf{C}) = r_C, \ \ \{\text{rank}(\mathbf{A}^{(i)}) = r_A^{(i)}, \mathbf{C}\mathbf{A}^{(i)^T} = \mathbf{0}\}_{i=1}^N,$$

where $\mathbf{C} = \left[\mathbf{C}^{(1)^T}, \cdots, \mathbf{C}^{(N)^T}\right]^T \in \mathbb{R}^{q \times d_h}$ and $\{\mathbf{A}^{(i)} \in \mathbb{R}^{d_m^{(i)} \times d_h}\}_{i=0}^N$ are the joint and individual components of $\mathbf{J}$ respectively. We approximate the solution of this optimization problem by utilizing the iterative JIVE method as proposed in [18].

After the decomposition, the row subspace $\mathbf{V}_C$ obtained by the SVD of the joint component $\mathbf{C} = \mathbf{U}_C\mathbf{S}_C\mathbf{V}_C$ captures global variation in the entire image (e.g. changing the gender). In contrast, the row subspace $\mathbf{V}_A^{(i)}$ obtained by the SVD of the individual component $\mathbf{A}^{(i)} = \mathbf{U}_A^{(i)}\mathbf{S}_A^{(i)}\mathbf{V}_A^{(i)^T}$ captures local modes of variation specific to region $\mathbf{m}^{(i)}$. Figure 2 (middle) illustrates the effect of JIVE on the row subspaces of the Jacobians.

Calculating the JIVE decomposition directly on the set of Jacobians $\mathbf{J} \in \mathbb{R}^{q \times d_h}$ is highly impractical. For instance, in a standard DDPM [11], with image size $q = 256 \cdot 256 \cdot 3$ and latent dimension $d_h = 8 \cdot 8 \cdot 512$ the Jacobian $\mathbf{J}$ has approximately 6B parameters. The most computationally expensive step in the JIVE algorithm (see [18]) is that of SVD. Clearly applying SVD on a 6B parameter Jacobian is prohibitive in practice. To make such a computation feasible we adopt a dimension-reducing transformation: $\mathbf{J}^{(i)} \rightarrow \mathbf{J}_\perp^{(i)}$ where $\mathbf{J}_\perp^{(i)} = \mathbf{S}^{(i)}\mathbf{V}^{(i)^T}$ is an $r \times d_h$ matrix with $r \ll q$, derived from the SVD of $\mathbf{J}^{(i)}$ as obtained from the *subspace iteration* 4.1. The above mentioned approach is valid since the Euclidian distance between the columns of $\mathbf{J}^{(i)}$ is preserved in $\mathbf{J}_\perp^{(i)}$ [18].

5 Experiments

In this section, we present a series of experiments to validate the proposed method. Initially, we describe our experimental setup in Section 5.1. Then, in Section 5.2 we showcase the effectiveness of the individual components on localized edits. In Section 5.3 we show that our method can identify meaningful editing directions from a single image that generalize to other images. Finally in Section 5.4 we qualitatively and quantitatively compare our approach with existing alternatives for attribute manipulation in unconditional DMs.

5.1 Experimental Setup

We conduct our experiments on three different datasets, namely CelebA-HQ [17], LSUN-churches [37], and METFACES [12], using an unconditional DDPM[1] [2] [3] as the base model. We highlight that all models are pre-trained and kept frozen. We find that a large joint and a small individual rank $r_C < r_{A_i}$ yield the best results for local editing. This aligns with our intuition since the global modes of variation are expected to be more than the local. For all experiments presented, we set the rank of our dimension-reducing transformation to $r = 50$, the joint and individual rank to $r_C = 30$ and $r_A = 5$ respectively, and obtain the editing directions at timestep $t = 0.6T$. Editing results derived from different timesteps can be found in the Appendix.

For quantitative evaluation, we use Fréchet Inception Distance (FID), Identity Similarity (ID), and Region of Interest Ratio (ROIR) [21]. FID is utilized to evaluate the fidelity of the generated images after the edit. To assess identity similarity (ID) before and after the edit we use the ArcFace model [5]. To quantify local editing, we use ROIR [21], which is the ratio of the distance between pixels of the original and edited images in the region of 'disinterest',

[1] https://huggingface.co/google/ddpm-ema-celebahq-256
[2] https://huggingface.co/google/ddpm-ema-church-256
[3] https://github.com/jychoi118/P2-weighting

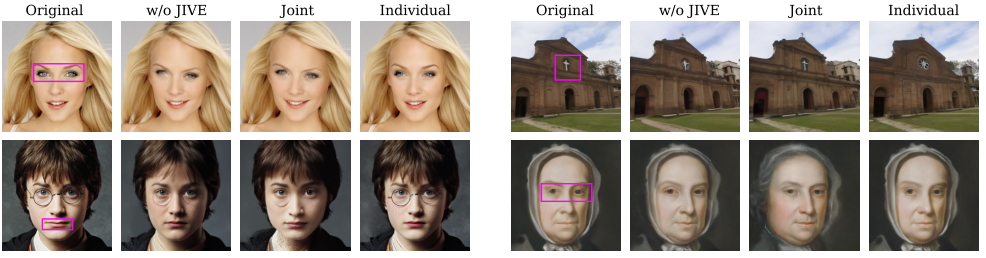

Figure 3: **Editing with the joint and individual components** for the CelebA-HQ, LSUN-Churches and METFACES datasets. Regions of interest are denoted by a pink rectangle. By decomposing the Jacobians of each region into a joint and individual component we can disentangle the global and the local semantic variation

over the same quantity in the region of interest. A small ROIR indicates localized edits, with large changes within the region of interest and small changes in the rest of the image:

$$\text{ROIR}(\mathcal{M}, \mathcal{X}, \mathcal{X}') = \frac{1}{N}\sum_{i=1}^{N}\frac{\|(\mathbf{1}-\mathcal{M})\cdot(\mathcal{X}_i - \mathcal{X}_i')\|}{\|\mathcal{M}\cdot(\mathcal{X}_i - \mathcal{X}_i')\|} \tag{6}$$

where $\mathcal{M} \in [0,1]^{H\times W\times C}$ is the mask specifying the region of interest, $\mathbf{1}$ is a 1-tensor and $\mathcal{X}, \mathcal{X}' \in \mathbb{R}^{N\times H\times W\times C}$ are a batch of original and edited images respectivly.

5.2 Editing within the Individual and Joint Subspaces

In this section, we validate the effectiveness of the JIVE decomposition on the Jacobians as presented in Section 4.2 on samples from CelebA-HQ, LSUN-Churches, and METFACES. In Figure 3 we show the effects of editing with latent vectors belonging to the row spaces of the Jacobian w/o JIVE, the individual component, and the joint component of the Jacobian respectively. First, we observe that when we edit a region $\mathbf{m}_i$ directly within the row space of the Jacobian i.e. $\mathbf{v}_i \in \mathbf{V}^{(i)}$, undesirable non-localized edits occur. For example, when editing the lips of Harry Potter, other attributes such as the glasses and the color of the cheeks are manipulated and when editing the church window, the background building is also altered. In the next column, we depict the effects of editing within the row space of the joint component $\mathbf{v}_i \in \mathbf{V}_C^{(i)}$. This results in global manipulations that affect the entire image, such as editing both eyes and mouth in the top left image from CelebA-HQ and changing the gender in the sample from METFACES. On the contrary, when using the vectors from the row space of the individual component $\mathbf{v}_i \in \mathbf{V}_A^{(i)}$ to edit the image region the manipulations are highly localized . In the last two rows of Table 1 we demonstrate this quantitatively for four attribute manipulations. When edits are derived from the individual component, FID and ROIR decrease, while ID increases indicating that we could achieve more precise control over a specific region, while better retaining image quality and identity similarity.

5.3 Qualitative results

Here we present qualitative results, as depicted in Fig. 4. We show that directions obtained from the individual component extracted from a region of interest by our method can perform various localized semantic edits. We highlight with a red rectangle, how semantic directions identified on a single reference image are transferred to the rest for each dataset. For CelebA-HQ samples, the eyes, eyebrows, and mouth regions are selected. For the eyes region, local

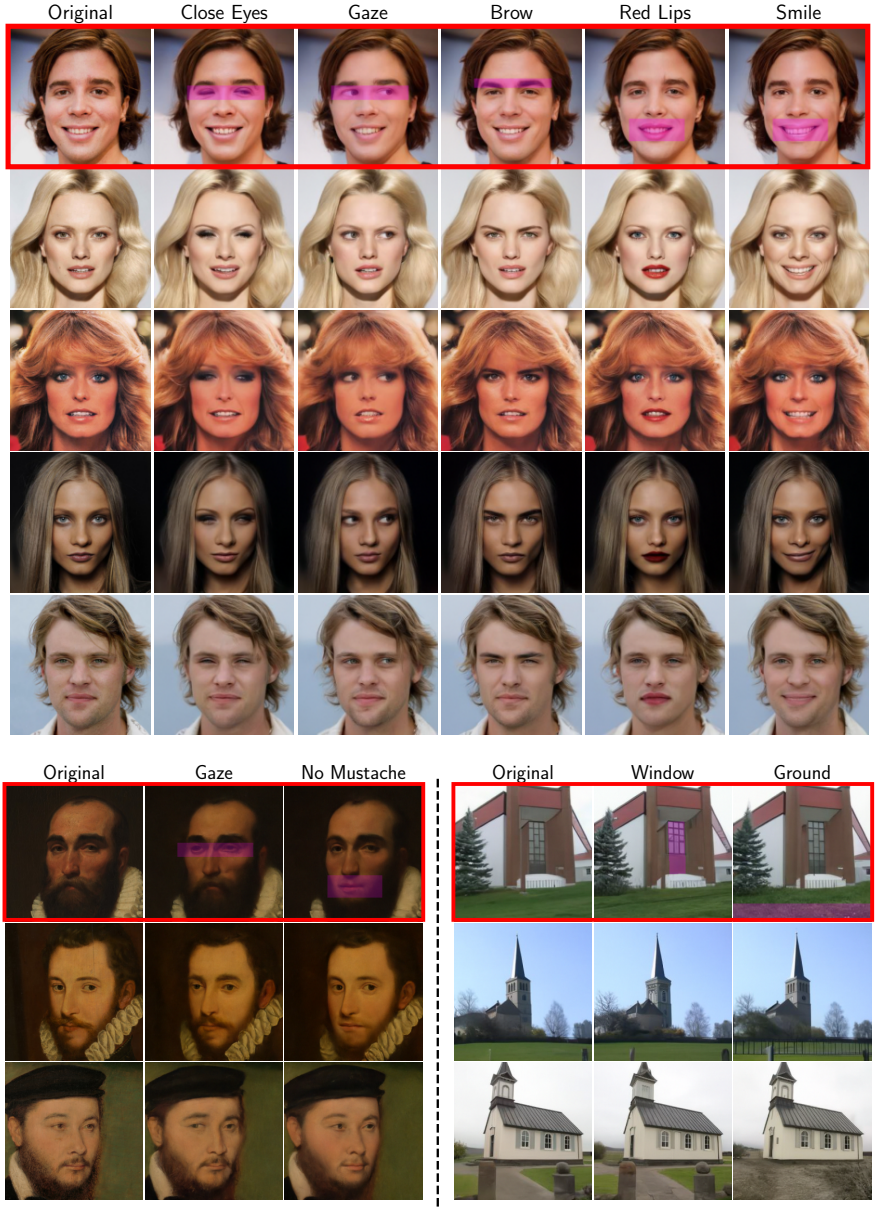

Figure 4: **Local editing results** on the CelebA-HQ (top), METFACES and LSUN-Churches (bottom). The region of interest is highlighted with pink rectangles. Our method can identify diverse semantic manipulations within a region while not affecting the rest of the image. Note that the latent vectors used to edit the images in each row are derived from the image in the first row.

semantic changes such as closing the eyes and changing the gaze are depicted in the second and third columns. In the fourth column, the eyebrows become thicker. In the last two

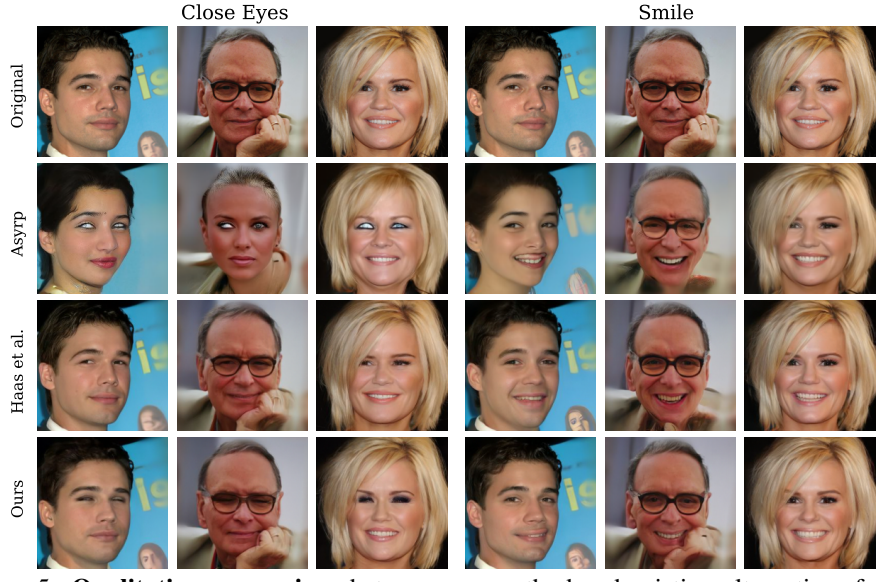

Figure 5: **Qualitative comparison** between our method and existing alternatives for two local edits.

Method	Close Eyes			Smile			Red Lips			Gaze		
	FID ↓	ID ↑	ROIR ↓	FID ↓	ID ↑	ROIR ↓	FID ↓	ID ↑	ROIR ↓	FID ↓	ID ↑	ROIR ↓
Asyrp [15]	60.63	0.40	3.95	67.9	0.57	4.32	78.27	0.32	3.75	84.17	0.12	6.04
Haas et al. [8]	51.42	**0.70**	5.18	52.16	0.68	4.30	**48.82**	0.69	6.61	52.28	0.68	4.05
Ours (wo/ JIVE)	52.53	0.67	3.66	54.20	0.72	3.62	51.46	0.73	4.32	51.23	0.68	3.53
Ours (w/ JIVE)	**49.16**	0.69	**2.87**	**51.90**	**0.78**	**2.73**	48.93	**0.74**	**3.26**	**48.11**	**0.71**	**3.07**

Table 1: **Quantitative results** between our method with and without the JIVE decomposition and existing alternatives for four localized edits on 5k CelebA-HQ samples.

columns where the mouth region is selected, we can manipulate the facial expression by adding a smile and changing the lip color. Similarly for METFACES, we present directions that manipulate the gaze in the eyes region and remove the mustache in the mouth region. Finally for LSUN-Churches, directions that alter the windows and the ground are presented.

5.4 Comparison with Other Methods

In this section, we compare our method both qualitatively and quantitatively with the state-of-the-art for attribute manipulation in unconditional DMs. Specifically, we compare our approach with Asyrp [15], and the method proposed by Haas et al. [8], two recently proposed methods that identify semantic directions in the latent space of unconditional DMs. We find the most relevant vectors that can control the eyes and smile according to their papers. Note that Asyrp is a supervised method, that uses CLIP [24] to achieve image edits. Also, the method of Haas et al. is equivalent to our method without using JIVE or constraining the Jacobians to a region of interest. As shown in Figure 5, our method retains the individual characteristics of the original image and the edits are better restricted to the region of interest than the other two methods. For the Close Eyes edit, Asyrp produces unrealistic artifacts in the eye region and even alters the gender of the first two images. While Haas et al. better retain the subject's identity, their method fails to fully close the eyes. For the Smile edit,

Asyrp swaps the first image subject's gender, and also significantly alters the second image, removing the hand. Haas et al. fail to produce edits as localized as our method, changing the subject's facial expression and characteristics outside of the specified region.

To quantify the comparison, we present in Table 1 quantitative experiments for four different manipulations on 5k real images from CelebA-HQ. Our method achieves comparable ID and FID to Haas et al. for the Close Eyes and Red Lips edits respectively, while outperforming both methods on all metrics for the rest of the attributes. As captured by the FID and ID metrics, our method produces edited images of higher fidelity, more closely resembling the original images. Finally, as captured by the ROIR metric, our method is better at producing localized edits that do not affect the rest of the image.

6 Conclusion and Future Work

In this work, we propose a method for localized semantic manipulation of real images using a pre-trained DM. Our method involves first associating specific regions of interest in an image to subspaces in the DMs latent space and then factorizing these subspaces to isolate their individual and joint variation. We find that the subspaces discovered from one image can be used to edit different images, making the computation of a new factorization unnecessary. By extensive qualitative and quantitative experiments, we establish that our method can produce meaningful edits that are localized to specific regions of interest while preserving the original image quality and identity better than previous methods. In future work we aim to explore the latent semantics of video diffusion models. The temporal dimension in video generation adds complexity to the latent space, making it an interesting research direction to identify latent subspaces corresponding to specific temporal moments.

Acknowledgement This work has been partially supported by project MIS 5154714 of the Natonal Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGeneraton EU Program and by a grant from The Cyprus Institute on Cyclone.

References

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18187–18197, 2021. URL https://api.semanticscholar.org/CorpusID:244714366.

[2] Jaewoong Choi, Junho Lee, Changyeon Yoon, Jung Ho Park, Geonho Hwang, and Myungjoo Kang. Do not escape from the manifold: Discovering the local coordinates on the latent space of gans. In *International Conference on Learning Representations*, 2021.

[3] Jaewoong Choi, Geonho Hwang, Hyunsoo Cho, and Myungjoo Kang. Finding the global semantic representation in gan through fréchet mean. In *The Eleventh International Conference on Learning Representations*, 2022.

[4] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *The Eleventh International Conference on Learning Representations*, 2022.

[5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[8] René Haas, Inbar Huberman-Spiegelglas, Rotem Mulayoff, and Tomer Michaeli. Discovering interpretable directions in the semantic latent space of diffusion models. *arXiv preprint arXiv:2303.11073*, 2023.

[9] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in neural information processing systems*, 33:9841–9850, 2020.

[10] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2022.

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

[12] 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.

[13] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6007–6017, June 2023.

[14] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022.

[15] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. In *The Eleventh International Conference on Learning Representations*, 2022.

[16] 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. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 289–299, 2023.

[17] 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.

[18] Eric F Lock, Katherine A Hoadley, James Stephen Marron, and Andrew B Nobel. Joint and individual variation explained (jive) for integrated analysis of multiple data types. *The annals of applied statistics*, 7(1):523, 2013.

[19] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11461–11471, 2022.

[20] 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 *International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022.

[21] James Oldfield, Christos Tzelepis, Yannis Panagakis, Mihalis Nicolaou, and Ioannis Patras. Panda: Unsupervised learning of parts and appearances in the feature maps of gans. In *The Eleventh International Conference on Learning Representations*, 2022.

[22] Yong-Hyun Park, Mingi Kwon, Jaewoong Choi, Junghyo Jo, and Youngjung Uh. Understanding the latent space of diffusion models through the lens of riemannian geometry. *Advances in Neural Information Processing Systems*, 36:24129–24142, 2023.

[23] 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.

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[25] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[26] 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.

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[28] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[29] Vikash Sehwag, Caner Hazirbas, Albert Gordo, Firat Ozgenel, and Cristian Canton. Generating high fidelity data from low-density regions using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11492–11501, 2022.

[30] 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.

[31] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[32] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020.

[33] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[34] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020.

[35] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, pages 32211–32252. PMLR, 2023.

[36] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. 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 (CVPR)*, pages 1921–1930, June 2023.

[37] 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 preprint arXiv:1506.03365*, 2015.

[38] Guanhua Zhang, Jiabao Ji, Yang Zhang, Mo Yu, Tommi Jaakkola, and Shiyu Chang. Towards coherent image inpainting using denoising diffusion implicit models. In *International Conference on Machine Learning*, pages 41164–41193. PMLR, 2023.

[39] Jiapeng Zhu, Ruili Feng, Yujun Shen, Deli Zhao, Zheng-Jun Zha, Jingren Zhou, and Qifeng Chen. Low-rank subspaces in gans. *Advances in Neural Information Processing Systems*, 34:16648–16658, 2021.

[40] Jiapeng Zhu, Yujun Shen, Yinghao Xu, Deli Zhao, and Qifeng Chen. Region-based semantic factorization in gans. In *International Conference on Machine Learning*, pages 27612–27632. PMLR, 2022.