ID-Booth: Identity-consistent Face Generation with Diffusion Models

Darian Tomašević[1], Fadi Boutros[2], Chenhao Lin[3], Naser Damer[2], Vitomir Štruc[4] and Peter Peer[1]
[1] University of Ljubljana, Faculty of Computer and Information Science, Ljubljana, Slovenia
[2] Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany
[3] Xi'an Jiaotong University, School of Cyber Science and Engineering, Xi'an, China
[4] University of Ljubljana, Faculty of Electrical Engineering, Ljubljana, Slovenia

***Abstract*— Recent advances in generative modeling have enabled the generation of high-quality synthetic data that is applicable in a variety of domains, including face recognition. Here, state-of-the-art generative models typically rely on conditioning and fine-tuning of powerful pretrained diffusion models to facilitate the synthesis of realistic images of a desired identity. Yet, these models often do not consider the identity of subjects during training, leading to poor consistency between generated and intended identities. In contrast, methods that employ identity-based training objectives tend to overfit on various aspects of the identity, and in turn, lower the diversity of images that can be generated. To address these issues, we present in this paper a novel generative diffusion-based framework, called ID-Booth. ID-Booth consists of a denoising network responsible for data generation, a variational auto-encoder for mapping images to and from a lower-dimensional latent space and a text encoder that allows for prompt-based control over the generation procedure. The framework utilizes a novel triplet identity training objective and enables identity-consistent image generation while retaining the synthesis capabilities of pretrained diffusion models. Experiments with a state-of-the-art latent diffusion model and diverse prompts reveal that our method facilitates better intra-identity consistency and inter-identity separability than competing methods, while achieving higher image diversity. In turn, the produced data allows for effective augmentation of small-scale datasets and training of better-performing recognition models in a privacy-preserving manner. The source code for the ID-Booth framework is publicly available at `https://github.com/dariant/ID-Booth`.**

I. Introduction

Deep learning models are nowadays utilized as backbones in a variety of recognition systems [1]. These models typically require sufficiently diverse and large-enough training datasets to achieve competitive performance. However, obtaining suitable datasets can be difficult in the field of biometrics, due to copyright, consent and privacy issues [22], [29]. Given the recent advancements in generative models, researchers are increasingly exploring the use of synthetic data to address the data needs of contemporary deep learning models. This is especially true in the area of face recognition, where synthetic data may be used to train face recognition models, augment existing datasets and enrich the variations present in real-world datasets [6].

State-of-the-art generative models are currently dominated by diffusion-based techniques, which offer unparalleled synthesis capabilities in terms of quality and diversity of the generated data, while enabling synthesis guided by text prompts [40]. Recently, diffusion models have also been utilized to produce datasets suitable for face recognition tasks, i.e., containing images of multiple identities with multiple samples each. To this end, approaches rely on identity-conditioning [34], [51], [52] and fine-tuning [42], [35] of pretrained diffusion models. Nevertheless, most solutions focus mainly on image reconstruction during training, resulting in poor consistency between the desired identities and the generated ones. To address this issue, PortraitBooth [35] recently extended the fine-tuning DreamBooth [42] method with an identity-based training objective. However, the proposed solution only considers the identity similarity of input samples and the generated samples during training. In turn, it tends to overfit on input identity features, including undesired characteristics, e.g., the pose, age, hair, accessories, thus lowering the diversity of generated images.

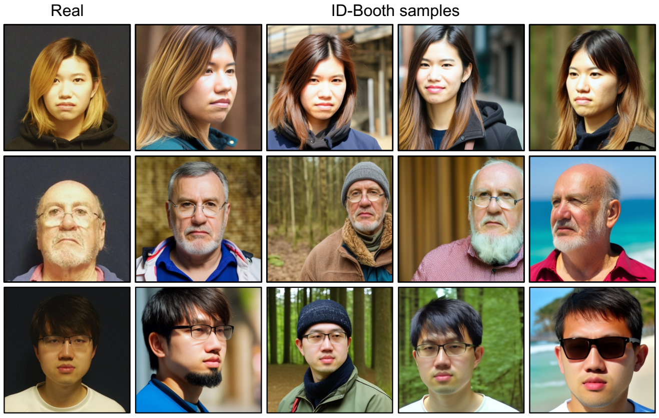

Fig. 1. **Samples generated with the proposed ID-Booth framework.** The framework enables fine-tuning of pretrained diffusion models for generating diverse identity-consistent images based on images gathered in a constrained setting with the consent of subjects.

In this paper, we present a solution for the outlined issues, in the form of a new generative framework, called ID-Booth. The proposed framework entails three main components, including (i) a denoising network that produces data based on input noise, (ii) a Variational Auto-Encoder (VAE) that maps images to and from a more efficient latent space on which the denoising network operates, and (iii) a text encoder that enables prompt-based conditioning of the denoising network. The proposed framework utilizes a novel triplet identity objective, which considers both positive and negative identity samples during training, to facilitate the generation of identity-consistent images while retaining the synthesis capabilities of pretrained models. Throughout the experiments, we explore the suitability of ID-Booth for addressing privacy concerns by generating diverse synthetic in-the-wild images of identities from the Tufts Face Database [33], which contains images gathered in a constrained setting with the consent of subjects, as shown in Figure 1. We perform fine-tuning of a state-of-the-art diffusion model conditioned on diverse prompts and compare synthesis results with DreamBooth [42] and PortraitBooth [35] in terms of image quality, fidelity and diversity as well as intra-identity consistency and inter-identity separability. Furthermore, we investigate the real-world utility of the produced synthetic samples for augmenting existing datasets to train modern face recognition models in a privacy-preserving manner. We demonstrate that our fine-tuning framework enables the generation of more diverse synthetic samples with better intra-identity consistency and inter-identity separability. As showcased by improved recognition performance, across five real-world verification benchmarks, this makes our approach more suitable for augmenting small-scale training datasets than DreamBooth [42] or PortraitBooth [35]. Overall, the paper makes the following contributions:

- We introduce ID-Booth, a novel framework for generating highly-diverse identity-consistent privacy-preserving face images.
- We propose a novel triplet identity learning objective for fine-tuning that improves identity consistency while retaining better image diversity.
- We demonstrate the suitability of the produced data for augmenting existing small-scale datasets and show that training with the mixed images leads to better performing face recognition models.

II. Related work

Image generation. The field of image synthesis has undergone rapid development since the introduction of deep generative models. Generative Adversarial Networks (GANs) [13] were the initial models to achieve the synthesis of convincing images, with a generator and a discriminator network. Extensive improvements followed, namely StyleGAN [24] facilitated higher image quality and better control over the generation process. However, the synthesis capabilities of GANs have nowadays been surpassed by recent diffusion models [11], which generate images by gradually removing noise from initial noisy samples. This denoising process is learned with a convolutional encoder-decoder by predicting the noise that is added to training samples at different scales [18]. Recently, Latent Diffusion Models (LDMs) [40] achieved improved efficiency and efficacy by moving the denoising process from the pixel space to a lower-dimensionality latent space of a pretrained variational autoencoder. Their remarkable synthesis capabilities and conditioning on text prompts via a pretrained text encoder have led to their broad adoption, namely of the open-source Stable Diffusion (SD) model [40]. Image resolution of these models has been further improved by utilizing a larger U-Net backbone along with two text encoders and additional conditioning schemes [36]. Recent approaches have also enhanced control over the generation process, e.g., ControlNet [54] conditions the model on segmentation masks or depth maps via an auxiliary trainable copy of the model, while IP-Adapter [52] utilizes image features as a condition through a decoupled cross-attention mechanism. Fine-tuning approaches have also been developed to incorporate new concepts into pretrained diffusion models by training on a minimal set of images [42].

Generating synthetic face recognition data. Generative models and synthetic data hold considerable potential in face recognition by enabling the creation of large-scale (training and test) datasets with predefined characteristics, facilitating augmentation in data-scarce application scenarios, and balancing data across different demographics [6]. To enable control over various characteristics of generated faces, Deng *et al.* [10] conditioned StyleGAN [24] on input 3D face priors. However, recognition models trained on the generated data achieved worse performance than those trained on real-world data. To tackle this issue, Qiu *et al.* [38] introduced identity and domain mixup of synthetic and real data during training. Boutros *et al.* [5] proposed to condition StyleGAN2 [23] on one-hot encoded identity labels. This improved intra-identity diversity at the cost of lowered inter-identity separability and a limited amount of possible identities. To address this, Tomašević *et al.* [48] instead utilized identity features from a pretrained face recognition model as the condition, in addition to enabling the generation of multispectral data.

Recently, Boutros *et al.* [4] achieved the generation of identity-specific images with latent diffusion models by conditioning the denoising network on face recognition features. The proposed contextual partial dropout also prevented overfitting on identities and enabled control over inter-identity separability and intra-identity diversity. Differently, more recent approaches relied on pretrained diffusion models [40] rather than training the models from scratch. Ruiz *et al.* [42] presented the DreamBooth method that can associate a new identity to a rare text token through fine-tuning on images of the identity. During training, face images generated by the pretrained model are also used to preserve prior synthesis capabilities. Arc2Face [34] instead replaces the identity token with recognition features and fine-tunes the model on a large-scale dataset. The textual-part of the prompt is frozen, so that control is tied primarily to the identity features, thus enabling more consistent generation of input identities. However, this comes at the cost of losing powerful prompt-based control. The recent IP-Adapter [52] has also been modified to use identity features as the condition, while retaining control of text prompts through decoupled cross-attention. InstantID [51] extends these capabilities by incorporating spatial control with an auxiliary ControlNet-based [54] module conditioned on facial landmarks and features. Despite advancements, identity consistency remained problematic, as the identity aspect was not considered in training objectives. To address this, Peng *et al.* [35] introduced PortraitBooth, which incorporates an identity-based objective into the fine-tuning of DreamBooth [42]. However, the solution only relies on the identity similarity of training images and generated noisy images, despite the success of more refined objectives on face recognition tasks [49]. As a result, the approach can overfit even on undesired characteristics of training identities, e.g., their pose, age or face accessories, which lowers the diversity of produced images. In contrast, our proposed ID-Booth framework utilizes a triplet objective that relies on the identity similarity between generated images and both training images (i.e., positive samples) and prior images produced by the initial model (i.e., negative samples). This enables better identity consistency, while better retaining synthesis capabilities of pretrained latent diffusion models.

III. Methodology

In this section we present the inner-workings and the main components of the novel ID-Booth framework that enables the generation of diverse high-fidelity identity-consistent facial images suitable for augmenting existing small-scale datasets captured with the consent of subjects.

A. The ID-Booth framework

The proposed diffusion-based ID-Booth framework consists of three primary components, as depicted in Figure 2. This includes (i) the denoising network, responsible for enabling data generation through diffusion, (ii) the Variational Auto-Encoder (VAE) that is used to map images to and from a more efficient latent space, and (iii) the text encoder, which enables prompt-based control of the generation process. Fine-tuning of pretrained diffusion models is then achieved with three training objectives, (i) the conventional reconstruction loss on a small set of input samples, (ii) the prior preservation loss, focused on combating overfitting through the reconstruction of images generated by the model before training, and (iii) the triplet identity loss, which utilizes a pretrained face recognition model to guide the diffusion model toward better similarity of generated identities and desired identities in input samples rather than random identities in prior images. Details of each component and objective are provided below.

Denoising network. At the core of the diffusion model lies the denoising network, which is trained to reverse a noising process that gradually degrades training images by adding noise at different scales. This entails the corruption of a real data sample $x_0 \sim p(x_0)$ into its noised versions $x_1, ..., x_T$ through a Markov chain of length T, as follows:

$$x_t = \mathcal{N}(\sqrt{\alpha_t} x_{t-1}, 1 - \alpha_t), \quad \forall t \in 1, ..., T, \tag{1}$$

where $\alpha_1, ..., \alpha_T$ represent a fixed variance schedule. However, any step of the noised sample can also be efficiently produced directly from the input x_0 [18] as:

$$x_t = \sqrt{\bar{\alpha}_t} x_0, (1 - \bar{\alpha}_t)\epsilon, \tag{2}$$

with $\bar{\alpha}_t := \prod_{s=0}^{t} \alpha_s$, which enables uniform sampling of t. Through training, the denoising network (i.e., typically a U-Net network [41]), learns to estimate the real data distribution from a noise-filled standard Gaussian distribution. This entails gradually denoising a noisy image $x_T \sim \mathcal{N}(0, \mathbf{I})$ to less noisy samples x_t until a denoised data sample x_0 is reached. To this end, the denoising network $\epsilon_\theta(x_t, t)$ predicts the noise ϵ that is added at step t with Equation (2).

Variational Auto-Encoder (VAE). To greatly improve efficiency, the noising and denoising processes are carried out in the latent space of a pretrained Variational Auto-Encoder (VAE) instead of the pixel space [40]. This is achieved by first mapping the input sample x_0 to the latent input z_0 through the encoder model $\mathcal{E}$ of the VAE. Noising with Equation (2) is then performed to obtain noised samples z_t on which the denoising network ϵ_θ is trained. During inference, synthetic images can then be generated by randomly sampling a noisy sample z_T in the latent space, denoising it with the predictor ϵ_θ, and then mapping the denoised sample z_0 back to the pixel space with the VAE decoder $\mathcal{D}$.

Text encoder. To enable control over the generation process, the denoising network is also conditioned on input text prompts [18]. The text prompt is first tokenized and mapped to corresponding token embeddings, which are then encoded through a pretrained CLIP text encoder [39]. Encoded prompts c are then passed as conditions to the denoising network through the cross-attention mechanism [7].

B. Training objectives of ID-Booth

Pretrained diffusion models provide unparalleled text-guided synthesis capabilities, owing to training on various datasets of unprecedented scale [40]. However, their knowledge of very specific concepts and styles remains limited. This is also true for their ability to create images of a desired identity as prompting for a specific non-celebrity identity can be difficult or even impossible.

To facilitate the generation of identity-specific images, we propose to fine-tune a pretrained diffusion model on a small set of input images of a desired identity. Our proposed ID-Booth framework utilizes three separate training objective, to improve identity consistency while retaining the synthesis capabilities of pretrained models. This includes the reconstruction loss $\mathcal{L}_{REC}$, the prior preservation loss $\mathcal{L}_{PR}$ and a triplet-identity loss $\mathcal{L}_{TID}$, which are combined as follows to form the overall objective:

$$\mathcal{L}_{Total} = \mathcal{L}_{REC} + \lambda_{PR}\mathcal{L}_{PR} + \lambda_{TID}\mathcal{L}_{TID}, \tag{3}$$

as illustrated in Figure 2. Here, the balancing weight λ_{PR} is set to 1.0, while λ_{TID} is defined as $1 - \frac{t}{T^2}$ to gradually reduce the identity influence at higher timesteps as image blurriness increases. Detailed descriptions of the three training objectives are provided below.

To further retain the capabilities of pretrained models, while still enabling fine-tuning on new identities, our ID-Booth framework also relies on the use of the Low-Rank Adaptation method (LoRA) [19]. Thus, instead of fine-tuning the entire diffusion model, all existing weights remain frozen while new low-rank trainable layers are introduced in the denoising network. This allows for better retention of synthesis capabilities, while enabling faster training and more efficient storage of fine-tuned model weights.

Reconstruction loss. The first training objective of our ID-Booth framework is aimed at image reconstruction and is based on the reweighted optimization objective conventionally used for training diffusion models [18]. Since denoising

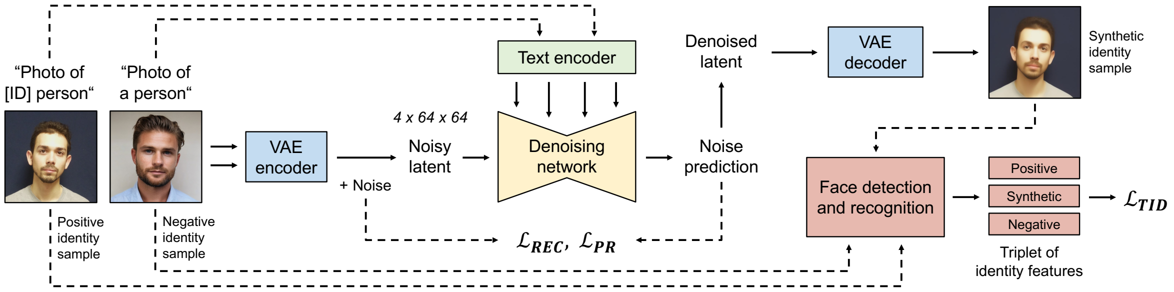

Fig. 2. **Overview of the ID-Booth framework.** The framework utilizes three training objectives to fine-tune a pretrained diffusion model. $\mathcal{L}_{REC}$ and $\mathcal{L}_{PR}$ are aimed at the reconstruction of training and prior images. Differently, the proposed triplet identity objective $\mathcal{L}_{TID}$ focuses on the identity similarity between generated samples and both training and prior samples to improve identity consistency without impacting the capabilities of the pretrained model.

is performed in the latent space of a pretrained VAE, the loss is based on the noise ϵ that is added to sample z_0 at timestep t and the noise that is estimated by the denoising network ϵ_θ considering the noisy latent sample z_t, the timestep t, and the text prompt condition c. Formally, this reconstruction loss $\mathcal{L}_{REC}$ can be defined as follows:

$$\mathcal{L}_{REC} = \mathbb{E}_{z\sim\mathcal{E}(x),\epsilon\sim\mathcal{N}(0,1),t,c}\left[\|\epsilon - \epsilon_\theta(z_t, t, c)\|_2^2\right]. \quad (4)$$

Prior preservation loss. Fine-tuning a diffusion model on a small set of images with only the reconstruction objective $\mathcal{L}_{REC}$ often leads to overfitting on input data and the loss of prior knowledge, e.g., the concept of what a person is. To address this, our ID-Booth framework utilizes an additional training objective aimed at the preservation of prior concepts [42]. To this end, a set of prior images $x_{pr,0}$ are generated by the initial pretrained model prior to training, with prompts related to the novel concept to be introduced. Following the initial reconstruction objective, these prior samples are used to form the prior preservation loss $\mathcal{L}_{PR}$ following the DreamBooth approach [42]:

$$\mathcal{L}_{PR} = \mathbb{E}_{z_{pr},c_{pr},\epsilon',t'}\left[\epsilon_{pr} - \epsilon_\theta(z_{pr,t'}, t', c_{pr})\|_2^2\right], \quad (5)$$

where the pr notation represents factors related to prior images generated with the initial model.

Triplet identity loss. Despite the suitability of $\mathcal{L}_{REC}$ and $\mathcal{L}_{PR}$ for fine-tuning, both objectives are focused solely on image reconstruction and do not target the consistency of generated identities. This is the case for both consistency with desired input identities and consistency among generated samples. To address this, we propose to incorporate the identity aspect into the training process through the similarity of identity features extracted from images with a pretrained face recognition model. However, to enable the inspection of generated identities during training, suitable face images must be produced at each training step. To this end, we use the predicted noise $\epsilon_\theta(z_t, t, c)$ and the latent noisy sample z_t to estimate the denoised latent $\hat{z}_0$ as:

$$\hat{z}_0 = \frac{z_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta}{\sqrt{\bar{\alpha}_t}}. \quad (6)$$

Afterward, we can decode the estimated denoised latent $\hat{z}_0$ to the estimated input image with $\hat{x}_0 = \mathcal{D}(\hat{z}_0)$. Next, we extract the facial region with a face detection model for both the estimated and the input training image, denoted as $\hat{x}_0^f$ and x_0^f respectively. If the facial region exist, we obtain the identity feature representations of each image with a pretrained face recognition model φ.

To guide the generative model toward better identity consistency, we propose to form a triplet identity objective. The objective utilizes identity features of the reconstructed sample $\hat{x}_0$ as the anchor, the input image x_0 as a positive example of an identity and prior images $x_{pr,0}$ as a negative example. Formally, our proposed triplet identity objective $\mathcal{L}_{TID}$ can be defined using cosine similarity cos as follows:

$$\mathcal{L}_{TID}{=}max\{cos(\varphi(x_0^f),\varphi(\hat{x}_0^f)){-}cos(\varphi(x_{pr,0}^f),\varphi(\hat{x}_0^f)){+}m,0\}, \quad (7)$$

where the notations introduced before apply. In addition, m represents a non-negative margin, i.e., the minimum difference between positive and negative similarities that is required for the loss to be zero. The proposed triplet-objectives also addresses the risk of overfitting on unintentional characteristics of training samples, e.g., the pose, age, hair or accessories, which might leak into the identity embeddings. This is achieved through negative identity examples, which often share similar characteristics with positive examples.

IV. Experiments and Results

Dataset preparation. To fine-tune ID-Booth, we utilize the Tufts Face Database (TFD) [33], which contains images captured in a constrained laboratory setting with the consent of subjects. In total, the dataset includes over $10,000$ images of 113 human subjects captured across various light spectra. We focus on images captured with four visible field cameras under constant diffused light in a semi-circle around the subjects. During preprocessing, we remove heavily blurred images and extreme side-profile images lacking key facial features (e.g., two eyes), then crop them to focus on the face region, resulting in 2299 images of 107 subjects. Next, we use eye landmarks, detected with the Multi-Task Cascaded Convolutional Neural Network (MTCNN) [53] to align the faces through an affine transform, and then resize the images to 512×512. For evaluation we rely on the Flickr Faces High-Quality (FFHQ) [24] dataset of $70,000$ diverse in-the-wild unlabeled face images, which we also resize to 512×512.

Implementation details. We evaluate the suitability of our framework on the state-of-the-art diffusion model Stable

Diffusion 2.1 (SD-2.1) [40], which is capable of generating high-quality and diverse images at a resolution of 512×512 through 1000 denoising timesteps, specified by the discrete denoising scheduler with $\beta_{end} = 0.012$, $beta_{start} = 8.5 \times 10^{-4}$ [18]. We fine-tune the SD-2.1 model on images of each identity in the Tufts Face Database (TFD) [33]. To this end, we utilize the training objectives specified by either DreamBooth [42], focused primarily on image reconstruction, PortraitBooth [42], which includes a simple two-point identity objective, or by our proposed ID-Booth framework, that balances identity consistency and image diversity through $\mathcal{L}_{Total}$. The identity objectives are based on features of a ResNet-100 face recognition model trained with the ArcFace loss [9] from face regions of noisy samples detected with MTCNN [53]. The detection of faces also acts as the decision factor for when identity-based training objectives are applied. To minimize the effect on the synthesis capabilities of the pretrained model, we utilize the Low-Rank Adaptation (LoRA) [19] method, which freezes the diffusion model but introduces new trainable layers instead. Specifically, we add two linear layers of rank 4 to each cross-attention block, initialized with a Gaussian distribution. We also generate 200 images with the initial SD-2.1 model and the prompt `photo of a person`, which are used for preservation of prior concepts through $\mathcal{L}_{PR}$ [42]. We then perform fine-tuning with images of a desired identity and the prompt `photo of [ID] person`, where `[ID]` represents a rare text token that will be tied to the new identity, in our case `sks`, following existing works [42]. We utilize an initial learning rate of 10^{-4} and the AdamW optimizer [27] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$ and a weight decay of 0.01, along with the half-precision floating point format to lower VRAM usage. The fine-tuning process is stopped after 32 epoch (i.e., 6400 steps), based on our initial observations and existing works [42], [35].

Data generation. Each fine-tuned SD model is used to generate two synthetic datasets, one with 21 images per identity, as is the case in TFD [33], and one with 100 images per identity to investigate the scalability of our approach. Data generation is performed with a guidance scale of 5.0 and 30 inference denoising steps with the same discrete denoising scheduler as during training. The goal is to generate diverse synthetic images of desired identities under various scenarios. To produce images that resemble real-world in-the-wild datasets [24], we utilize a prompt that defines a face image of a specific identity as well as the environment the image is taken:

```
face [P] photo of [G] [ID] person, [B] background
```

Here `[ID]` represents the identity token, while `[G]` defines the gender of the person, i.e., `female` or `male`. To generate diverse images we also select the environment through `[B]`, which we randomly sample from the following list:

```
forest, city street, bus, office, factory, beach,
laboratory, construction site, hospital, night club
```

To also produce a variety of poses we randomly select whether the image should be a portrait or a side-portrait, represented by `[P]`. In addition, we rely on a negative prompt to ensure the synthesis of more realistic images:

```
cartoon, render, illustration, painting, drawing,
black and white, bad body proportions, landscape
```

An ablation study of the main prompt components is available in the supplementary material.

Evaluation methodology. We evaluate the suitability of the proposed ID-Booth framework by comparing its synthesis capabilities to those of DreamBooth [42] and PortraitBooth [35]. Other diffusion-based frameworks that produce identity-specific images, e.g., Arc2Face [34] and InstantID [51], are not considered as they are trained on large-scale web-scraped face recognition datasets, without the consent of subjects. Meanwhile, our experiments entail fine-tuning on a limited amount of images from TFD [33] gathered with suitable consent. To evaluate the produced images, we compare them to the diverse real-world images of FFHQ [24]. Here, we consider either entire images or only the face regions of a resolution 112×112, which are aligned and cropped based on face landmarks detected by MTCNN [53]. The quality of images is then determined with the use of Fréchet Distance [17] and Kernel Distance [2], measured on features extracted with the pretrained DINOv2-ViT-L/14 model [32] rather than the typical Inception-v3 [47] model, which has been shown to be unsuitable, due to poor correlation with human evaluators [46] and the limitations of the ImageNet dataset [8]. We also evaluate the fidelity and diversity of images separately, with the use of Density and Coverage [31], measured on features of DINOv2-ViT-L/14 [32]. To compute these scores we utilize the generated datasets with 100 samples per identity and compare them to 10.000 samples from FFHQ [24]. In addition, we analyze the intra-identity diversity of samples with the Vendi score [12] computed on the extracted features, which differently from previous measures does not require a reference dataset. Similarly, we rely on the Certainty Ratio Face Image Quality Assessment (CR-FIQA) [3] to evaluate the quality of each face image through relative classifiability with a pretrained ResNet-101 [15]. We also analyze intra-identity diversity by evaluating the pitch, yaw and roll of faces in the images with the 6DRepNet [16] head pose estimator.

Recognition experiment details. As part of our experiments, we also investigate intra-identity consistency and inter-identity separability based on genuine and imposter distributions. These are formed using the cosine similarity of identity features of synthetic samples and either samples of the corresponding identity (genuine pair) or a different identity (imposter pair), from either TFD [33] or the synthetic dataset. The identity features are extracted with a Resnet-101 [15] recognition model trained with the ArcFace loss [9] on the MSV1MV3 dataset [14] from face regions of the images detected with MTCNN [53]. To allow for a fair comparison with samples of TFD [33], we form the distributions with synthetic datasets that also consist of 21 samples per identity. For each dataset combination, we form all possible genuine pairs along with an equal amount of

TABLE I

Quantitative evaluation of quality, fidelity and diversity of synthetic images. Quality is assessed with Fréchet Distance [17] and Kernel Distance [2], while fidelity and diversity are measured through Density and Coverage [31]. Results are computed by comparing distributions of features extracted with DINOv2-ViT-L/14 [32] from synthetic images and real-world images of FFHQ [24], considering either entire images or only the face region. Vendi score [12] is used to evaluate intra-identity diversity, while CR-FIQA [3] measures face image quality of each sample, without a reference dataset.

		Fréchet Distance ↓		**Kernel Distance ↓**		**Density ↑**		**Coverage ↑**		**Vendi score per ID ↑**		**CR-FIQA**
Data from	**Method**	Entire	Face	Entire	Face	Entire	Face	Entire	Face	Entire	Face	Face region
TFD [33]	Real data	2035.615	1679.317	7.056	5.779	0.195	0.623	0.043	0.120	2.536	3.132	2.131 ± 0.094
FFHQ [24]	Real data	38.703	33.799	0.001	0.001	1.028	1.007	0.972	0.970	–	–	2.090 ± 0.134
SD-2.1	No fine-tuning	1123.226	1075.064	2.201	2.881	0.422	0.413	0.137	0.204	–	–	2.080 ± 0.238
	DreamBooth [42]	1374.696	1371.129	4.134	4.484	**0.698**	**0.575**	0.128	0.131	7.264	6.705	2.187 ± 0.134
	PortraitBooth [35]	1182.511	1202.684	3.000	3.568	0.575	0.510	0.149	0.154	12.192	9.614	2.149 ± 0.173
	ID-Booth (ours)	**1144.651**	**1159.537**	**2.778**	**3.346**	0.536	0.502	**0.157**	**0.166**	**13.510**	**10.430**	2.143 ± 0.181

(↓ / ↑) – Lower / Higher is better; (**Bold**) – Best result; (Underline) – Second best result

TABLE II

Evaluation of diversity through pose estimation. Reported are the mean and standard deviation of standard deviation values of pitch, yaw and roll measured across samples of each identity with the 6DRepNet [16] head pose estimator.

			Pose estimation	
Data from	**Method**	**Pitch (σ per ID)**	**Yaw (σ per ID)**	**Roll (σ per ID)**
TFD [33]	Real data	2.015 ± 0.718	26.297 ± 4.609	2.285 ± 1.114
SD-2.1	No fine-tuning	7.486 ± 1.487	24.909 ± 2.596	4.835 ± 1.656
	DreamBooth [42]	4.681 ± 1.232	26.118 ± 5.261	3.248 ± 1.449
	PortraitBooth [35]	6.249 ± 2.423	33.159 ± 6.735	5.241 ± 2.563
	ID-Booth (ours)	$\mathbf{6.641 \pm 2.662}$	$\mathbf{33.527 \pm 7.569}$	$\mathbf{5.637 \pm 2.920}$

(↓ / ↑) – Lower / Higher is better; (**Bold**) – Best result; (Underline) – Second best result

randomly sampled imposter pairs. We report the mean and standard deviation of distributions along with established metrics, including Equal Error Rate (EER), False Match Rate at a False Non-Match Rate of 1.0% (FMR100) or 0.01% (FMR1000), False Non-Match Rate at a False Match Rate of 1.0% (FNMR100) or 0.01% (FNMR1000), and the Fisher Discriminant Ratio (FDR) [21]. Lastly, we use the produced data to augment the TFD [33] dataset, which is then used to train a ResNet-50 [15] recognition model with the AdaFace loss [25]. For training we utilize a batch size of 128 and the Stochastic Gradient Descent (SGD) optimizer with 0.9 momentum, a weight decay of 5×10^{-4}, and a dropout ratio of 0.4. The learning rate is initially set to 0.1 and is lowered by a factor of 10 after the 22nd, the 30th, and the 35th epoch. Training is stopped once no improvement in 5 epochs is observed on the LFW [20] benchmark. The performance of the trained model is then evaluated on five state-of-the-art verification benchmarks, including Labeled Faces in the Wild (LFW) [20], its Cross-Age and Cross-Pose subsets CA-LFW [56] and CP-LFW [55], Celebrities in Frontal-Profile in the Wild (CFP-FP) [44] and AgeDB-30 [30].

Experimental hardware. The experiments were conducted on a cluster of 4 Nvidia A100 SXM4 40GB GPUs as well as a Desktop PC with an AMD Ryzen 7 7800X3D CPU with 128 GB of RAM and an Nvidia RTX 4090 GPU.

A. Evaluation of generated images

Image quality. We begin our evaluation by assessing the overall quality of images, produced by either the proposed ID-Booth framework, or its two main competitors, DreamBooth [42] and PortraitBooth [35]. To this end, utilize the Fréchet Distance [17] and Kernel Distance [2] computed between synthetic features and features of the real-world FFHQ [24] dataset extracted with DINOv2-ViT-L/14 [32]. From results reported in Table I and samples in Figures 1 and 3, we can discern that with the SD-2.1 model and our defined prompts we can generate images that better match the quality of in-the-wild FFHQ [24] images than the real-world images of TFD [33], which were gathered in a constrained environment. Comparing results of the different fine-tuning methods, we see that our ID-Booth framework achieves the best quality results, scoring closest to the non fine-tuned model, while enabling identity-specific generation. This is the case both when evaluating entire images or only their face regions. Interestingly, both approaches with identity-based training objectives (i.e., PortraitBooth [35] and ID-Booth) also generate images of a better quality than DreamBooth [42]. In addition, we evaluate the quality of each face region with CR-FIQA [3]. Here, however, high face quality is not necessarily as desired as having a mix of high and low quality images, which can lead to better performing recognition models. This difference can also be observed on real-world datasets, where in-the-wild images of FFHQ [24] achieve a lower mean but higher standard deviation than constrained images of TFD [33]. Similarly, compared to DreamBooth [42] and PortraitBooth [35], our

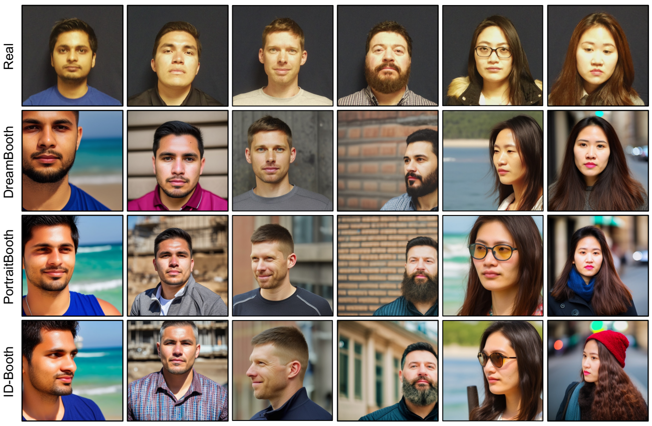

Fig. 3. **Comparison of generated image samples.** ID-Booth facilitates better identity consistency than DreamBooth [42] and better image diversity than PortraitBooth [35], which limits the variety of facial features and poses.

ID-Booth framework achieves a higher standard deviation but lower mean of CR-FIQA [3] scores, that are closer to those of diverse samples of the non fine-tuned models.

Image fidelity and diversity. Next, we analyze the produced images in terms of fidelity, i.e., the degree to which they resemble real samples, and diversity, i.e., how well they cover the variability of real samples [43]. To this end, we rely on Density and Coverage [31], respectively, reported in Table I, in addition to qualitative samples in Figures 3 and 4. Comparing different datasets with FFHQ [24], we can observe that images of TFD [33] lack the fidelity and diversity expected of in-the-wild images. Synthetic samples produced by the non-finetuned SD-2.1 model with diverse prompts offer a notable improvement in these areas. In comparison, fine-tuning approaches achieve a higher fidelity of entire images and face regions, but crucially also result in lower diversity of face regions. Among the approaches, DreamBooth [42] scores the highest in terms of density (i.e., fidelity), while our ID-Booth achieves the highest coverage (i.e., diversity) on both entire images and detected face regions. This can also be observed in Figures 3 and 4, where ID-Booth generates images that offer more consistent identities, while allowing for a larger variety of poses, ages, accessories and other facial features, unlike DreamBooth [42] or PortraitBooth [35].

Intra-identity diversity. To further investigate the produced images, we also analyze the intra-diversity of samples with the per-class Vendi score [12]. As reported in Table I, synthetic images generated by all fine-tuning methods offer more intra-identity diversity than the constrained samples of TFD [33]. Our ID-Booth achieves the largest intra-identity diversity among the fine-tuning approaches, both of entire images and only face regions, while ensuring better identity consistency. This can be seen in Figure 4, where ID-Booth samples of the same identity contain a larger variety of poses and face accessories. To obtain deeper insight, we also analyze the pitch, yaw and roll of faces with the 6DRepNet [16] head pose estimator. In Table II we report the mean and standard deviation of standard deviation values obtained from pose distributions of each identity. Results reveal that ID-Booth generates samples with the largest variety of poses per identity, especially in terms of pitch and roll, which represents an important aspect of overall intra-identity diversity. In comparison, DreamBooth [42] and PortraitBooth [35] often default to more front-facing poses, as seen in Figures 3 and 4.

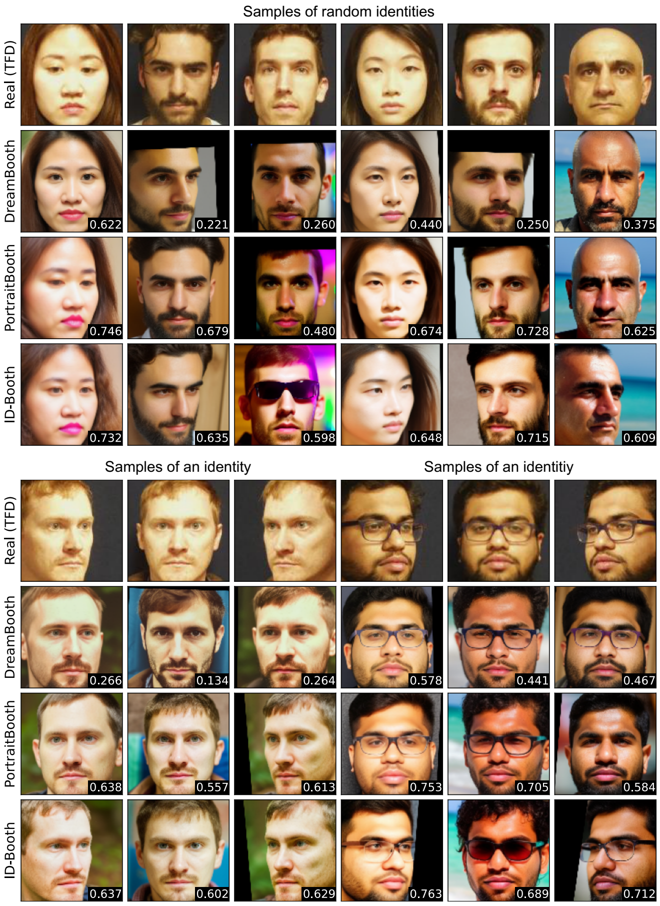

Fig. 4. **Comparison of identity consistency.** ID-Booth achieves better identity consistency than DreamBooth [42], while retaining more diverse synthesis capabilities and ensuring better intra-identity diversity than PortraitBooth [35]. Reported is the cosine similarity of synthetic and real identity features extracted with the pretrained ArcFace recognition model [9].

B. Recognition-based experiments

Identity consistency and separability. To determine the suitability of generated images for forming recognition datasets we must also examine the consistency and separability of identities in the images. To this end, we form genuine and imposter distributions either only among synthetic identities or between synthetic and real-world identities, based on the similarity of features extracted with the pretrained ArcFace-based recognition model [9]. From verification results in Table III that describe these distributions, we can discern that both PortraitBooth [35] and ID-Booth achieve better identity consistency and separability than DreamBooth [42], due to the use of identity-based training objectives. This is the case both either among synthetic or between synthetic and real identities across all verification metrics. The only exception is the FNMR1000 score in the latter scenario, which can likely be attributed to a handful of outliers, especially when considering the lower FNMR100 scores and improved FDR values. Compared to PortraitBooth [35], our ID-Booth achieves lower FNMR scores among synthetic identities along with lower FMR and FNMR scores between synthetic and real identities, indicating fewer outliers. It also achieves a notably higher FDR score in the second scenario, which in combination with above observations signifies better intra-identity consistency and inter-identity separability. This is further supported by ID-Booth samples in Figure 4, which showcase better consistency between generated and real identities or among different synthetic samples of the same identity. Overall, these results highlight crucial characteristics of ID-Booth that demonstrate its suitability for augmenting existing datasets in a privacy-preserving manner by producing identity-consistent in-the-wild images of real-world identities from the training dataset for which we have consent.

TABLE III

Evaluation of consistency and separability between synthetic and real-world identities. Reported are verification measures of genuine and imposter distributions among synthetic or between synthetic and real images, constructed based on the cosine similarity of identity features obtained with a pretrained ArcFace-based recognition model [9].

Data setting	**Method**	**EER** ↓	**FMR100 / 1000** ↓	**FNMR100 / 1000** ↓	**Imposter** $\mu \pm \sigma$ ↓	**Genuine** $\mu \pm \sigma$ ↑	**FDR** ↑
among TFD [33]	Real data	0.002	0.002 / 0.002	0.001 / 0.003	0.021 ± 0.0725	0.871 ± 0.070	70.969
among SD-2.1	DreamBooth [42]	0.055	0.153 / 0.337	0.297 / $\underline{0.919}$	0.103 ± 0.093	$\mathbf{0.499 \pm 0.141}$	5.509
	PortraitBooth [35]	$\underline{0.043}$	$\underline{0.096}$ / $\underline{0.230}$	$\underline{0.269}$ / 0.967	$\underline{0.065 \pm 0.084}$	$\underline{0.495 \pm 0.151}$	**6.210**
	ID-Booth (ours)	**0.042**	**0.095 / 0.217**	**0.249 / 0.896**	$\mathbf{0.059 \pm 0.082}$	0.486 ± 0.153	$\underline{6.073}$
SD-2.1 vs. TFD	DreamBooth [42]	0.046	0.087 / 0.184	0.286 / **0.684**	0.019 ± 0.072	0.406 ± 0.155	5.132
	PortraitBooth [35]	$\underline{0.028}$	$\underline{0.048}$ / $\underline{0.112}$	$\underline{0.133}$ / 0.848	$\underline{0.017 \pm 0.073}$	$\underline{0.465 \pm 0.153}$	$\underline{6.987}$
	ID-Booth (ours)	**0.027**	**0.044 / 0.091**	**0.110** / $\underline{0.838}$	$\mathbf{0.017 \pm 0.072}$	$\mathbf{0.465 \pm 0.148}$	**7.402**

(↓ / ↑) – Lower / Higher is better; (**Bold**) – Best result; ($\underline{\text{Underline}}$) – Second best result

TABLE IV

Verification performance of recognition models trained on real and synthetic data. Reported is the accuracy of a ResNet-50 model trained with AdaFace loss [25] across 5 real-world verification benchmarks. LFW [20] is used for validation.

Training setting		**Verification accuracy on benchmarks** ↑					
Dataset (samples per ID)	**Method**	**LFW**	**AgeDB-30**	**CA-LFW**	**CFP-FP**	**CP-LFW**	**Average accuracy**
TFD (21)	Real data	0.688	0.500	0.557	0.593	0.540	0.575 ± 0.063
TFD (21) + SD-2.1 (21)	DreamBooth [42]	$\underline{0.755}$	0.531	0.602	0.603	$\underline{0.576}$	0.613 ± 0.075
	PortraitBooth [35]	0.753	$\underline{0.561}$	**0.615**	$\underline{0.618}$	0.576	$\underline{0.624 \pm 0.068}$
	ID-Booth (ours)	**0.765**	**0.595**	$\underline{0.612}$	**0.621**	**0.587**	$\mathbf{0.636 \pm 0.066}$
TFD (21) + SD-2.1 (100)	DreamBooth [42]	0.766	0.561	0.606	**0.646**	0.592	0.634 ± 0.071
	PortraitBooth [35]	$\underline{0.787}$	$\underline{0.565}$	$\underline{0.627}$	0.631	$\underline{0.605}$	$\underline{0.643 \pm 0.076}$
	ID-Booth (ours)	**0.790**	**0.609**	**0.635**	$\underline{0.632}$	**0.606**	$\mathbf{0.654 \pm 0.069}$

(↑) – Higher is better; (**Bold**) – Best result; ($\underline{\text{Underline}}$) – Second best result

Training face recognition models. Lastly, we also explore the utility of the generated data in a real-world scenario for training deep face recognition models. As part of our experiments, we utilize the produced synthetic datasets to augment the small-scale real-world Tufts Face Database (TFD) [33] with in-the-wild syntethic images of its identities, images of which were gathered with consent in a constrained laboratory setting. The augmented datasets are then used to train a ResNet-50 [15] recognition model with the AdaFace loss [25], following the procedure described in Section IV. The suitability of the synthetic datasets obtained with different fine-tuning methods is then evaluated based on the performance of the trained recognition model on five state-of-the-art face verification benchmarks.

We first explore augmentation with synthetic datasets consisting of 21 samples per identity, denoted as SD-2.1 (21), which matches the scale of TFD [33]. Results reported in Table IV reveal that this form of augmentation enables the training of drastically better performing recognition models compared to training on only real-world samples of TFD [33]. Among the fine-tuning approaches, our ID-Booth framework achieves the highest overall accuracy, with improvements over DreamBooth [42] and PortraitBooth [35] being especially evident on the AgeDB-30 [30] and CP-LFW [55] benchmarks. In total, the use ID-Booth for augmentation results in a 6.1% average accuracy increase over the recognition model trained on the non-augmented dataset. In the second set of experiments, we investigate the augmentation capabilities of larger synthetic datasets consisting of 100 samples per identity, denoted as SD-2.1 (100). From results in Table IV we can discern notable accuracy improvements across with all fine-tuning approaches, showcasing their scalability. Similar to previous results, our ID-Booth framework achieves higher overall verification accuracy than DreamBooth [42] or PortraitBooth [35], with a total average augmentation improvement of 7.9%. As before, the most noticeable improvement is observed on the cross-age benchmarks (i.e., AgeDB-30 [30] and CA-LFW [56]). Overall, the accuracy improvements obtained with our ID-Booth framework can be attributed to the novel triplet identity objective, which facilitates better identity consistency and higher intra-identity diversity, especially in terms of age and pose, than either DreamBooth [42] or PortraitBooth [35].

V. Conclusion

In this paper, we presented ID-Booth, a new diffusion-based fine-tuning framework for generating high-quality identity-consistent images, suitable for augmenting existing small-scale datasets in a privacy-preserving manner. To this end, ID-Booth relies on a novel triplet identity training objective that improves both intra-identity consistency and inter-identity separability, while retaining image diversity of pretrained state-of-the-art diffusion models. Throughout the experiments, we showcase that training deep recognition models on datasets augmented with synthetic samples of ID-Booth results in better performance across five verification benchmarks than when performing augmentation with synthetic samples of existing approaches or training only on real-world data. With regards to future work, we aim to investigate the applicability of identity-based objectives in the training of conditioning approaches and exploring the creation of larger-scale datasets.

VI. Acknowledgements

Supported in parts by the Slovenian Research and Innovation Agency (ARIS) through the Research Programmes P2-0250 (B) "Metrology and Biometric Systems" and P2–0214 (A) “Computer Vision”, the ARIS Project J2-50065 "DeepFake DAD", and the ARIS Young Researcher Programme.

ETHICAL IMPACT STATEMENT

This research primarily focuses on investigating the utility of synthetic data for face recognition tasks, which has potential benefits for privacy-preserving biometric systems [6], [29]. The proposed ID-Booth framework is designed to generate high-quality synthetic facial images that ensure identity consistency while preserving image diversity, making them suitable for dataset augmentation. The generated data can help mitigate privacy concerns associated with real face datasets by reducing the need for large-scale data collection, which often includes web-scraping without the proper consent of subjects in the images [22]. Furthermore, this work limits the need for the distribution and storage of personally identifiable facial images.

However, the use of generative models for facial image synthesis also carries ethical considerations. The potential misuse of identity-consistent synthetic faces includes unauthorized impersonation, deepfake-related fraud, and misuse in surveillance applications. Furthermore, biases inherent in the training data of latent diffusion models [40] could be reflected in the generated images, potentially leading to demographic disparities.

To mitigate these risks, our work follows strict ethical guidelines and established protocols from prior works on synthetic face generation [6]. The training dataset used in our research consists of images from the Tufts Face Database [33], which were captured with the consent of subjects in a constrained laboratory setting. Thus, we do not train on or generate images of real individuals that have not given their consent. Other datasets [24] and benchmarks [20], [30], [44], [55], [56] used solely for evaluation are also publicly available, thus ensuring compliance with standard ethical practices in data handling.

We advocate for the responsible use of ID-Booth by emphasizing its intended applications in privacy-preserving synthetic dataset generation rather than identity spoofing or deceptive practices. Future work should focus on bias mitigation strategies, ensuring fair representation across demographics, and improving safeguards against potential misuse. Further exploration of automated detection methods for synthetic faces could also enhance the transparency and accountability of generative models.

References

[1] X. Bai, X. Wang, X. Liu, Q. Liu, J. Song, N. Sebe, and B. Kim. Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments. *Pattern Recognition*, 120:108102, 2021.

[2] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations (ICLR)*, 2018.

[3] F. Boutros, M. Fang, M. Klemt, B. Fu, and N. Damer. CR-FIQA: Face image quality assessment by learning sample relative classifiability. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5836–5845, 2023.

[4] F. Boutros, J. H. Grebe, A. Kuijper, and N. Damer. IDiff-Face: Synthetic-based face recognition through fizzy identity-conditioned diffusion model. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19650–19661, 2023.

[5] F. Boutros, M. Huber, P. Siebke, T. Rieber, and N. Damer. SFace: Privacy-friendly and accurate face recognition using synthetic data. In *IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–11. IEEE, 2022.

[6] F. Boutros, V. Struc, J. Fierrez, and N. Damer. Synthetic data for face recognition: Current state and future prospects. *Image and Vision Computing*, page 104688, 2023.

[7] C.-F. R. Chen, Q. Fan, and R. Panda. CrossViT: Cross-attention multi-scale vision transformer for image classification. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 357–366, 2021.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[9] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2019.

[10] Y. Deng, J. Yang, D. Chen, F. Wen, and X. Tong. Disentangled and controllable face image generation via 3D imitative-contrastive learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5154–5163, 2020.

[11] P. Dhariwal and A. Nichol. Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:8780–8794, 2021.

[12] D. Friedman and A. B. Dieng. The Vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2024.

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2672–2680, 2014.

[14] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In *Springer European Conference on Computer Vision (ECCV)*, pages 87–102, 2016.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[16] T. Hempel, A. A. Abdelrahman, and A. Al-Hamadi. Toward robust and unconstrained full range of rotation head pose estimation. *IEEE Transactions on Image Processing (TIP)*, 33:2377–2387, 2024.

[17] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6626–6637, 2017.

[18] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[19] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

[20] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[21] Information technology – Biometric performance testing and reporting – Part 1: Principles and framework. Standard, International Organization for Standardization, 2021.

[22] C. Jasserand. Massive facial databases and the GDPR: The new data protection rules applicable to research. In *Data Protection and Privacy: The Internet of Bodies*, pages 169–188. 2018.

[23] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila. Training generative adversarial networks with limited data. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 12104–12114, 2020.

[24] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410,

2019.

[25] M. Kim, A. K. Jain, and X. Liu. AdaFace: Quality adaptive margin for face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18750–18759, 2022.

[26] T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen, and T. Aila. Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

[27] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, pages 1–18, 2019.

[28] D. Maio, D. Maltoni, R. Cappelli, J. L. Wayman, and A. K. Jain. FVC2000: Fingerprint verification competition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 24(3):402–412, 2002.

[29] B. Meden, P. Rot, P. Terhörst, N. Damer, A. Kuijper, W. J. Scheirer, A. Ross, P. Peer, and V. Štruc. Privacy–enhancing face biometrics: A comprehensive survey. *IEEE Transactions on Information Forensics and Security (TIFS)*, 16:4147–4183, 2021.

[30] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou. AgeDB: The first manually collected, in-the-wild age database. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 51–59, 2017.

[31] M. F. Naeem, S. J. Oh, Y. Uh, Y. Choi, and J. Yoo. Reliable fidelity and diversity metrics for generative models. In *PMLR International Conference on Machine Learning (ICML)*, pages 7176–7185, 2020.

[32] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. HAZIZA, F. Massa, A. El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

[33] K. Panetta, Q. Wan, S. Agaian, S. Rajeev, S. Kamath, R. Rajendran, S. P. Rao, A. Kaszowska, H. A. Taylor, A. Samani, et al. A comprehensive database for benchmarking imaging systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 42(3):509–520, 2018.

[34] F. P. Papantoniou, A. Lattas, S. Moschoglou, J. Deng, B. Kainz, and S. Zafeiriou. Arc2Face: A foundation model for id-consistent human faces. In *Springer European Conference on Computer Vision (ECCV)*, pages 241–261, 2024.

[35] X. Peng, J. Zhu, B. Jiang, Y. Tai, D. Luo, J. Zhang, W. Lin, T. Jin, C. Wang, and R. Ji. PortraitBooth: A versatile portrait model for fast identity-preserved personalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27080–27090, 2024.

[36] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations (ICLR)*, 2024.

[37] N. Poh and S. Bengio. A study of the effects of score normalisation prior to fusion in biometric authentication tasks. Technical report, IDIAP, 2004.

[38] H. Qiu, B. Yu, D. Gong, Z. Li, W. Liu, and D. Tao. Synface: Face recognition with synthetic data. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10880–10890, 2021.

[39] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *PMLR International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021.

[40] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.

[41] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer, 2015.

[42] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22500–22510, 2023.

[43] M. S. Sajjadi, O. Bachem, M. Lucic, O. Bousquet, and S. Gelly. Assessing generative models via precision and recall. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018.

[44] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs. Frontal to profile face verification in the wild. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9, 2016.

[45] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[46] G. Stein, J. Cresswell, R. Hosseinzadeh, Y. Sui, B. Ross, V. Villecroze, Z. Liu, A. L. Caterini, E. Taylor, and G. Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

[47] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016.

[48] D. Tomašević, F. Boutros, N. Damer, P. Peer, and V. Štruc. Generating bimodal privacy-preserving data for face recognition. *Engineering Applications of Artificial Intelligence (EAAI)*, 133:108495, 2024.

[49] D. S. Trigueros, L. Meng, and M. Hartnett. Enhancing convolutional neural networks for face recognition with occlusion maps and batch triplet loss. *Image and Vision Computing*, 79:99–108, 2018.

[50] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu. CosFace: Large margin cosine loss for deep face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5265–5274, 2018.

[51] Q. Wang, X. Bai, H. Wang, Z. Qin, and A. Chen. InstantID: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024.

[52] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

[53] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[54] L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3836–3847, 2023.

[55] T. Zheng and W. Deng. Cross-Pose LFW: A database for studying cross-pose face recognition in unconstrained environments. *Beijing University of Posts and Telecommunications, Tech. Rep*, 5(7), 2018.

[56] T. Zheng, W. Deng, and J. Hu. Cross-Age LFW: A database for studying cross-age face recognition in unconstrained environments. *arXiv preprint arXiv:1708.08197*, 2017.

[57] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3D solution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 146–155, 2016.

Supplementary material

A. Additional results and ablation studies

Ablation of training objectives. To determine the suitability of the selected configuration for our framework we first perform an ablation study regarding the training objectives utilized for fine-tuning the diffusion model. Following the evaluation methodology in Section IV, we report in Table V results related to quality, fidelity and diversity on synthetic datasets consisting of 100 samples per identity. In addition, we provide qualitative samples in Figure 5 that demonstrate the effects of the different loss functions. As can be discerned from the samples, fine-tuning the Stable Diffusion 2.1 (SD-2.1) [40] model with only $\mathcal{L}_{REC}$ leads to generated samples that closely resemble training samples from the Tufts Face Database (TFD) [33]. Despite the utilized prompts, which specify the environment and subject pose as described in Section IV, the generated images often contain a uniform background, similar to the backdrop used in TFD [33]. These observations are supported by results in Table V, where we can observe that the synthetic images are more similar to images of FFHQ [24] than the real-world images of TFD [33] across all reported metrics. However, many of the images also contain minor artifacts including blurry faces, as also observed through considerably higher standard deviation of CR-FIQA [3] scores. The density and coverage [31] scores, which measure fidelity and diversity, are also rather low, especially when considering entire images. In combination with poor prompt adherence, these signs point to overfitting on characteristics of training samples from the Tufts Face Database (TFD) [33].

The addition of the prior preservation objective $\mathcal{L}_{PR}$ resolves some of these issues by utilizing additional training images, which are generated by the pretrained model before fine-tuning. This prevents overfitting on undesired image characteristics of TFD [33] (e.g., the background), thus facilitating better prompt adherence and in turn the generation of more diverse image samples, as depicted in Figure 5. While this is not reflected in quality-based metrics, we do see a noticeable improvement in terms of density (i.e., fidelity) [31] on entire images, which point to backgrounds that resemble in-the-wild images. The face regions also become sharper, as denoted by improved CR-FIQA [3] scores. However, these improvements come at the price of identity consistency both among synthetic and between synthetic and real samples, as discerned from verification measures of genuine and imposter distributions reported in Table VII.

To address the identity-based issues raised by the higher diversity of samples enabled by $\mathcal{L}_{PR}$, we propose to utilize a triplet identity learning objective $\mathcal{L}_{TID}$. As seen by results in Table VII, the proposed identity-based objective achieves better results across the majority of measures, noticeably improving both intra-identity consistency and inter-identity separability. At the same time, the $\mathcal{L}_{TID}$ also ensures the generation of images that better match in-the-wild images of FFHQ [24], in terms of quality and diversity, as revealed by improvements in Fréchet Distance [17], Kernel Distance [2], Coverage [31] and per-class Vendi score [12] in Table V. Overall, the combination of the three objectives ensures the generation of diverse high-quality images of desired identities.

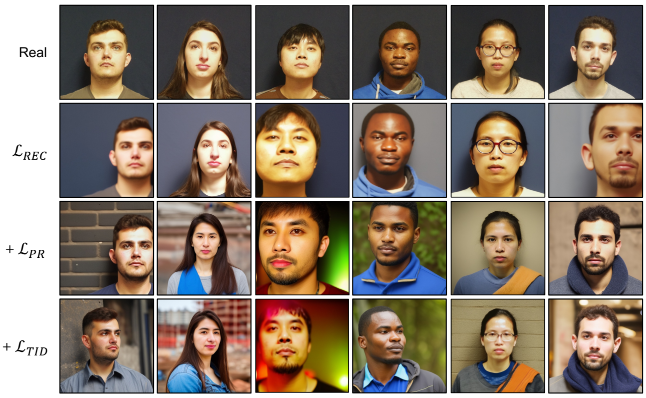

Fig. 5. **Ablation study of ID-Booth training objectives.** Shown are sample images generated by the ID-Booth framework trained with different training objectives. Training only with $\mathcal{L}_{REC}$ generates images similar to the training set, disregarding the given prompts. $\mathcal{L}_{PR}$ improves the diversity of samples but lowers identity consistency. Our proposed $\mathcal{L}_{TID}$ presents an objective that improves both diversity and identity consistency.

Ablation of inference prompts. As part of our experiments, we rely on a variety of prompts to facilitate the generation of diverse images that best match in-the-wild images of FFHQ [24] while still retaining identity consistency. In this section, we investigate the effects of each prompt component presented in Section IV. This includes the background `[B]`, the negative prompt, the subject gender `[G]` and the subject portrait pose `[P]`. We begin the experiments with the base prompt `face portrait photo of [ID] person`, a more face-focused version of the training prompt, which had to be kept simple to ensure that the identity was correctly linked to the ID token during fine-tuning [42]. As can be observed from qualitative samples in Figure 6, generating images with solely the base prompt results in images that often portray subjects in a similar constrained environment as in the training images of TFD [33]. The generated identities are consistent, as seen by verification results similar to TFD [33] in Table VII. However, the quality, fidelity and diversity of images, reported in Table V, is far from the desired characteristics of in-the-wild images of FFHQ [24], despite being closer than constrained images of TFD [33]. The addition of the background prompt component `[B]` drastically improves scores related to quality, fidelity and diversity of generated data in Table V. The intra-identity diversity of samples is also notably improved, both in terms of the Vendi score [12] as well as the standard deviation of head poses, reported in Table VI. However, as can be discerned from samples in Figure 6, the additional complexity of the prompt also introduces prominent artifacts ranging from unnatural backgrounds and face features to gender changes. In turn, this also immensely lowers intra-identity consistency and inter-identity separability, as reported in Table VII.

The use of an additional negative prompt, that guides the generation process away from undesired styles addresses some of these issues, as seen by improvements across all

TABLE V

Ablation study of different training objectives and prompt components of ID-Booth through quantitative evaluation of quality, fidelity and diversity. Quality is assessed with Fréchet Distance [17] and Kernel Distance [2], while fidelity and diversity are measured through Density and Coverage [31]. Results are computed by comparing distributions of features extracted with DINOv2-ViT-L/14 [32] from synthetic images and real-world images of FFHQ [24], considering either entire images or only the face region. Vendi score [12] is used to evaluate intra-identity diversity, while CR-FIQA [3] measures each synthetic sample separately, both without a reference dataset.

		Fréchet Distance ↓		**Kernel Distance ↓**		**Density ↑**		**Coverage ↑**		**Vendi score per ID ↑**		**CR-FIQA**
Data from	**Loss / Prompt**	Entire	Face	Entire	Face	Entire	Face	Entire	Face	Entire	Face	Face region
TFD [33]	–	2035.615	1679.317	7.056	5.779	0.195	0.623	0.043	0.120	2.536	3.132	2.131 ± 0.094
FFHQ [24]	–	38.703	33.799	0.001	0.001	1.028	1.007	0.972	0.970	–	–	2.090 ± 0.134
ID-Booth	$\mathcal{L}_{REC}$	1256.490	1258.637	3.187	3.657	0.330	0.551	0.119	0.130	9.754	8.462	2.079 ± 0.226
	$+ \mathcal{L}_{PR}$	1374.696	1371.129	4.134	4.484	**0.698**	**0.575**	0.128	0.131	7.264	6.705	2.187 ± 0.134
	$+ \mathcal{L}_{TID}$	**1144.651**	**1159.537**	**2.778**	**3.346**	0.536	0.502	**0.157**	**0.166**	**13.510**	**10.430**	2.143 ± 0.181
ID-Booth	Base prompt	1849.488	1889.947	7.418	7.639	0.348	0.310	0.034	0.031	3.674	3.185	2.136 ± 0.107
	+ `[B]`	1394.284	1454.935	3.923	4.631	0.392	0.362	0.096	0.097	13.534	11.761	2.108 ± 0.175
	+ Negative prompt	1201.857	1250.269	3.261	4.016	0.543	0.475	0.151	0.141	**14.876**	**12.173**	2.158 ± 0.159
	+ `[G]`	1185.778	1230.030	3.359	4.085	**0.603**	**0.531**	0.153	0.153	11.794	9.358	2.176 ± 0.132
	+ `[P]`	**1144.651**	**1159.537**	**2.778**	**3.346**	0.536	0.502	**0.157**	**0.166**	13.510	10.430	2.143 ± 0.181

(↓ / ↑) – Lower / Higher is better; (**Bold**) – Best result; (Underline) – Second best result

TABLE VI

Ablation of ID-Booth prompt components through pose estimation. Reported are the mean and standard deviation of standard deviation values of pitch, yaw and roll measured across samples of each identity with the 6DRepNet [16] head pose estimator.

			Pose estimation	
Data from	**Prompt**	**Pitch (σ per ID)**	**Yaw (σ per ID)**	**Roll (σ per ID)**
TFD [33]	–	2.015 ± 0.718	26.297 ± 4.609	2.285 ± 1.114
ID-Booth	Base prompt	2.738 ± 0.681	7.254 ± 3.470	1.269 ± 0.448
	+ `[B]`	5.147 ± 0.962	15.592 ± 4.307	2.522 ± 0.949
	+ Negative	4.787 ± 0.925	16.096 ± 4.135	2.370 ± 0.892
	+ `[G]`	4.414 ± 0.862	17.114 ± 4.504	2.360 ± 0.913
	+ `[P]`	$\mathbf{6.641 \pm 2.662}$	$\mathbf{33.527 \pm 7.569}$	$\mathbf{5.637 \pm 2.920}$

(↓ / ↑) – Lower / Higher is better; (**Bold**) – Best result; (Underline) – Second best result

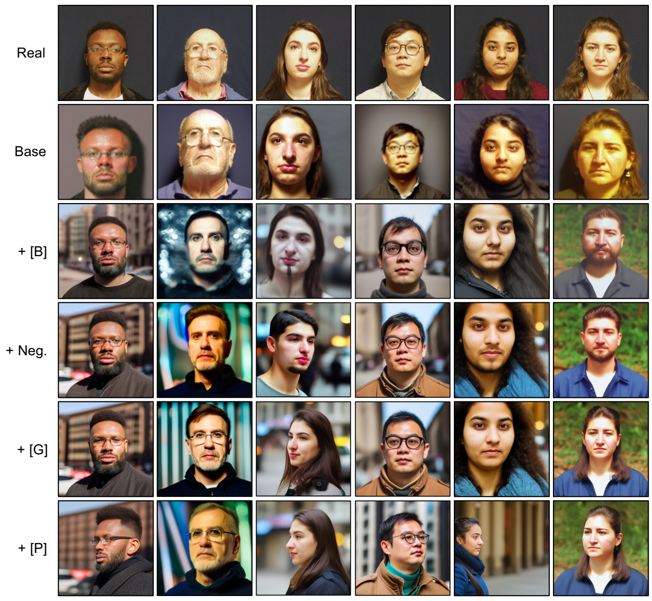

Fig. 6. **Ablation study of prompt components used with ID-Booth.** Specifying the background `[B]` greatly improves diversity, but introduces artifacts. The negative prompt improves photorealism, while `[G]` solves issues related to gender. Adding the side-portrait component through `[P]` ensures the generation of more diverse poses.

measures in Table V. Furthermore, the negative prompt actually improves intra-identity diversity measured by the Vendi score [12] as well as identity consistency slightly. Differently, specifying the gender with `[G]` of the subject drastically influences and improves identity consistency, as reported in Table VII and fixes the unintentional gender changes as observed in samples of Figure 6, which is also reflected in lower intra-identity diversity. Lastly, to also address the lack of pose diversity in the samples, we utilize the final prompt component `[P]`, which specifies whether the image should be a portrait or a side-portrait. With this small change we can greatly influence the intra-identity diversity of head poses, as seen by drastic improvements in Table VI in terms of the pitch, yaw and roll. In turn, this also improves results related to overall image diversity in Table V. This addition also slightly negatively impacts the identity consistency, however, this is to be expected, which is to be expected as the features used for computing identity similarity are often affected by the head pose. Overall, the final constructed prompt ensures the generation of diverse high-quality images that better match the desired characteristics of in-the-wild images, while still ensuring identity consistency.

Real-world time requirements. Fine-tuning with our proposed ID-Booth framework takes 31 minutes on average for each identity of TFD [33] with an Nvidia A100 GPU. Similarly, PortraitBooth [35] requires 30 minutes on average for fine-tuning, likely due to the simpler training objective. Differently, DreamBooth [42], which does not utilize an identity-based objective, requires only 20 minutes for fine-tuning on average, as it does not require the decoding of latent samples during training.

B. Details regarding the evaluation methodology

Measuring image quality, fidelity and diversity. To evaluate the synthesis capabilities of the fine-tuned models we compare the produced synthetic images with in-the-wild face images of FFHQ [24]. To this end, we utilize the following performance measures:

- **Fréchet Distance [17]** which estimates the overall quality of synthetic images, by evaluating the difference between distributions of image features. For this, the

TABLE VII

ABLATION STUDY OF TRAINING OBJECTIVES AND PROMPT COMPONENTS OF ID-BOOTH IN TERMS OF IDENTITY CONSISTENCY AND SEPARABILITY. REPORTED ARE VERIFICATION MEASURES OF GENUINE AND IMPOSTER DISTRIBUTIONS AMONG SYNTHETIC OR BETWEEN SYNTHETIC AND REAL IMAGES, CONSTRUCTED BASED ON THE COSINE SIMILARITY OF IDENTITY FEATURES OBTAINED WITH A PRETRAINED ARCFACE-BASED RECOGNITION MODEL [9]. RESULTS OF $\mathcal{L}_{REC}$ OR THE BASE PROMPT ARE NOT INCLUDED WHEN DETERMINING THE BEST RESULTS AS THEY PRODUCE SAMPLES THAT TOO CLOSELY MATCH THE REAL-WORLD TRAINING SAMPLES OF TFD [33].

Data setting	Loss / Prompt	EER ↓	FMR100 / 1000 ↓	FNMR100 / 1000 ↓	Imposter $\mu \pm \sigma$ ↓	Genuine $\mu \pm \sigma$ ↑	FDR ↑
among TFD [33]	Real data	0.002	0.002 / 0.002	0.001 / 0.003	0.021 ± 0.0725	0.871 ± 0.070	70.969
among ID-Booth	$\mathcal{L}_{REC}$	0.027	0.042 / 0.096	0.226 / 0.885	0.056 ± 0.079	0.534 ± 0.141	8.704
	+ $\mathcal{L}_{PR}$	0.055	0.153 / 0.337	0.297 / 0.919	0.103 ± 0.093	**0.499 ± 0.141**	5.509
	+ $\mathcal{L}_{TID}$	**0.042**	**0.095 / 0.217**	**0.249 / 0.896**	**0.059 ± 0.082**	0.486 ± 0.153	**6.073**
ID-Booth vs. TFD	$\mathcal{L}_{REC}$	0.008	0.007 / 0.018	0.003 / 0.220	0.021 ± 0.073	0.557 ± 0.109	16.709
	+ $\mathcal{L}_{PR}$	0.046	0.087 / 0.184	0.286 / **0.684**	0.019 ± 0.072	0.406 ± 0.155	5.132
	+ $\mathcal{L}_{TID}$	**0.027**	**0.044 / 0.091**	**0.110** / 0.838	**0.017 ± 0.072**	**0.465 ± 0.148**	**7.402**
among ID-Booth	Base prompt	0.002	0.001 / 0.002	0.000 / 0.009	0.062 ± 0.079	0.799 ± 0.100	33.634
	+ [B]	0.068	0.143 / 0.244	0.561 / 0.942	0.060 ± 0.080	0.454 ± 0.174	4.238
	+ Negative prompt	0.064	0.137 / 0.265	0.520 / 0.932	0.064 ± 0.082	0.485 ± 0.179	4.577
	+ [G]	**0.036**	**0.069 / 0.147**	**0.198 / 0.799**	0.070 ± 0.087	**0.533 ± 0.149**	**7.210**
	+ [P]	0.042	0.095 / 0.217	0.249 / 0.896	**0.059 ± 0.082**	0.486 ± 0.153	6.073
ID-Booth vs. TFD	Base prompt	0.004	0.003 / 0.007	0.000 / 0.118	0.018 ± 0.073	0.662 ± 0.095	28.963
	+ [B]	0.038	0.074 / 0.128	0.274 / 0.744	0.019 ± 0.074	**0.462 ± 0.160**	6.307
	+ Negative prompt	0.036	0.067 / 0.142	0.238 / 0.673	**0.014 ± 0.075**	0.466 ± 0.163	6.389
	+ [G]	**0.021**	**0.030 / 0.067**	**0.067 / 0.571**	0.017 ± 0.073	0.478 ± 0.138	**8.736**
	+ [P]	0.042	0.095 / 0.217	0.249 / 0.896	0.059 ± 0.082	0.486 ± 0.153	6.073

(↓ / ↑) – Lower / Higher is better; (**Bold**) – Best result; (Underline) – Second best result

original implementation (i.e., FID) utilizes an Inception-v3 network [47] pretrained on ImageNet [8] as a feature extractor. Differently, we rely on features extracted with DINOv2-ViT-L/14 [32], which offers a representation space that is more consistent with human evaluators and thus better suited for evaluating generative approaches [46].

- **Kernel distance [2]** represents an alternative to Fréchet Distance by utilizing the maximum mean discrepancy to measure the distance between distributions. Throughout our experiments we rely on the third degree polynomial kernel, following existing works [46] and compute the distance on image features of DINOv2-ViT-L/14 [32].
- **Density and Coverage [31]**, which measure the fidelity and diversity, respectively, by considering the distance between nearest neighbour embeddings of images. Density and coverage address issues with previous methods, namely precision and recall [26], by providing a measure that is more resilient to outliers. While the original implementation relied on features extracted with the VGG16 network [45] pretrained on ImageNet [8], we instead utilize features of DINOv2-ViT-L/14 [32].
- **Vendi score [12]**, which measures dataset diversity without the need for a reference dataset. However, when conditioned on the class (i.e., the identity) it can also be used to quantify intra-class diversity of samples. It can therefore be used alongside Coverage [31] to gain better insight into the diversity of generated data. To compute it we rely on image features extracted with DINOv2-ViT-L/14 [32].
- **Certainty Ratio Face Image Quality Assessment (CR-FIQA) [3]** measure, which is designed specifically for evaluating the quality of face images. It measures the quality through the relative classifiability of a given face image with a pretrained ResNet-101 network [15].

Here, it should be noted that the fine-tuned diffusion models produce images that often contain more context than just the face region, differently from the FFHQ dataset [24]. Thus, to a allow for a fair evaluation of specifically the face region we preprocess the generated images, following the preprocessing steps of FFHQ [24]. This includes first detecting facial landmarks with the Multi-Task Cascaded Convolutional Neural Network (MTCNN) [53] and then defining an affine transform to align them to a set of predefined positions. Finally, images are cropped to a resolution of 112×112, suitable for the AdaFace-based recognition model [25]. For Fréchet Distance [17], Kernel Distance [2], as well as density and coverage [31], which utilize both synthetic and real-world distributions for evaluation, we utilize the entire synthetic datasets with 100 samples per identity and $10,000$ samples from FFHQ [24]. To obtain a baseline for the scores, we also compare samples from the Tufts Face Database [33] with samples from FFHQ [24], as well as compare two sets of $10,000$ samples from FFHQ [24] between each other.

To further analyze intra-identity diversity in a more explainable manner, we also investigate the pitch, yaw and roll of samples for each identity with the state-of-the-art **6DRepNet [16]** head pose estimator, which is pretrained on the 300W-LP dataset [57].

Assesment of identity consistency and separability. We also investigate the generated images in terms of the identity aspect in order to better understand the consistency and separability of identities of generated datasets. For this purpose we utilize genuine and imposter score distribution plots, based on the cosine similarity of features extracted with the pretrained ArcFace recognition model [9]. Below we provide descriptions of the verification measures that we utilize throughout the experiments:

- **Equal Error Rate (EER)** [28], which is the point on the Receiver Operating Characteristics (ROC) curve, where the False Match Rate (FMR) equals the False Non-Match Rate (FNMR).
- **FMR100** and **FMR1000**, which report the lowest the False Non-Match Rate (FNMR) achieved at a False Match Rate (FMR) of 1.0% or 0.1% respectively.
- **FNMR100** and **FNMR1000** that represent the lowest the False Match Rate (FMR) achieved at a False Non-Match Rate (FNMR) of 1.0% or 0.1% respectively.
- **Fisher Discriminant Ratio (FDR)** [37], which quantifies the separability of genuine and imposter distributions.

Recognition experiments. In the experiments, we train the AdaFace recognition model [50] on the produced synthetic datasets, as described in Section IV. To determine their suitability, we evaluate the performance of the model on five real-world verification benchmarks. These include:

- **Labeled Faces in the Wild (LFW)** [20], which is an unconstrained web-scraped verification dataset of $13,233$ face images of 5749 identities.
- **Cross-Age Labeled Faces in the Wild (CA-LFW)** [56], which is a subset of LFW [20] with 7156 images of 2996 identities, aimed at evaluating verification performance across a given age gap.
- **Cross-Pose Labeled Faces in the Wild (CP-LFW)** [55], which is a LFW [20] subset that is suited specifically for evaluating cross-pose verification performance. It includes 5984 face images of 2296 identities captured in various poses.
- **AgeDB-30** [30], which is a dataset of in-the-wild face images, suited for evaluating verification performance across a 30 year age gap. The dataset comprises $16,488$ images of 568 identities.
- **Celebrities in Frontal-Profile in the Wild (CFP-FP)** [44], which is a verification dataset that is aimed at evaluating cross-pose performance, in particular of frontal and profile poses. In total, it contains 7000 images of 500 identities, each with 10 frontal and 4 profile images.

Each benchmark is formed with 3000 genuine and 3000 imposter image pairs of a given verification dataset, with an image resolution of 112×112. To limit the influence of race and gender, the CA and CP verification pairs are sampled from the same race and gender.