Face2Diffusion for Fast and Editable Face Personalization

Kaede Shiohara Toshihiko Yamasaki
The University of Tokyo
{shiohara, yamasaki}@cvm.t.u-tokyo.ac.jp

Abstract

Face personalization aims to insert specific faces, taken from images, into pretrained text-to-image diffusion models. However, it is still challenging for previous methods to preserve both the identity similarity and editability due to overfitting to training samples. In this paper, we propose Face2Diffusion (F2D) for high-editability face personalization. The core idea behind F2D is that removing identity-irrelevant information from the training pipeline prevents the overfitting problem and improves editability of encoded faces. F2D consists of the following three novel components: 1) Multi-scale identity encoder provides well-disentangled identity features while keeping the benefits of multi-scale information, which improves the diversity of camera poses. 2) Expression guidance disentangles face expressions from identities and improves the controllability of face expressions. 3) Class-guided denoising regularization encourages models to learn how faces should be denoised, which boosts the text-alignment of backgrounds. Extensive experiments on the FaceForensics++ dataset and diverse prompts demonstrate our method greatly improves the trade-off between the identity- and text-fidelity compared to previous state-of-the-art methods. Code is available at https://github.com/mapooon/Face2Diffusion.

1. Introduction

Text-to-image (T2I) diffusion models trained on web-scale data such as GLIDE [30], DALL-E2 [35], Imagen [40], and StableDiffusion [36] have shown the impressive image generation ability aligned with a wide range of textual conditions, outperforming previous generative models (*e.g.*, [1, 19, 31]). Therefore, the next challenge of the community is to explore how to insert specific concepts (*e.g.*, someone's face) from images that T2I models do not know, leading to the pioneering personalization methods such as TextualInversion [11] and DreamBooth [39]. However, they require heavy fine-tuning that takes several tens of minutes per concept. To reduce the training costs and improve the editability, some studies [24, 47] propose efficient training strategies that optimize mainly cross-attention layers in denoising UNet. Others use pretrained image encoders (*e.g.*, CLIP [33]) to represent new concepts as textual embeddings [4, 12, 54]. Most recently, some studies [5, 50, 51] achieve tuning-free personalization via large-scale pretraining. For example, ELITE [50] proposes a two-stage training strategy that optimizes its mapping networks and cross-attention layers on the OpenImages dataset [25].

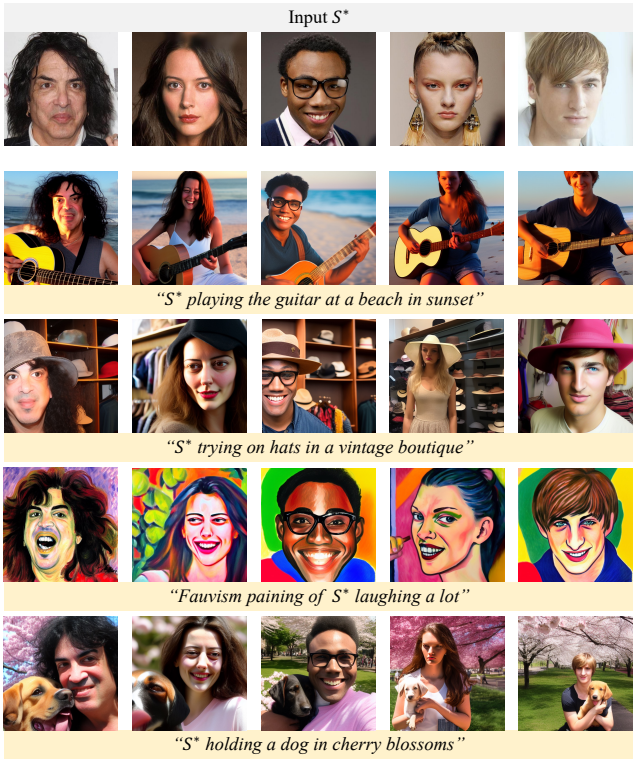


Figure 1. **Our Results.** Face2Diffusion satisfies challenging text prompts that include multiple conditions while preserving input face identities without individual test-time tuning.

In particular, personalization for faces, which we call "face personalization" in this paper, draws attention from the community because of its potential applications, *e.g.*, content creation. CelebBasis [54] represents face embeddings on a basis constructed by celebrity names known to T2I models. FastComposer [51] proposes localizing cross-

Figure 2. **Typical overfitting to input data.** The original StableDiffusion [36] is capable of generating text-aligned images with plausible backgrounds, camera poses, and diverse face expressions. Nevertheless, a previous method [24] fails in disentangling these identity-irrelevant information from the input sample.

attention to solve the identity blending problem where T2I models mix-up multiple people in one image. DreamIdentity [5] improves the identity similarity of generated images by utilizing multi-scale features from a pretrained face recognition model [6]. However, despite the great efforts in personalization of T2I models, we found that it is still challenging to generate well-aligned images with a wide range of text prompts while preserving the subjects' identities because of overfitting to training images.

In this paper, we propose Face2Diffusion (F2D) for more editable face personalization. The core idea behind our F2D is that properly removing identity-irrelevant information from the training pipeline helps the model learn editable face personalization. We analyze the overfitting problem on a previous method [24] and classify the typical overfitting into three types: camera poses, face expressions, and backgrounds, and present the following solutions for the problem: 1) Multi-scale identity (MSID) encoder is presented to provide purer identity features disentangled from camera pose information while keeping the benefits of multi-scale feature extraction. 2) Expression guidance disentangles face expressions from identity features, making it possible to control face expressions by text prompts and reference images. These techniques enable F2D to disentangle backgrounds, camera poses, and face expressions, performing high-fidelity face personalization as shown in Fig. 1. 3) Class-guided denoising regularization (CGDR) constrains backgrounds of injected faces to be denoised in the same manner as its super-class word (*i.e.*, *"a person"* for face personalization) to improve the text-fidelity on backgrounds.

We compare our F2D with the nine previous state-of-the-art methods [5, 11, 12, 24, 39, 47, 50, 51, 54] on 100 faces from the FaceForensics++ [38] dataset with 40 diverse text prompts we collected. Our method ranks in the top-3 in five of the six metrics in terms of the identity-fidelity [21, 27, 41] and text-fidelity [10, 14, 55], and outperforms previous methods in the harmonic and geometric means of the six metrics, reflecting the superiority of F2D in the total quality of face personalization.

2. Related Work

2.1. Text-to-Image Diffusion Models.

Text-to-image (T2I) diffusion models [30, 35, 36, 40] are a kind of generative models based on diffusion models [17, 44, 45] that generate images by gradual denoising steps in the image space [30, 35, 40] or latent one [36]. Specifically, T2I models generate images from encoded texts by a large language model [34] or vision-language model [33]. ADM [8] guides denoised images to be desired classes using image classifiers. Classifier-free guidance [16] achieves to guide diffusion models without any classifiers, overcoming the limitation of guided diffusion that requires noise-robust image classifiers.

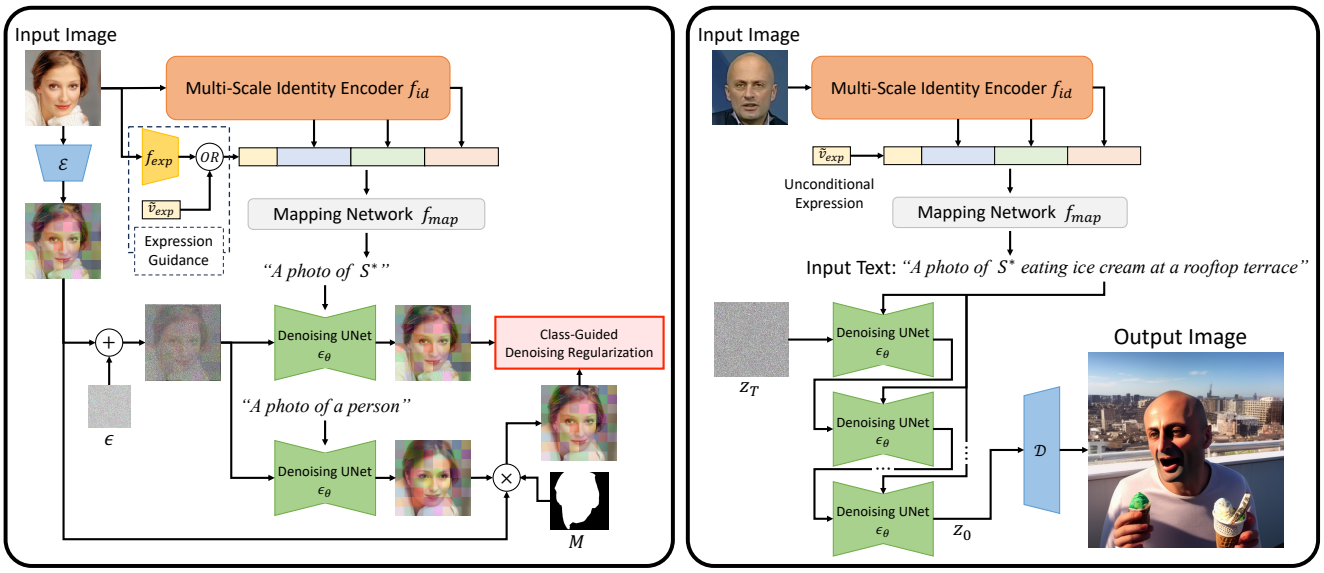
2.2. Personalization of T2I models

The high-fidelity generation ability of T2I models drives the community to explore personalization methods, or how to insert unseen concept into T2I models. Here, we review existing works by classifying them into three types:

Direct optimization methods. Early attempts optimize a learnable word embedding for a new concept [11] or fine-tune the whole T2I model [39] by learning to reconstruct input images from a new concept word embedding. CustomDiffusion [24] identifies that cross-attention layers of the denoising UNet are more influential to invert new concepts and proposes a lightweight fine-tuning method. Perfusion [47] replaces the key value of a new concept embedding with that of its super-category's one in the cross-attention layers to mitigate overfitting to training samples.

Encoder-based optimization methods. To improve editability and training efficiency, some studies [4, 12, 54] propose to utilize pretrained image encoders such as CLIP [33] to represent new concepts as textual embeddings. E4T [12] proposes a two-stage personalization method that includes pretraining of a word embedder and weight offsets of attention layers on domain-oriented datasets [18, 19, 53] and few-step (5-15) fine-tuning per concept. DisenBooth [4] learns to disentangle a new concept from irrelevant information by reconstructing an concept image from another image.

Optimization-free methods. Motivated by the impressive results of the encoder-based methods, recent studies [5, 28, 43, 50, 51] focus on feed-forward inversion without fine-tuning per concept. ELITE [50] introduces a two-stage pretraining framework that optimizes global and local mapping networks and cross-attention layers of its T2I model. FastComposer [51] replaces the inserted embeddings of input text prompts with their super-class embeddings in early denoising steps during inference to improve the text-fidelity. DreamIdentity [5] leverages generated images by its T2I model to train itself on various image-text pairs without manual collection.



(a) Training (b) Inference

Figure 3. **Overview of Face2Diffusion.** (a) During training, we input a face image into our novel multi-scale identity encoder f_{id} and an off-the-shell 3D face reconstruction model f_{exp} to extract identity and expression features, respectively. The concatenated feature is projected into the text space as a word embedding S^* by a mapping network f_{map}. The input image is also encoded by VAE's encoder $\mathcal{E}$ and then a Gaussian noise ϵ is added to it. We constrain the denoised latent feature map to be the original one in the foreground and to be a class-guided denoised result in the background. (b) During inference, the expression feature is replaced with an unconditional vector $\tilde{v}_{exp}$ to diversify face expressions of generated images. After injecting the face embedding S^* into an input text, the original denoising loop of StableDiffusion is performed to generate an image conditioned by the input face identity and text.

3. Face2Diffusion

Our goal is to represent input faces as face emeddings S^* in the text space of CLIP [33] to generate target subjects conditioned by text prompts on StableDiffusion [36]. The important observation behind our work is that previous methods suffer from three types of overfitting: backgrounds, camera poses, and face expressions. Fig. 2 shows a failure case of a previous method [24]. It can be observed that the method tends to generate similar backgrounds, camera poses, and face expressions due to overfitting to the input sample despite the capability of the original StableDiffusion generating suitable and diverse scenes. Motivated by this overfitting problem, we propose Face2Diffusion (F2D) for more editable face personalization. We visualize the overview of F2D in Fig. 3. F2D consists of three important components to solve the overfitting promblem: In Sec. 3.2, we introduce multi-scale identity (MSID) encoder to disentangle camera poses from face embeddings by removing identity-irrelevant information from a face recognition model [6]. In Sec. 3.3, we present expression guidance that disentangles face expressions from face embeddings to diversify face expressions aligned with texts. In Sec. 3.4, we present class-guided denoising regularization (CGDR) that forces the denoising manner of face embeddings to follow that of its super-class (*i.e.*, *"a person"*) in the background.

3.1. Preliminary

StableDiffusion. We adopt StableDiffusion [36] (SD) as our base T2I model. SD first learns perceptional image compression by VAE [22]. Then, SD learns conditional image generation via latent diffusion. Given a timestep t, noisy latent feature z_t, and text feature $\tau(p)$ where τ and p are the CLIP text encoder [33] and a text prompt, respectively, a denoising UNet [37] ϵ_θ predicts the added noise ϵ. The training objective is formulated as:

$$\mathcal{L}_{ldm} = \|\epsilon - \epsilon_\theta(z_t, t, \tau(p))\|_2^2. \quad (1)$$

Leveraging large-scale image-text pair datasets (*e.g.*, LAION-5B [42]), SD learns semantic relationships between images and texts.

Encoder-based face personalization. Previous works [5, 12, 51, 54] show that pretrained image encoders are usefull to represent face identities as textual embeddings S^*. For a face encoder f_{id} and a mapping network f_{map}, S^* for an input image x is computed as follows:

$$S^* = f_{map}(f_{id}(x)). \quad (2)$$

Then, S^* is used as a tokenized word to generate identity-injected images as in Fig. 3 (b). f_{map} is optimized using the reconstruction loss Eq. 1 per subject [54] or face datasets for instant encoding without test-time tuning [5, 51].

3.2. Multi-Scale Identity Encoder

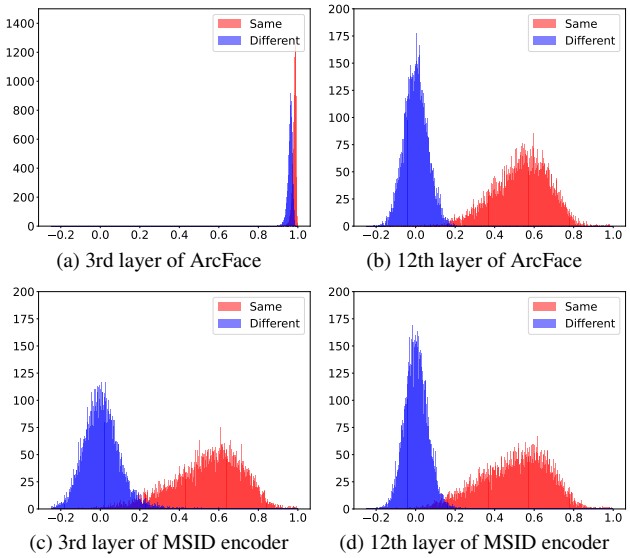
We start by considering how to encode faces in the CLIP text space. Previous works [5, 54] reveal that face recognition models [6, 48] provide well-aligned identity representations for face personalization. Specifically, DreamIdentity [5] proposes to extract multi-scale features from shallow layers to deep ones to improve the identity similarity.

However, such a straight-forward multi-scale method suffers from overfitting to input samples. This is because features from shallow layers include a lot of low level information (*e.g.*, camera pose, expression, and background) that is irrelevant to identity [50]. This accidentally makes models dependent on shallow features to minimize the reconstruction loss Eq. 1, which results in a low editability. Figs. 4(a) and (b) visualize identity similarity distributions of ViT [9]-based ArcFace [6] on a shallow (3rd) layer and the deepest (12th) one, respectively. The horizontal and vertical axes represent the identity similarity and frequency, respectively. We uniformly sample an {Anchor, Positive (Same), Negative (Different)} triplet for each identity from the VGGFace2 [2] dataset. It can be observed that the shallow features are un-aligned and insufficient to discriminate identities; AUC on shallow features is only 93.99% although that on the deepest is 99.97%. This result implies that shallow features contain specific information of input images, which causes overfitting in face personalization.

To solve this problem, we present a multi-scale identity (MSID) encoder for face personalization. In the pre-training of a face recognition model ArcFace [6], we encourage not only the deepest layer but also shallower ones to discriminate identities. We denote the concatenated multi-scale identity vector as $v_{id} = [v_{id}^1, v_{id}^2, ..., v_{id}^D]$ where v_{id}^i is the i-th level feature. Given an image and its class label y, we formulate our multi-scale loss $\mathcal{L}_m$ as the original ArcFace loss [6]:

$$\mathcal{L}_m = -\log \frac{e^{s\cos(W_y^T v_{id}+m)}}{e^{s\cos(W_y^T v_{id}+m)} + \sum_{k=1,k\neq y}^{K} e^{s\cos W_k^T v_{id}}}, \tag{3}$$

where W_k is the k-th column of the weight matrix W. m and s are the margin and scale of ArcFace [6], respectively. As shown in Figs. 4(c) and (d), the multi-scale features by our encoder are well-aligned and discriminative both on shallow and deep layers; AUC on shallow features is improved from 93.99% to 99.46%. The disentangled multi-scale identity features prevent overfitting while providing strong identity information for face personalization. Differently from a previous work in image classification tasks [32] that targets only two deepest layers, we train a wide range of levels because we focus on the disentanglement of multi-scale features from shallow to deep layers for face personalization, rather than the discriminability of the deepest feature for classification.



(a) 3rd layer of ArcFace (b) 12th layer of ArcFace

(c) 3rd layer of MSID encoder (d) 12th layer of MSID encoder

Figure 4. **Identity similarity distributions of ArcFace and MSID encoder.** Shallow features of ArcFace (a) are un-aligned and low discriminability although deep features (b) are well aligned. Our MSID encoder succeeds in removing identity-irrelevant information both from the shallow layer (c) and deep one (d) to prevent overfitting.

3.3. Expression Guidance

Although our MSID encoder greatly improves the text-fidelity while preserving the identity similarity, we found that it is still difficult to control face expressions due to overfitting to input samples. To tackle this problem, we propose expression guidance to disentangle face expressions from identities. We utilize a 3D face reconstruction model [7], denoted as f_{exp}, to extract expression features.

For an input face image x, we simply inject expression information by concatenating $v_{exp} = f_{exp}(x)$ and $v_{id} = f_{id}(x)$ before the mapping network f_{map}. The expression-guided feature vector S^* is computed as:

$$S^* = \begin{cases} f_{map}([v_{id}, v_{exp}]) & (p = 0.8), \\ f_{map}([v_{id}, \tilde{v}_{exp}]) & (p = 0.2), \end{cases} \tag{4}$$

where we drop v_{exp} with a probability of 0.2 and input an alternative learnable vector $\tilde{v}_{exp}$ for unconditional generation. After training, we generate target subjects aligned with diverse expression-related prompts using the learned unconditional vector $\tilde{v}_{exp}$. Accessorily, F2D can perform conditional generation with expression references v_{exp}. Please see the supplementary material for more details.

3.4. Class-Guided Denoising Regularization

Overfitting to training samples sometimes prevent generated images from aligned with text conditions in the background. Delayed subject conditioning (DSC) [51] tackles this problem by replacing the identifier S^* with its super-

Figure 5. **Qualitative comparison with previous methods.** Our method generates authentic images aligned with input texts and identities in challenging scenes whereas previous methods compromise an either. Enlarged images are included in the supplementary materials.

class word (*e.g.*, *"a person"*) in early denoising steps and switching the super-class word with S^* from the middle of the denoising loop during inference. However, because some of the face identities emerge in early denoising steps, DSC degrades the identity similarity of generated images to input faces. To disentangle background information from face embeddings without identity degradation, we propose class-guided denoising regularization (CGDR). The important observation behind CGDR is that the class word *"a person"* is well disentangled from specific backgrounds as various ones are generated in Fig. 2. CGDR forces the denoising manner of face embeddings to follow that of the class prompt in the background. First, we input an identity-injected prompt p = *"A photo of S^*"* and class prompt p_c = *"A photo of a person"* into our model to predict the noise for the same noisy latent z_t:

$$\hat{\epsilon} = \epsilon_\theta(z_t, t, \tau(p)), \tag{5}$$

$$\hat{\epsilon_c} = \epsilon_\theta(z_t, t, \tau(p_c)). \tag{6}$$

Then, we constrain the original predicted noise $\hat{\epsilon}$ to be $\hat{\epsilon}_c$ in the background and to be the actually added noise ϵ in the foreground as follows:

$$\mathcal{L} = \|\hat{\epsilon} - \{\epsilon \odot M + \hat{\epsilon}_c \odot (1 - M)\}\|_2^2, \tag{7}$$

where M is a segmentation mask by a face-parsing model [52] to divide the identity region from others and $\odot$ represents pixel-wise multiplication. By this single objective function, F2D learns to encode face identities, strongly mitigating overfitting to the backgrounds of input images.

Notably, our training strategy keeps the weights of T2I models original because we optimize only f_{map} during training of F2D. Therefore, our method fundamentally does not struggle with the language drift [26] and catastrophic forgetting [23] encountered by some of previous methods (*e.g.*, [4, 39, 43, 51]) that modify the original weights.

4. Experiments

4.1. Implementation Detail

Multi-scale identity encoder. We pretrain ViT [9] equipped with 12 transformer encoder layers on the MS1M [13] dataset for 30 epochs. The batch size and learn-

Method	Identity-Fidelity (↑)			Text-Fidelity (↑)			Identity×Text (↑)	
	AdaFace	SphereFace	FaceNet	CLIP	dCLIP	SigLIP	hMean	gMean
Subject-driven								
TextualInversion [11]	0.1654	0.2518	0.3197	0.1877	0.1075	0.1283	0.0320	0.0575
DreamBooth [39]	0.3482(2)	0.4084	0.4774	0.2351	0.1480	0.3452(3)	0.0689	0.1116
CustomDiffusion [24]	0.4537(1)	0.5624(1)	0.6458(1)	0.2023	0.1490	0.2182	0.1078	0.1700(3)
Perfusion [47]	0.0925	0.1478	0.1887	0.2490	0.1686	0.3433	0.0342	0.0575
E4T [12]	0.0433	0.1190	0.1725	0.2510(2)	0.1784	0.2671	0.0370	0.0589
CelebBasis [54]	0.1601	0.2724	0.3762	0.2563(1)	0.1847(3)	0.3624(2)	0.1138(3)	0.1542
Subject-agnostic								
FastComposer [51]	0.1736	0.3271	0.4725	0.2495(3)	0.2149(1)	0.3168	0.1655(2)	0.2120(2)
ELITE [50]	0.0924	0.1925	0.3232	0.1755	0.1250	0.0671	0.0308	0.0624
DreamIdentity* [5]	0.2847	0.4252(2)	0.5523(2)	0.1924	0.1463	0.1539	0.0849	0.1326
Face2Diffusion (Ours)	0.3143(3)	0.4215(3)	0.5313(3)	0.2486	0.2020(2)	0.3856(1)	0.1749(1)	0.2252(1)

Table 1. **Comparison with previous methods.** The top-3 values are ranked in brackets. * denotes our own implementation. CustomDiffusion [24] suffers from overfitting to input faces while FastComposer [51] cannot represent identities sufficiently, resulting in lower Identity×Text scores. Our model ranks in the top-3 in five of the six metrics and achieves the best Identity×Text scores.

ing rate are set to 1024 and 10^{-3}, respectively. We select a depth set of $\{3, 6, 9, 12\}$ for multi-scale feature extraction.

Face2Diffusion. We adopt the two-word embedding method [5, 54] where $S^* = [S_1^*, S_2^*]$ is predicted by two independent mapping networks f_{map}^1 and f_{map}^2. Each mapping network f_{map}^i consists of two-layer MLP where each layer has a linear layer, dropout [46], and leakyReLU [29], followed by an additional linear layer that projects the features into the text space. We train our model on FFHQ [19] for 100K iterations. Note that only f_{map}^1 and f_{map}^2 are updated and other networks are frozen. Only horizontal flip is used for data augmentation. The batch size and learning rate are set to 32 and 10^{-5}, respectively.

4.2. Setup

Evaluation data. We adopt the FaceForensics++ [38] dataset. We randomly select 100 videos and then extract a frame from each video. We align and crop images following the FFHQ [19] dataset. We also collect 40 human-centric prompts including diverse scenes such as job, activity, expression, and location. The prompt list is included in the supplementary material. We generate all combinations of face images and prompts, *i.e.*, 4,000 scenes. Following previous works [4, 39], we generate four images for each {image, prompt} set for robust evaluation. We report averaged scores over all samples.

Compared methods. We compare our method with nine state-of-the-art personalization methods including six subject-driven methods, TextualInversion [11], DreamBooth [39], CustomDiffusion [24], Perfusion [47], E4T [12], and CelebBasis [54], and three subject-agnostic methods, FastComposer [51], ELITE [50], and DreamIdentity [5]. For all the methods including our Face2Diffusion, we use StableDiffusion-v1.4 [36], Euler ancestral discrete scheduler [20] with 30 denoising steps, and classifier-free guidance [16] with a scale of 7.0 for fair comparison. Please refer to the supplementary material for more details.

	CustomDiffusion	CelebBasis	FastComposer	Ours
FID (↓)	86.18	69.87	77.62	**69.33**
#Params (↓)	5.71×10^7	**1024**	8.88×10^8	1.20×10^7
Time (↓)	140 sec	220 sec	0.026 sec	**0.006 sec**

Table 2. **Comparison on FID and the computation costs.**

Metrics. We consider three evaluation metrics following:

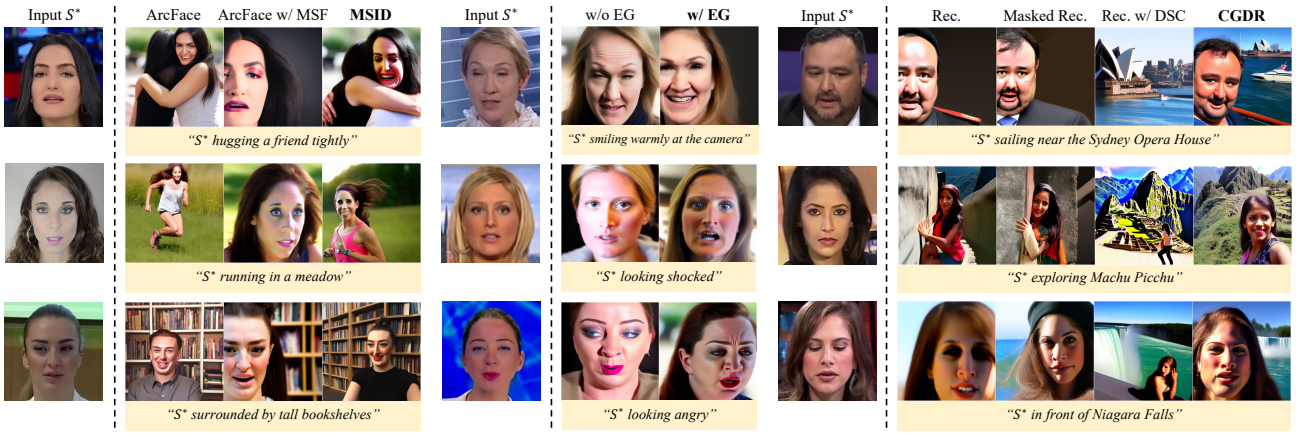
1) Identity-fidelity represents how a generated image is similar to an input image in terms of face identity. We adopt AdaFace [21], SphereFace [27], and FaceNet [41], and compute the cosine similarity between extracted features from input images and generated images. We assign 0.0 to images where no face is detected.

2) Text-fidelity expresses how a generated image is aligned with a text prompt except the face identity. We adopt CLIP score [14], directional CLIP (dCLIP) score [10], and SigLIP [55] to compute image-text similarity. We use a reference prompt "*A photo of a person*" and a reference image generated by "*A photo of S^**" for dCLIP.

3) Because the goal of face personalization is to satisfy "injecting a face identity" and "being aligned with a text prompt" simultaneously, it is important to evaluate these performance simultaneously on each image. To achieve this, we introduce Identity×Text score that reflects the total quality of face personalization. We compute the harmonic mean (hMean) and geometric one (gMean) of the six metrics above.

4.3. Comparison with Previous Methods

Qualitative result. We show the generated samples by personalization methods in Fig. 5. Some of the subject-driven methods such as DreamBooth [39] and CustomDiffusion [24] that require per-face otimization generate highly similar faces to input ones. However, they struggle with



(a) Effect of MSID encoder. (b) Effect of expression guidance. (c) Effect of CGDR.

Figure 6. **Comprehensive ablation study.** We omit the common prefix *"A photo of"* due to the space limitation. (a) Our MSID encoder disentangles camera poses while keeping the identity similarity. (b) Our expression guidance mitigates overfitting to input face expressions. (c) Our CGDR improves the text-fidelity mainly on backgrounds.

Encoder	AdaFace	CLIP	hMean
ArcFace [6]	0.2421	**0.2758**	0.2135
ArcFace w/ Multi-Scale Feat. [5]	**0.3264**	0.2069	0.1549
MSID Encoder (Ours)	0.3143	0.2486	**0.2252**

Table 3. **Effect of MSID encoder.** The original ArcFace [6] remains at a low identity similarity. The multi-scale features [5] trades the text-fidelity for the identity-fidelity. Our method improves the text-fidelity while preserving the identity-fidelity.

overfitting to identity-irrelevant attributes of input images, which results in ignoring some text conditions, *e.g.*, *"in a vintage boutique"*. On the other hand, CelebBasis [54] and FastComposer [51], that are specialized for face personalization, are better-aligned with input texts but comprise the identity-fidelity due to their strong regularizations. Also, we observe that FastComposer tends to produce inconsistent images between foregrounds and backgrounds, as reported in a previous study [54]. This is because DSC adopted in the method switches the input prompts in the middle of the denoising loop. Compared to these methods, our Face2Diffusion consistently satisfies both input face identities and text conditions, which demonstrates the effectiveness of our approach.

Quantitative result. Then, we evaluate the methods quantitatively in Table 1. CustomDiffusion [24] consistently achieves the best results in the identity metrics (AdaFace, SphereFace, and FaceNet); however, it faces the poor editablity due to overfitting to input images. In contrast, CelebBasis [54] and FastComposer [51] obtain better editability compared to others but generated faces have less similarity to input faces. As a result, these methods remain at low Identity×Text scores. Our method ranks in the top-3 in five of the six metrics and outperforms the state-of-the-art methods in Identity×Text scores. This result clearly indicates that our model improves the trade-off between identity- and text-fidelity.

Method	AdaFace	CLIP	hMean
w/o Expression Guidance	**0.3338**	0.2315	0.2032
w/ Expression Guidance (Ours)	0.3143	**0.2486**	**0.2252**

Table 4. **Effect of expression guidance.**

We also report FID [15], the numbers of trained parameters, and encoding time on a single NVIDIA A100 GPU in Table 2 for more comparisons of the top models in Table 1. Because there is no target distribution (real images) for the test prompts, we compute FID between the generated images and human-centric images whose captions include a human-related word (*"person"*,*"man"*, or *"woman"*) from the CC12M [3] dataset. We observe that our method achieves the best FID, demonstrating its higher-fidelity and diversity than the previous state-of-the-arts. For the computation costs, although CelebBasis has less trainable parameters than the others, it is much slower than our method because CelebBasis requires optimization per subject. Our method achieves the fastest encoding time.

4.4. Ablation Study and Analysis

Effect of MSID encoder. First, we compare our MSID encoder with the original ArcFace [6] and its multi-scale features [5] in Fig. 6(b) and Table 3. We observe that the original ArcFace [6] cannot represent subjects' identities sufficiently (0.2421 in AdaFace). Also, it sometimes mistakes subjects' gender because of the excessively abstracted identity features from the deepest layer, as shown in the third row of the figure. On the contrary, multi-scale features [5] cannot disentangle camera poses of input images, resulting in a low editability (0.2069 in CLIP). Our MSID encoder successfully improves the editability (from 0.2069 to

Method	AdaFace	CLIP	hMean
Reconstruction	<u>0.3133</u>	0.2053	0.1411
Masked Reconstruction	0.3012	0.1990	0.1282
Reconstruction w/ DSC [51]	0.1651	**0.2851**	<u>0.1596</u>
CGDR (Ours)	**0.3143**	<u>0.2486</u>	**0.2252**

Table 5. **Effect of CGDR.** The best and the second-best values are in **bold** and <u>underlined</u>, respectively. The original reconstruction training and simple masking strategy result in the low editability. DSC trades the identity-fidelity for the text-fidelity. Our method balances the trade-off and achieves the best Identity×Text score.

0.2486 in CLIP) by preventing overfitting to camera poses while maintaining the benefits of leveraging multi-scale features (0.3143 vs. 0.3264 in AdaFace) and achieves the best Identity×Text score (0.2252 in hMean).

Effect of expression guidance. We evaluate our method with and without expression guidance in Fig. 6(c) and Table 4. We can see that our expression guidance improves the text-fidelity (from 0.2315 to 0.2486 in CLIP) especially on expression-related prompts (*e.g.*, *"S^* looking shocked"*), as shown in the figure. We also observe that the identity-fidelity of our model is slightly lower than our model without expression guidance (0.3154 vs. 0.3338 in AdaFace). This is because generated images have diverse face expressions aligned with text prompts, which sometimes degrades the quantitative identity similarity of generated images.

Effect of CGDR. We compare our CGDR with three baselines in Fig. 6(a) and Table 5: "Reconstruction (Rec.)" is our model trained with the original reconstruction loss (Eq. 1) instead of our CGDR (Eq. 7). "Masked Reconstruction (Masked Rec.)" is our model trained with a reconstruction loss computed within facial regions using segmentation masks. Moreover, we compare Delayed Subject Conditioning (DSC) [51] that uses an alternative prompt whose identifier S^* is replaced with its class name (*i.e.*, *a person*) in early denoising steps during inference. It is partially similar to our CGDR from the perspective of leveraging class-guided denoising. We implement this technique on the "Reconstrucion" model and denote it as "Reconstruction w/ DSC (Rec. w/ DSC)". We observe that the original reconstruction training causes overfitting to training samples, which results in a low editability (0.2053 in CLIP). And the masked reconstruction training does not solve the overfitting problem (0.1990 in CLIP). DSC [51] highly improves the text-fidelity (from 0.2053 to 0.2851 in CLIP) but brings a fatal degradation in the identity similarity (from 0.3133 to 0.1651 in AdaFace). This is because the class prompt *"a person"* is sometimes ignored when emphatic words coexist in the input prompts, as shown in Fig. 6(a). Our method improves the text-fidelity from the original reconstruction model (from 0.2053 to 0.2486 in CLIP) and brings a large improvement on the trade-off between identity- and text-fidelity (0.2252 in hMean), which clearly proves the effectiveness of our CGDR.

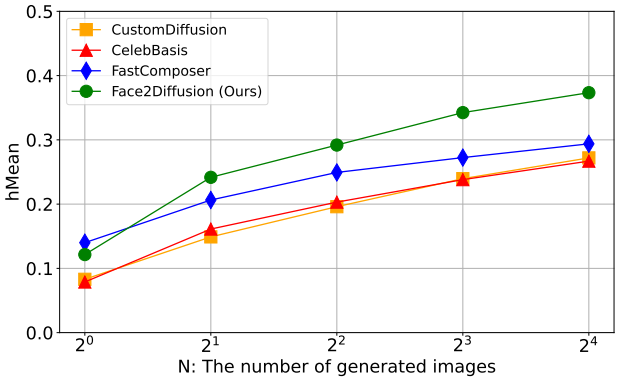


Figure 7. **Upper-bound analysis.**

Upper-bound analysis. Given a practical situation where users can redraw the initial noise several times to produce the most preferable image, it is more important to evaluate only the best image rather than all the generated images. To meet this demand, we conduct an upper-bound analysis where we evaluate only the single image that achieves the best hMean in the N generated images for each {image, prompt} set. We plot the upper-bound hMean scores when $N = \{1, 2, 4, 8, 16\}$ in Fig. 7. We can see that our method surpasses the previous methods by the larger margins as N increases while FastComposer shows a limited improvement. An important observation from Table 1 and Fig. 7 is that FastComposer produces somewhat agreeable results on average, but it is difficult to generate miracle samples. In contrast, our method is more likely to generate very desirable images when the initial noise is redrawn several times.

5. Conclusion

In this paper, we present Face2Diffusion for editable face personalization. Motivated by the overfitting problem on the previous personalization methods, we propose multi-scale identity encoder, expression guidance, and class-guided denoising regularization to disentangle camera poses, face expressions, and backgrounds from face embeddings, respectively. Extensive experiments indicate that our method greatly improves the trade-off between the identity-fidelity and text-fidelity, outperforming previous state-of-the-art methods.

Broader Impact. Face personalization aims to augment human-centric content creation. However, our method can be misused for malicious purposes, *e.g.*, to create fake news, as well as previous personalization methods. To mitigate the risk, we contribute to the research community by releasing the generated images for image forensics (*e.g.*, [49]).

Acknowledgements

This work is supported by JSPS KAKENHI (Grant Numbers JP23KJ0599 and JP22H03640) and Institute for AI and Beyond of The University of Tokyo.

References

[1] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *ICLR*, 2018. 1

[2] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *FG*, 2018. 4

[3] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021. 7

[4] Hong Chen, Yipeng Zhang, Xin Wang, Xuguang Duan, Yuwei Zhou, and Wenwu Zhu. Disenbooth: Disentangled parameter-efficient tuning for subject-driven text-to-image generation. *arXiv preprint arXiv:2305.03374*, 2023. 1, 2, 5, 6

[5] Zhuowei Chen, Shancheng Fang, Wei Liu, Qian He, Mengqi Huang, Yongdong Zhang, and Zhendong Mao. Dreamidentity: Improved editability for efficient face-identity preserved image generation. *arXiv preprint arXiv:2307.00300*, 2023. 1, 2, 3, 4, 6, 7

[6] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive Angular Margin Loss for Deep Face Recognition. In *CVPR*, 2019. 2, 3, 4, 7

[7] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3D Face Reconstruction with Weakly-Supervised Learning: From Single Image to Image Set. In *CVPR Workshop*, 2019. 4

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

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 4, 5

[10] Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *arXiv preprint arXiv:2108.00946*, 2021. 2, 6

[11] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2022. 1, 2, 6

[12] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Designing an encoder for fast personalization of text-to-image models. *arXiv preprint arXiv:2302.12228*, 2023. 1, 2, 3, 6

[13] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*, 2016. 5

[14] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: a reference-free evaluation metric for image captioning. In *EMNLP*, 2021. 2, 6

[15] 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. 2017. 7

[16] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS Workshop*, 2021. 2, 6

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

[18] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018. 2

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 1, 2, 6

[20] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *NeurIPS*, 2022. 6

[21] Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *CVPR*, 2022. 2, 6

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

[23] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *PNAS*, 2017. 5

[24] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023. 1, 2, 3, 6, 7

[25] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020. 1

[26] Jason Lee, Kyunghyun Cho, and Douwe Kiela. Countering language drift via visual grounding. In *EMNLP-IJCNLP*, 2019. 5

[27] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, 2017. 2, 6

[28] Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. *arXiv preprint arXiv:2307.11410*, 2023. 2

[29] Andrew L. Maas. Rectifier nonlinearities improve neural network acoustic models. 2013. 6

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

[31] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*, 2017. 1

[32] Shitala Prasad and Tingting Chai. Multi-scale arc-fusion based feature embedding for small-scale biometrics. *Neural Processing Letters*, 2023. 4

[33] 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 *ICML*, 2021. 1, 2, 3

[34] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2020. 2

[35] 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. 1, 2

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 3, 6

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 3

[38] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. Faceforensics++: Learning to detect manipulated facial images. In *ICCV*, 2019. 2, 6

[39] 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. In *CVPR*, 2023. 1, 2, 5, 6

[40] 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. In *NeurIPS*, 2022. 1, 2

[41] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 2, 6

[42] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022. 3

[43] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*, 2023. 2, 5

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

[45] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2020. 2

[46] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *JMLR*, 2014. 6

[47] Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-locked rank one editing for text-to-image personalization. In *SIGGRAPH*, 2023. 1, 2, 6

[48] 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 *CVPR*, 2018. 4

[49] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *ICCV*, 2023. 8

[50] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *ICCV*, 2023. 1, 2, 4, 6

[51] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*, 2023. 1, 2, 3, 4, 5, 6, 7, 8

[52] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. BiSeNet: Bilateral Segmentation Network for Real-time Semantic Segmentation. In *ECCV*, 2018. 5

[53] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, 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. 2

[54] Ge Yuan, Xiaodong Cun, Yong Zhang, Maomao Li, Chenyang Qi, Xintao Wang, Ying Shan, and Huicheng Zheng. Inserting anybody in diffusion models via celeb basis. In *NeurIPS*, 2023. 1, 2, 3, 4, 6, 7

[55] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023. 2, 6