Personalized Face Inpainting with Diffusion Models by Parallel Visual Attention

Jianjin Xu[1] Saman Motamed[1,3] Praneetha Vaddamanu[1] Chen Henry Wu[1]
Christian Haene[2] Jean-Charles Bazin[2] Fernando de la Torre[1]
[1]*Carnegie Mellon University*
[2]*Independent Researcher* [3]*INSAIT, Sofia University*

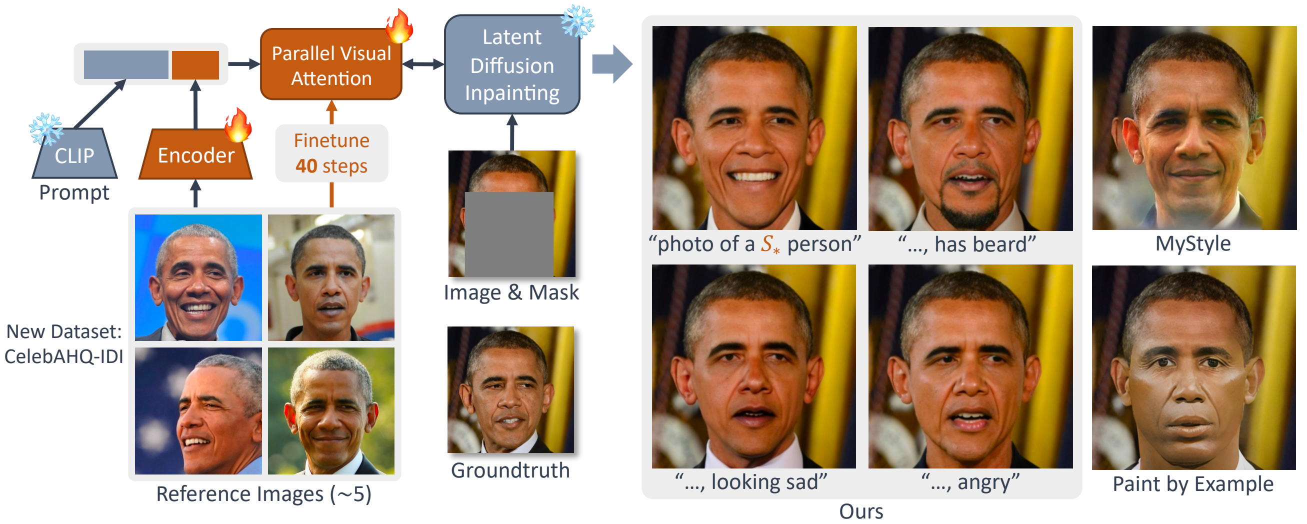

Figure 1. We address the task of identity-preserving language-controllable face inpainting by adding Parallel Visual Attention (PVA) to a pretrained diffusion model. PVA enhances the diffusion model to condition on the reference images, thereby preserving the identity. PVA achieves the best identity similarity and image quality compared to several baselines including MyStyle [35] and Paint by Example [62], even when editing the image with a prompt that changes the expression or iconic changes (e.g., beard, lipstick, changing hair style).

Abstract

Face inpainting is important in various applications, such as photo restoration, image editing, and virtual reality. Despite the significant advances in face generative models, ensuring that a person's unique facial identity is maintained during the inpainting process is still an elusive goal. Current state-of-the-art techniques, exemplified by MyStyle, necessitate resource-intensive fine-tuning and a substantial number of images for each new identity. Furthermore, existing methods often fall short in accommodating user-specified semantic attributes, such as beard or expression.

To improve inpainting results, and reduce the computational complexity during inference, this paper proposes the use of Parallel Visual Attention (PVA) in conjunction with diffusion models. Specifically, we insert parallel attention matrices to each cross-attention module in the denoising network, which attends to features extracted from reference images by an identity encoder. We train the added attention modules and identity encoder on CelebAHQ-IDI, a dataset proposed for identity-preserving face inpainting. Experiments demonstrate that PVA attains unparalleled identity resemblance in both face inpainting and face inpainting with language guidance tasks, in comparison to various benchmarks, including MyStyle, Paint by Example, and Custom Diffusion. Our findings reveal that PVA ensures good identity preservation while offering effective language-controllability. Additionally, in contrast to Custom Diffusion, PVA requires just 40 fine-tuning steps for each new identity, which translates to a significant speed increase of over 20 times.

1. Introduction

The task of reconstructing absent regions in face images (i.e., face inpainting) is key to various fields, such as virtual reality, photo editing, and photo restoration. In recent years, as masks became a common sight due to COVID-19, many photographs captured at social gatherings or tourist attractions featured individuals wearing masks. There is a growing interest in digitally removing these masks to reveal the person's true appearance beneath. Beyond simply filling in the covered areas, there's a demand for manipulating the restored image in ways like altering facial expressions through descriptive language—a task defined as identity-preserving, language-controllable face inpainting. This technology also has broader implications, like removing sunglasses in a personal photo, inpaiting the eyes in VR meetings, and other variety of restoration or editing purposes.

Maintaining a person's recognizable features is crucial in face inpainting tasks. Take the example of individuals aiming to upload restored images on social platforms or estimate the face of a user with VR glasses; in these applications it's essential that the user can still be recognized. Standard face inpainting tools that lack the ability to retain the user's likeness would be of minimal practical benefit. On the other hand, current strategies for identity-conserving face inpainting often rely on one or several reference photos of the individual. However, the integration of these reference photos into the inpainting process poses a significant technical challenge that has yet to be fully solved.

A SOTA technique for identity-preserving face inpainting is MyStyle [35], which finetunes a pre-existing StyleGAN model using several reference photos of a particular individual. Nevertheless, MyStyle necessitates more than 40 images to maintain image quality; reducing the number of images results in lower quality outputs. In contrast, newer methods that use diffusion processes for personalizations [13,46], like Custom Diffusion [28], have shown remarkable success in customizing a diffusion model using as few as five images. These models excel at producing highly accurate and linguistically controllable images of specific subjects, including people. While these approaches have not yet been specifically adapted for inpainting tasks in existing research, our study has taken the initiative to apply these personalization techniques to face inpainting. We found that they deliver promising results. However, a significant drawback of these methods is the computational cost. For example, adapting MyStyle to a new individual takes several minutes, whereas diffusion-based personalization methods require upwards of four hours on an A4000 GPU.

This study presents a new approach to reduce the computational expense associated with identity-preserving and language-controllable face inpainting. We introduce an auxiliary channel alongside diffusion models, coupled with a new component, the PVA. This new workflow is illustrated in Fig. 1. For each cross-attention [57] in the denoising UNet [45], we introduce a new set of $\{\mathbf{Q}', \mathbf{K}', \mathbf{V}'\}$ matrices that attend to visual features extracted with a vision encoder. In training, we freeze the denoising UNet and only train the vision encoder and those new matrices. As a result, we only need 40 steps of finetuning for a new subject in inference, which costs less than 1 minute on a single GPU, resulting in over 20 times acceleration over Custom Diffusion [28].

Assessing the quality of identity-preserving face inpainting requires a dataset that offers a variety of identities, multiple reference images for each identity, and high-resolution images. Currently, there isn't a definitive benchmark for such evaluations. CelebA-IDD [12] could have been a candidate, but it falls short in resolution and is no longer accessible. To bridge this gap, we have developed a new benchmark dataset named CelebAHQ-IDI. This dataset is curated from the existing CelebAHQ dataset [29], sorted based on the availability of multiple reference images for individual identities. Additionally, we have created semantic occlusion masks that conceal parts of the face, such as the lower half or the eyes and eyebrows, to mimic real-life occlusions. This dataset will serve as the new standard for evaluating identity-preserving face inpainting tasks.

PVA was trained and tested using a subset of the CelebAHQ-IDI dataset, specifically CelebAHQ-IDI-5, which provides five reference images for each identity. We trained PVA on the training partition of CelebAHQ-IDI-5 and evaluated it on the test partition, which contains identities that PVA had not previously encountered. The effectiveness of PVA in inpainting was assessed from two critical perspectives: how well it preserved the identity and the overall quality of the image output. These were quantitatively measured using a pretrained face recognition network to determine identity preservation and Frechet Inception Distance (FID [16]) and Kernel Inception Distance (KID) [6] scores for image quality. PVA's performance was benchmarked against five other methods, including the notable MyStyle and Custom Diffusion models. Our results indicated that PVA surpassed all the baseline methods, achieving the highest scores in identity preservation and the lowest in FID, confirming its superior performance in both preserving identity and maintaining high image quality.

Finally, to assess the language-directed editing capabilities of our method, we crafted 15 distinct prompts aimed at altering facial expressions, actions, and accessories, among other features. We quantified the degree of alignment between the modified image and the textual prompt using the CLIP score metric [15]. Our findings reveal a balancing act between maintaining the subject's identity and achieving the desired linguistic edits. Our PVA approach achieved the best results in preserving identity. At the same time, it offered language control on par with other methods like Textual Inversion and Custom Diffusion.

2. Related Work

Generative Adversarial Networks (GANs). The classic architecture of GANs [14, 40] consists of a generator and a discriminator. The discriminator is trained to distinguish generated images and guides the training of the generator. Currently, StyleGAN [24, 25] models hold SOTA generation quality on aligned image domains, like face, car, cat, *etc*. [47, 48]. Besides image generation, GANs also empower various applications like image-to-image translation [9, 20, 36, 42, 67], image inpainting [12, 33, 35, 65], and image editing [1, 5, 43, 49, 50, 60, 61, 66]. The Pivotal Tuning Inversion (PTI) [43] is an important technique to adapt a generic GAN to a customized object for image editing purposes. MyStyle [35] builds on top of PTI and finetunes a GAN on the images of a specific person. MyStyle achieves good results in identity-preserving inpainting, super-resolution, and editing. However, it requires 40+ images for each person and fewer images result in a severe loss of image quality. In comparison, our method only requires 5 images.

Diffusion models. Diffusion models [17, 54, 55] generate data by reversing a data corruption process. Recently, latent diffusion models [44] have been shown to be effective for high-resolution image synthesis tasks. Among these tasks, text-to-image generation [34, 41] aims at generating faithful images based on a text prompt. To adapt a generic text-to-image diffusion model to generate images of a specific object, recent work proposed to finetune text embeddings [13] or the diffusion model itself [28, 46]. However, previous adaptation methods require finetuning for hundreds or thousands of steps, making it inefficient in practical applications. In comparison, our adaptation method only needs 40 steps of finetuning for the adaptation.

Identity-preserving face inpainting. Most existing methods for identity-preserving face inpainting use a GAN architecture. The GAN is usually augmented with a pathway that incorporates features from an exemplar image [12, 30, 33, 65]. Dolhansky *et al*. [12] proposes the ExGAN architecture for eye inpainting only. Zhao *et al*. [65] is trained with 128px only images. EXE-GAN [33] proposes to encode an exemplar image into the style vector of a StyleGAN-like inpainting network. However, it only shows qualitative results for identity-preserving inpainting and has no quantitative evaluation. At the time of this work, EXE-GAN has not been open-sourced and we could not compare to it. The closest existing works to ours are MyStyle and the diffusion-based personalization algorithms. We use them as baselines for comparisons.

3. Method

3.1. Problem Definition

We address the task of identity-preserving language-controllable face inpainting. Given a set of reference images ($\mathcal{R}_p$) and a set of inference images and masks ($\mathcal{I}_p$) for identity p, we aim to inpaint the set of masked images ($\mathcal{C}_p$) such that the inpainted images can still be perceived as identity p. We define $\mathcal{R}_p = \{\mathbf{x}_r\}_{i=1}^{N_p^{\text{ref}}}$, where N_p^{ref} is the number of reference images for identity p; $\mathcal{I}_p = \{(\mathbf{x}_i, \{\mathbf{m}_{i,j}\}_{j=0}^{N_{p,i}^{\text{mask}}})\}_{i=1}^{N_p^{\text{infer}}}$, where $\mathbf{m}_{i,j}$ is the j-th corruption mask for the i-th image in identity p; and $\mathcal{C}_p = \{\mathbf{x}_i \odot \mathbf{m}_{i,j}\}$. Additionally, we consider providing language control over the inpainted content.

3.2. Background

Denoising Diffusion Probabilistic Models. DDPM models data as a sequence $\{\mathbf{x}_t\}_{t=1}^T$ where Gaussian noise is gradually added into the original data $\mathbf{x}_0$. In time step t, Gaussian noise with variance β_t is injected,

$$\mathbf{x}_{t+1} = \sqrt{1-\beta_t}\mathbf{x}_t + \sqrt{\beta_t}\boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $0 < \beta_t < 1$. As t increases, the original data $\mathbf{x}_0$ gradually disappears and $\mathbf{x}_t$ approximates a normal distribution. Multiple diffusion steps can be combined and expressed in a concise formulation:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t, \tag{2}$$

where $\bar{\alpha}_t = \Pi_{i=1}^t(1-\beta_i)$, indicating the down-weighting factor of the data term as a result of diffusion.

To generate new data points, DDPM reverses the corruption process by learning a denoising network, $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$, that tries to predict the noise added to $\mathbf{x}_t$. The network is trained to minimize the Denoising Score Matching (DSM) loss,

$$\mathcal{L}_{\text{DSM}} = \mathop{\mathbb{E}}_{\mathbf{x}_0, t, \boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\|_2^2\right], \tag{3}$$

where $\mathbf{x}_t$ is perturbed from $\mathbf{x}_0$ under noise $\boldsymbol{\epsilon}$ and t is uniformly sampled from all possible time steps. There are plenty of sampling algorithms available for DDPM [4, 64] and one of the most commonly used algorithms is DDIM [53]. We use DDIM sampling throughout our experiments.

Latent Diffusion Models. LDM [44, 52] is proposed to reduce the training cost of DDPM by building diffuison models on the latent space of Variational Auto-Encoder (VAE) [27]. If we denote the encoder by $\mathbf{E}_V(\cdot)$ and the decoder by $\mathbf{D}_V(\cdot)$, the encoding and decoding process are $\mathbf{z}_t = \mathbf{E}_V(\mathbf{x}_t)$ and $\mathbf{x}_t = \mathbf{D}_V(\mathbf{z}_t)$, respectively. The LDM is often conditioned on text prompts through cross-attention [57] modules, which attend to the text features. The text features are extracted with a pretrained text encoder (usually CLIP [39]), denoted by $\mathbf{y}_i = \mathbf{E}_T(T_i)$. In the inpainting task, LDM is also conditioned on the occlusion mask and the occluded image. The image and mask are directly concatenated to the input of LDM, denoted by $\tilde{\mathbf{z}}_t = \mathbf{z}_t || u_\downarrow(\mathbf{m}) || \mathbf{E}_V(\mathbf{m} \odot \mathbf{x}_0)$, where $u_\downarrow(\mathbf{m})$ means to downsample the mask $\mathbf{m}$ to the resolution of $\mathbf{z}_0$. The encoder for the masked image is the same as the one used for clean images.

Our method is built on top of the Latent Diffusion Inpainting (LDI) model, which is conditioned on texts, images, and masks. LDI is initialized from a pre-trained latent diffusion checkpoint and then trained with the inpainting LDM objective:

$$\mathcal{L}_{\text{LDM}} = \mathop{\mathbb{E}}_{\mathbf{z}_0, \mathbf{y}, \mathbf{m}, t, \boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\tilde{\mathbf{z}}_t, \mathbf{y}, t)\|_2^2\right]. \tag{4}$$

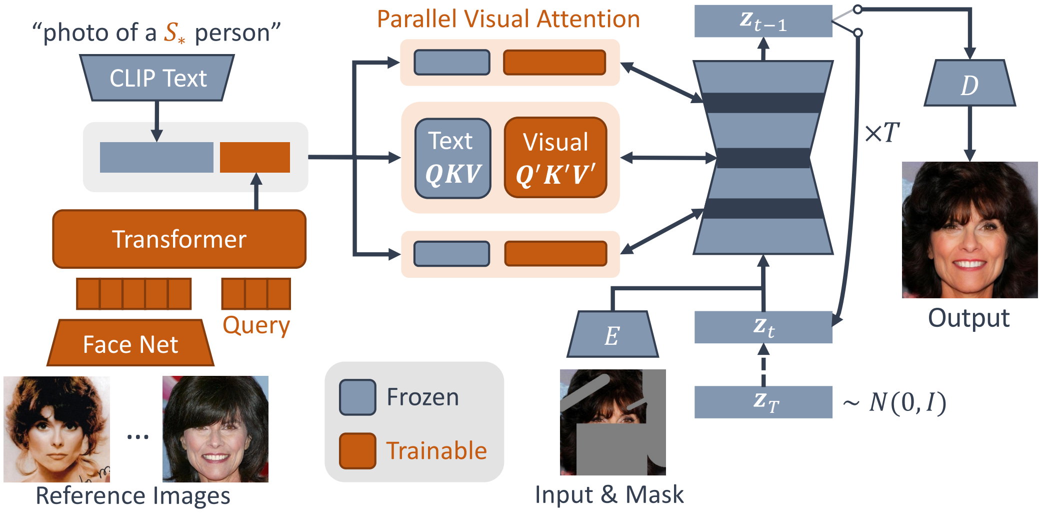

Figure 2. The proposed PVA pathway for incorporating reference images into a pretrained diffusion model. Each PVA module is modified from a cross-attention module (shown as a dark stripe on the U-Net) by adding a new set of QKV matrices. All pretrained parameters of the denoising U-Net are frozen in training.

During training, the conditioning is dropped 10% of the time for classifier-free guidance [18].

Personalized Diffusion Models. There are three typical methods for the personalization of diffusion models at the time of submission. Textual Inversion (TI) [13] finetunes the text embedding for a personalized token initialized from the original category token. For example, the photo for a person can be described as "A photo of S_*", where the embedding of S_* is the parameter to be optimized. DreamBooth [46] finetunes the whole diffusion model and uses prompts with rare tokens as modifiers, *e.g*., "A photo of S_* person". Custom Diffusion [28] uses the same style of prompts as DreamBooth, but only finetunes the cross-attention modules and additionally tunes the embedding of the rare token.

It is observed that the diffusion models easily overfit the reference images, reducing the quality and diversity of generated images. Therefore, both DreamBooth and Custom Diffusion need to be trained with a prior regularization loss to fight against the overfitting issue. This means that an additional set of regularization images needs to be collected, making the algorithm more complicated.

3.3. Parallel Visual Attention Pathway

We observe two main limitations in existing methods. First, high inference costs. Whenever a new object is personalized, they necessitate a computationally expensive finetuning process. Second, additional data costs due to the prior regularization loss. We propose to reduce these costs with the Parallel Visual Attention (PVA) pathway. The PVA pathway consists of two components. First, a feed-forward encoder that extracts identity features from reference images. Second, the PVA module that allows the denoising network to condition on visual features without changing existing parameters. The pipeline is shown in Fig. 2.

3.3.1 Identity Encoder

It is common practice to accelerate in-inference optimizations by feed-forward networks [19, 22, 59, 66]. A well-known example is GAN inversion [1, 25, 42, 66], where an encoder is trained to predict the latent code given an image or segmentation. The predicted latent code is a good initialization and thus is able to accelerate the inversion process. Inspired by this, we also use a feed-forward model to accelerate the finetuning process in model personalization.

Visual feature conditioning. There are various ways to incorporate visual features into diffusion models. Image-augmented diffusion models [7, 8, 51] concatenate the image features to text features directly. Paint by Example [62] removes the text features and only keeps the visual features. ControlNet [63] trains a siamese network to learn the residual to refine the original diffusion network. However, it assumes the condition to be spatially aligned with the generated image, such as edge maps and depth maps. In our task, the reference images might have a different pose than the image to be inpainted. Therefore, we choose to adopt the concatenation scheme.

Identity feature extractor. Existing methods mostly use a pretrained CLIP vision encoder as the visual feature extractor [7, 8, 51] due to its versatility. However, CLIP is not trained to distinguish the nuanced differences in human faces. So we use a pretrained face recognition network (referred to as FaceNet for convenience) as the feature extractor.

The extracted features from the FaceNet are further processed by a transformer [57]. As shown in Fig. 2, the inputs of the transformer are the FaceNet features from M images and N_{query} trainable query tokens. The output features of query tokens are treated as visual features and later concatenated to text features. We discard positional encoding in input because the extracted features should be invariant to the ordering of reference images.

3.4. Parallel Visual Attention

Personalizing a pretrained diffusion model on a few images of a specific object is prone to overfitting, resulting in uniform backgrounds, reduced diversity, degraded quality, *etc*. [46]. We think the main cause of overfitting is the large capability of finetuned parameters. Thus, we propose to finetune as fewer parameters as possible and even freeze the pretrained model.

Our solution is the Parallel Visual Attention module. The PVA module is modified from the cross-attention module and we first introduce its formulation here. For convenience, we use the formulation of single-head attention which can be trivially extended to the multi-head case. A cross-attention module consists of $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{D \times D}$ matrices. Given the flattened input feature map $\mathbf{F}_l \in \mathbb{R}^{HW \times D}$ at layer l and the conditional features $\mathbf{G} \in \mathbb{R}^{L \times D}$, the cross-attention computes the output as

$$\mathbf{F}_{l+1} = \mathbf{MGV}, \tag{5}$$

$$\mathbf{M} = \text{Softmax}\left[\frac{\mathbf{F}_l\mathbf{Q} \cdot (\mathbf{GK})^T}{\sqrt{D}}\right]. \tag{6}$$

The PVA module introduces a new set of attention matrices, $\{\mathbf{Q}', \mathbf{K}', \mathbf{V}'\}$, which attend to the visual features and compete with the attention on text features. PVA computes the output as follows,

$$\mathbf{F}_{l+1} = \mathbf{M}\left[\mathbf{G}_T\mathbf{V}, \mathbf{G}_V\mathbf{V}'\right], \tag{7}$$

$$\mathbf{M} = \text{Softmax}([\mathbf{S}_T, \mathbf{S}_V]), \tag{8}$$

$$\mathbf{S}_T = \frac{\mathbf{F}_l\mathbf{Q} \cdot (\mathbf{G}_T\mathbf{K})^T}{\sqrt{D}}, \tag{9}$$

$$\mathbf{S}_V = \frac{\mathbf{F}_l\mathbf{Q}' \cdot (\mathbf{G}_V\mathbf{K}')^T}{\sqrt{D}}, \tag{10}$$

where $[\cdot, \cdot]$ denotes tensor concatenation, and $\mathbf{G}_T$, $\mathbf{G}_V$ denotes the text features and visual features, respectively. The PVA module computes the textual and visual attention scores, $\mathbf{S}_T$ and $\mathbf{S}_V$, using two separate sets of attention matrices, and then applies softmax on the concatenated scores. The output is obtained by weighted sum over text features and visual features transformed by $\mathbf{V}$ or $\mathbf{V}'$, separately.

A notable characteristic of the PVA module is that when there is no visual feature, the PVA module falls back to the original cross-attention module, and the denoising network becomes identical to the pretrained one.

# Ref.	# Infer.	Total Images	# IDs
1	23704	28208	4504
2	19254	26364	3555
3	15694	24328	2878
4	12799	22335	2384
5	10396	20211	1963
10	3387	10857	747
15	868	4468	240
20	120	1040	46

Table 1. Statistics of CelebAHQ-IDI dataset with different numbers of available reference images. "# Ref." and "# Infer." refer to the number of reference images for each identity and the number of total inference images, respectively.

3.5. Training

We train the embedding of the special token, the PVA modules, the transformer, and the FaceNet. The training objective augments the LDM objective (Eq. (4)) with extra visual feature conditions,

$$\mathcal{L}_{\text{LDM}} = \mathop{\mathbb{E}}_{\mathbf{z}_0,\mathbf{y},\mathbf{m},\{\mathbf{x}_r\},t,\boldsymbol{\epsilon}}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\tilde{\mathbf{z}}_t, \mathbf{y}, \mathbf{E}_I(\{\mathbf{x}_r\}), t)\right\|_2^2\right], \tag{11}$$

where $\{\mathbf{x}_r\}$ is the set of reference images for the input image $\mathbf{x}_0$ and $\mathbf{E}_I$ is the identity encoder (Sec. 3.3). PVA does not need condition dropping because when the condition is dropped, the model falls back to the frozen pretrained model.

In training, we also need to provide plausible captions for images. Previous methods [13, 28, 46] use templates to generate coarse descriptions of an object, *e.g*., "A nice photo of ...". However, image inpainting is inherently ambiguous, such as recovering the expression of a person occluded by a mask. We believe using detailed captions would be beneficial for training. We use the captions provided by the CelebAHQ-Dialog dataset [21], which contains detailed language descriptions of various facial attributes, including gender, age, expression, *etc*. As we do not assume to have access to detailed captions in inference, we alternatively sample from generic prompts and detailed prompts in training.

3.6. Inference

In inference, we can either deploy the trained model directly or run a lightweight finetuning to fit the unseen identity better. This design choice is ablated in Sec. 4.2.3. The finetuning uses the same objective as in training, but only tunes the PVA modules. The finetuning process consists of just 40 iterations, which takes less than 1 minute on a single GPU.

4. Experiment

4.1. Setup

Pretrained models. We use a pretrained latent diffusion inpainting (LDI) model. For the FaceNet as a part of the iden-

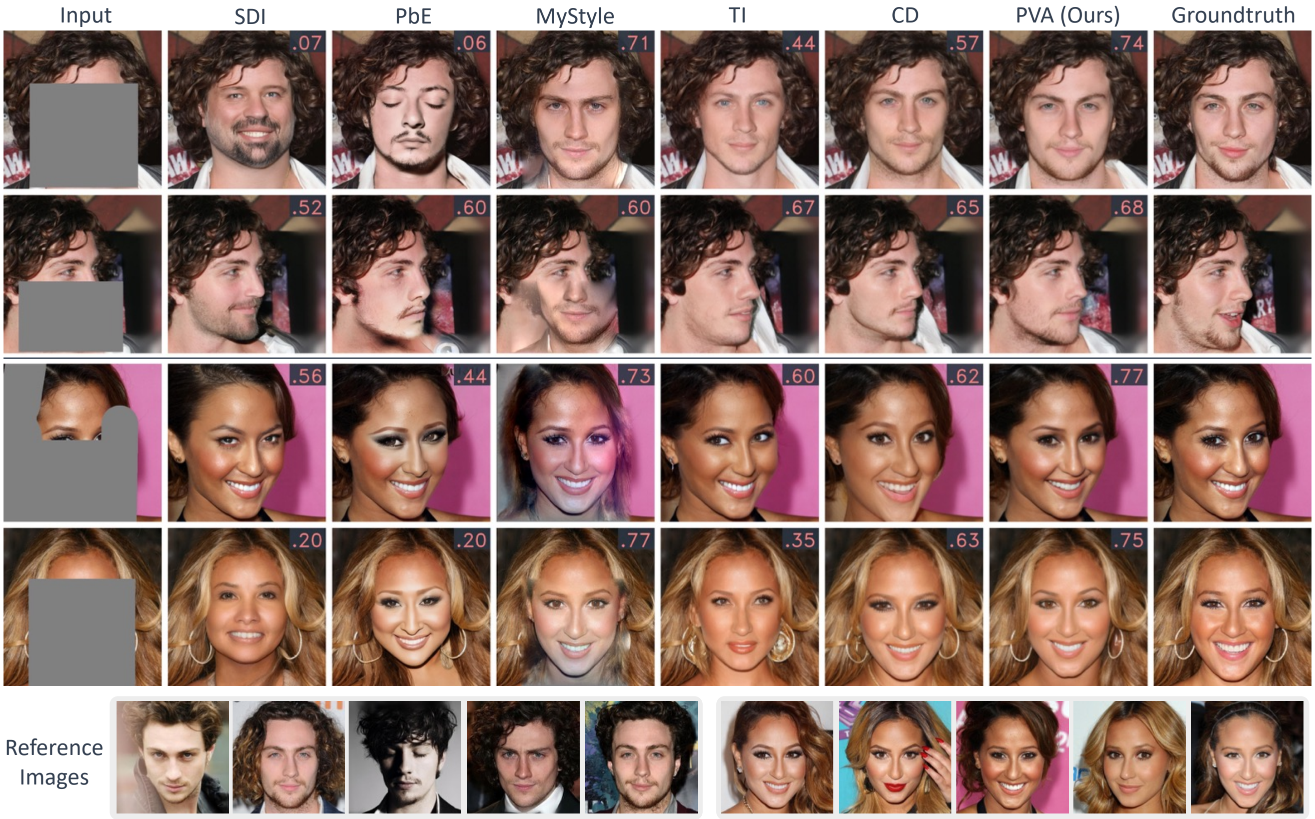

Figure 3. Inpainting results of PVA and baselines. The column tabs, “LDI, PbE, MyStyle, TI, CD” denotes Latent Diffusion Inpainting, Paint by Example [62], MyStyle [35], Textual Inversion [13], and Custom Diffusion [28], respectively. The upper right numbers of each image are the identity similarity (↑) between the inpainted image and the groundtruth. All the methods use the same 5 reference images shown in the bottom row. Rows 1 and 2 use the reference images on the left and Rows 3 and 4 use those on the right.

tity encoder, we use the ArcFace [10] R50 network trained on MS1MV3 dataset [11]. To calculate the identity similarity, we use a different pretrained network, the CosFace [58] R100 network trained on the Glint360K [2] dataset. Both pretrained networks are obtained from InsightFace [3].

Dataset. We used CelebAHQ, CelebAHQ-Dialog [21], and the images-of-celebs dataset[1]. We constructed a new dataset, CelebAHQ-IDI for identity-preserving face inpainting. CelebAHQ-IDI was built from the images in CelebAHQ and the identity annotation in CelebA [31]. To allow a fair comparison between algorithms, we reorganized the images according to the number of reference images. In this work, we mainly use the CelebAHQ-IDI-5 set, which has 5 reference images per identity. We also constructed several types of semantic occlusion masks that covered “lower face”, “eye & brow”, “whole face”, and “random” regions. Some examples of the masks could be found in Fig. 3. The statistics of CelebAHQ-IDI are shown in Tab. 1. For the detailed construction pipeline, please refer to the appendix.

Evaluation. We compared PVA to 5 baselines, the original Latent Diffusion Inpainting (LDI) [44], Paint by Example (PbE) [62], MyStyle [35], Textual Inversion (TI) [13] and Custom Diffusion (CD) [28]. We did not compare to DreamBooth [46] because it needed to store the whole finetuned Diffusion model for each identity, resulting in prohibitive storage consumption. Custom Diffusion could be regarded as an equivalence for DreamBooth as they are technically similar and have close performance [28].

We evaluate the performance of face inpainting in two aspects, ID similarity and image quality. Identity similarity was measured by the cosine similarity between the features of the inpainted images and the groundtruth images. The features were extracted by the CosFace R100 pretrained model. The image quality was measured by the Frechet Inception Distance (FID) [16] and Kernel Inception Distance (KID) [6,23] between features of inpainted images and groundtruth images. We used the clean-fid [37] implementation and InceptionV3 [56] as the feature extractor.

For the evaluation of language controllability, we created 15 edit prompts covering different facial expressions, makeup, action, and accessories. We ran inpainting algo-

[1] https://github.com/images-of-celebs/images-of-celebs

Method	FT. Time	ID ↑	FID ↓	KID ↓ $\times 10^{-3}$
LDI	-	0.359	8.24	**2.717**
Paint by Example	-	0.430	11.2	6.089
MyStyle	$\sim$ 15min	0.696	27.7	5.029
Textual Inversion	$\sim$ 6h	0.644	13.8	8.404
Custom Diffusion	$\sim$ 3h	0.729	13.9	5.870
PVA (Ours)	$\sim$ 1min	**0.741**	**8.22**	4.289

Table 2. Quantitative comparisons of identity similarity and image quality on the CelebAHQ-IDI-5 dataset. The finetuning costs of each method are indicated in Col "FT. Time", measured in single GPU (RTX A4000) time. "ID" stands for identity similarity.

rithms on images with "whole face" masks, ensuring a sufficient degree of freedom for editing. The level of controllability was measured by the text alignment between the inpainted images and the target prompts. We used the CLIP score [15] on the ViT-B/32 backbone as the metric for text alignment. A few examples of the prompts can be found in Fig. 4. The full list of prompts is described in the appendix.

Implementation details. We mainly conducted training and evaluation on the CelebAHQ-IDI-5 dataset. We trained our model for 200K iterations using the AdamW [26, 32] optimizer with batch size 16, learning rate 1.6×10^{-5}, and weight decay 10^{-2}. In finetuning, we used the same settings and trained for just 40 steps. We used the same sampling method across all diffusion-based methods, which was DDIM sampler with 100 steps and $\eta = 0.7$. We used PyTorch [38] to implement the algorithms. Other implementation details are described in the appendix. [2]

4.2. Results

We evaluated the performance of PVA on the face inpainting task and language-controlled inpainting task. We also ablated some design choices in PVA in Sec. 4.2.3.

4.2.1 Identity-Preserving Face Inpainting

Qualitative results. The qualitative comparisons of PVA to baselines are shown in Fig. 3. It was observed that our method PVA outperformed all baselines. In terms of identity similarity, all baselines had noticeable shifts in identity. The original diffusion model, LDI, produced plausible inpainting but the inpainted person was very different since it had no information about the identity. Paint by Example captured some characteristics of the identity like the small mustache (Row 1 Col 4) but failed to preserve the identity. All images inpainted by MyStyle had noticeable inconsistencies, in particular the side view face photo (Row 2 Col 3). Textual Inversion and Custom Diffusion performed relatively better among the baselines, yet they both failed to inpaint the jaw correctly, *e.g.*, in Row 1 Col 5 and 6. In comparison, PVA achieved the best identity similarity and image quality among all methods.

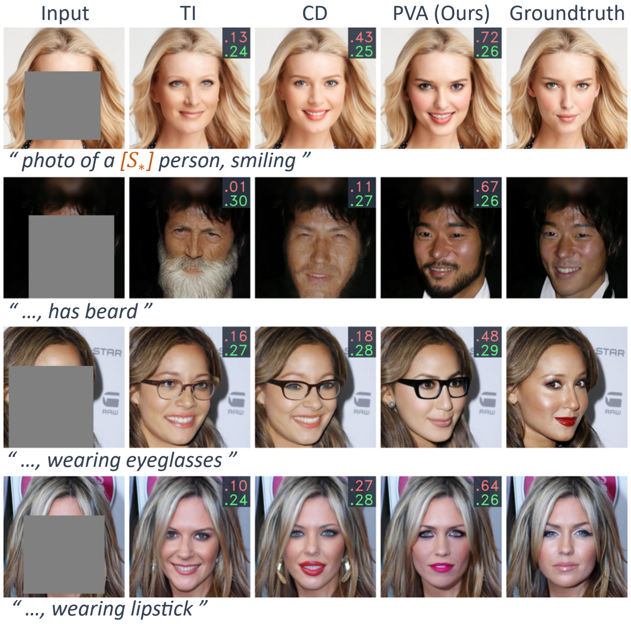

Figure 4. Qualitative comparisons of identity-preserving language-controlled inpainting. Prompts for editing are shown at the bottom of each row. We annotate the ID similarity (line 1) to the groundtruth and the CLIP similarity (line 2) to the text prompt at the upper right of each inpainted image.

Quantitative results. The evaluation results of identity similarity and image quality are presented in Tab. 2. It was observed that PVA outperformed all baselines on identity similarity and FID. Although PVA was a bit below LDI on KID, the LDI had the lowest identity similarity, indicating that LDI could not preserve the identity. The most competitive baseline was Custom Diffusion, achieving a similarity score of 0.729, yet still lower than the 0.753 score of PVA. Moreover, Custom Diffusion also scored worse than PVA in FID and KID, showing that it has inferior image quality than PVA. In conclusion, PVA outperformed all baselines on identity-preserving face inpainting on CelebAHQ-IDI-5.

4.2.2 Language Controllability

As MyStyle and Paint by Example do not support language control, we compared PVA to Textual Inversion and Custom Diffusion on language controllability.

Qualitative results. We present language-controlled inpainting examples in Fig. 3. Results showed that our method could control the inpainted content while preserving the identity. In comparison, Textual Inversion and Custom Diffusion lost the target identity severely while editing the image.

[2] All experiments and data processing activities are conducted at Carnegie Mellon University (CMU).

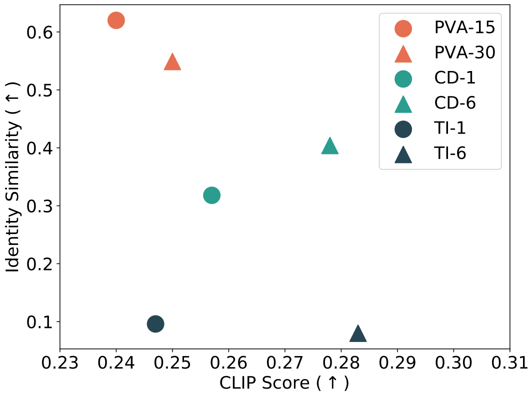

Figure 5. Identity similarity v.s. CLIP score for inpainting with language control. The suffix "-x" is the classifier-free guidance strength.

Quantitative results. The CLIP score and identity similarity for inpainted images are shown in Fig. 5. We had two observations. First, all methods had lower identity similarity compared to the inpainting-only task. PVA, Custom Diffusion, and Textual Inversion scored around 0.6, 0.4, and 0.1, and the inpainting-only ones were 0.753, 0.729, and 0.638. It indicated that using language control, there was a trade-off between identity similarity and prompt similarity. Second, PVA demonstrated the best trade-off efficiency. The identity similarity of PVA is significantly higher than the baselines. Meanwhile, the CLIP score of PVA ranged between 0.24 and 0.25, which overlapped with CD-1 and TI-1. The most competitive baseline, Custom Diffusion, scored around 0.4 for identity similarity. We refer readers to the example shown in Fig. 4 Row 1 Col 3, which has a similar score of 0.43. One could easily recognize that image as a different person than the groundtruth. In conclusion, PVA preserves the identity significantly better while having language controllability matched to the baselines with guidance scale 1.

4.2.3 Ablation Study

We ablated three factors that influence the identity similarity and image quality of PVA.

1. Ablation on classifier-free guidance. The effect of classifier-free guidance is shown in Fig. 6. We observed that a larger guidance scale increased the identity similarity, but also negatively affected the image quality. This trade-off was present in all personalization techniques.

2. Ablation on finetuning. See the comparison between PVA and PVA† (without finetuning) in Fig. 6. It was observed that PVA without finetuning performed closely to Textual Inversion. With only 40 steps of finetuning, the identity similarity of PVA was significantly improved, outperforming the baselines that need over 1K optimization steps. Therefore, it was supported that the feed-forward component of PVA learned a good initialization for finetuning.

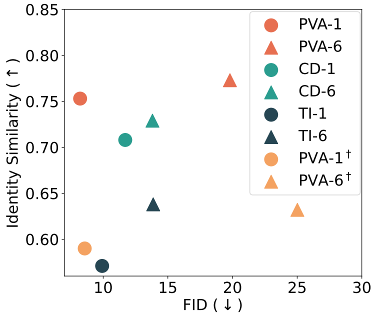

Figure 6. ID similarity and FID comparison of PVA and baselines. † indicates PVA without finetuning in inference.

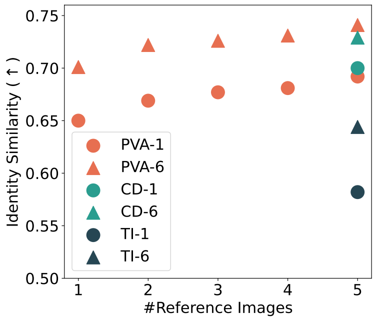

Figure 7. Ablation study on the number of reference images.

3. Ablation on the number of reference images. We trained PVA with different numbers of reference images and presented the results in Fig. 7. As shown in the figure, more reference images resulted in better identity similarity. Moreover, just one reference image already achieved a good result. We speculate that the pretrained diffusion model has learned to disentangle the person's identity from other attributes.

5. Limitation

We observe that the language controllability of PVA is sacrificed to a certain degree for better identity similarity. Though maintaining the identity similarity is indeed the top priority of this work, how to preserve the language control capability better remains an open question.

6. Conclusion

We address the problem of identity-preserving and language-controllable face inpainting. Our solution is the PVA pathway for diffusion models, which consists of an identity encoder and PVA modules. We also propose a new dataset, CelebAHQ-IDI, for benchmarking identity-preserving face inpainting. Results show that our method achieves the best identity similarity and image quality in face inpainting while being significantly faster than baselines. In terms of language-controllability, PVA achieved a similar degree of controllability while preserving the identity better.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019. 3, 4

[2] Xiang An, Jiankang Deng, Jia Guo, Ziyong Feng, XuHan Zhu, Jing Yang, and Tongliang Liu. Killing two birds with one stone: Efficient and robust training of face recognition cnns by partial fc. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4042–4051, June 2022. 6

[3] Xiang An, Xuhan Zhu, Yang Xiao, Lan Wu, Ming Zhang, Yuan Gao, Bin Qin, Debing Zhang, and Fu Ying. Partial fc: Training 10 million identities on a single machine. In *Arxiv 2010.05222*, 2020. 6

[4] Anonymous. DPM-solver++: Fast solver for guided sampling of diffusion probabilistic models. In *Submitted to The Eleventh International Conference on Learning Representations*, 2023. under review. 3

[5] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B. Tenenbaum, William T. Freeman, and Antonio Torralba. Visualizing and understanding generative adversarial networks. In *ICLR*, 2019. 3

[6] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *International Conference on Learning Representations*. 2, 6

[7] Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. Semi-parametric neural image synthesis. In *NIPS*, 2022. 4

[8] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022. 4

[9] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, pages 8789–8797, 2018. 3

[10] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019. 6

[11] Jiankang Deng, Jia Guo, Debing Zhang, Yafeng Deng, Xiangju Lu, and Song Shi. Lightweight face recognition challenge. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 6

[12] Brian Dolhansky and Cristian Canton Ferrer. Eye in-painting with exemplar generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7902–7911, 2018. 2, 3

[13] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2, 3, 4, 5, 6

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, pages 2672–2680, 2014. 2

[15] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 2, 7

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, pages 6626–6637, 2017. 2, 6

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NIPS*, 33:6840–6851, 2020. 3

[18] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NIPSW*, 2021. 4

[19] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pages 1501–1510, 2017. 4

[20] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 5967–5976, 2017. 3

[21] Yuming Jiang, Ziqi Huang, Xingang Pan, Chen Change Loy, and Ziwei Liu. Talk-to-edit: Fine-grained facial editing via dialog. In *ICCV*, 2021. 5, 6

[22] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711. Springer, 2016. 4

[23] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *arXiv preprint arXiv:2006.06676*, 2020. 6

[24] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 2

[25] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pages 8110–8119, 2020. 2, 4

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[27] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 3

[28] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. *arXiv preprint arXiv:2212.04488*, 2022. 2, 3, 4, 5, 6

[29] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*, pages 5549–5558, 2020. 2

[30] Kunjian Li and Qijun Zhao. If-gan: Generative adversarial network for identity preserving facial image inpainting and frontalization. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 45–52, 2020. 3

[31] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, December 2015. 6

[32] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

[33] Wanglong Lu, Hanli Zhao, Xianta Jiang, Xiaogang Jin, Min Wang, Jiankai Lyu, and Kaijie Shi. Diverse facial inpainting guided by exemplars. *arXiv preprint arXiv:2202.06358*, 2022. 3

[34] Alex 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. *arXiv preprint arXiv:2112.10741*, 2021. 3

[35] Yotam Nitzan, Kfir Aberman, Qiurui He, Orly Liba, Michal Yarom, Yossi Gandelsman, Inbar Mosseri, Yael Pritch, and Daniel Cohen-Or. Mystyle: A personalized generative prior. *arXiv preprint arXiv:2203.17272*, 2022. 1, 2, 3, 6

[36] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Gaugan: semantic image synthesis with spatially adaptive normalization. In *ACM SIGGRAPH 2019 Real-Time Live!*, pages 1–1. 2019. 3

[37] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *CVPR*, 2022. 6

[38] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8024–8035. Curran Associates, Inc., 2019. 7

[39] 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. 3

[40] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 2

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

[42] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2287–2296, 2021. 3, 4

[43] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *TOG*, 42(1):1–13, 2022. 3

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 3, 6

[45] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2

[46] 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 preprint arXiv:2208.12242*, 2022. 2, 3, 4, 5, 6

[47] Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected gans converge faster. volume 34, pages 17480–17492, 2021. 3

[48] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *SIGGRAPH*, pages 1–10, 2022. 3

[49] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *PAMI*, 44(4):2004–2018, 2022. 3

[50] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *CVPR*, pages 1532–1540, 2021. 3

[51] Shelly Sheynin, Oron Ashual, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, and Yaniv Taigman. Knn-diffusion: Image generation via large-scale retrieval. *arXiv preprint arXiv:2204.02849*, 2022. 4

[52] Jie Shi, Chenfei Wu, Jian Liang, Xiang Liu, and Nan Duan. Divae: Photorealistic images synthesis with denoising diffusion decoder. *arXiv preprint arXiv:2206.00386*, 2022. 3

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

[54] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *NIPS*, 32, 2019. 3

[55] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2020. 3

[56] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016. 6

[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 3, 4

[58] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018. 6

[59] Jianjin Xu, Zheyang Xiong, and Xiaolin Hu. Frame difference-based temporal loss for video stylization. *arXiv preprint arXiv:2102.05822*, 2021. 4

[60] Jianjin Xu, Zhaoxiang Zhang, and Xiaolin Hu. Extracting semantic knowledge from gans with unsupervised learning. *arXiv preprint arXiv:2211.16710*, 2022. 3

[61] Jianjin Xu and Changxi Zheng. Linear semantics in generative adversarial networks. In *CVPR*, pages 9351–9360, 2021. 3

[62] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. *arXiv preprint arXiv:2211.13227*, 2022. 1, 4, 6

[63] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 4

[64] Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. *arXiv preprint arXiv:2204.13902*, 2022. 3

[65] Yajie Zhao, Weikai Chen, Jun Xing, Xiaoming Li, Zach Bessinger, Fuchang Liu, Wangmeng Zuo, and Ruigang Yang. Identity preserving face completion for large ocular region occlusion. *arXiv preprint arXiv:1807.08772*, 2018. 3

[66] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. Indomain gan inversion for real image editing. In *Proceedings of ECCV (ECCV)*, 2020. 3, 4

[67] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *CVPR*, pages 2223–2232, 2017. 3