LCM-Lookahead for Encoder-based Text-to-Image Personalization

Rinon Gal[1,2*], Or Lichter[2*], Elad Richardson[2*], Or Patashnik[2], Amit H. Bermano[2], Gal Chechik[1], and Daniel Cohen-Or[2]

[1]NVIDIA [2]Tel Aviv University

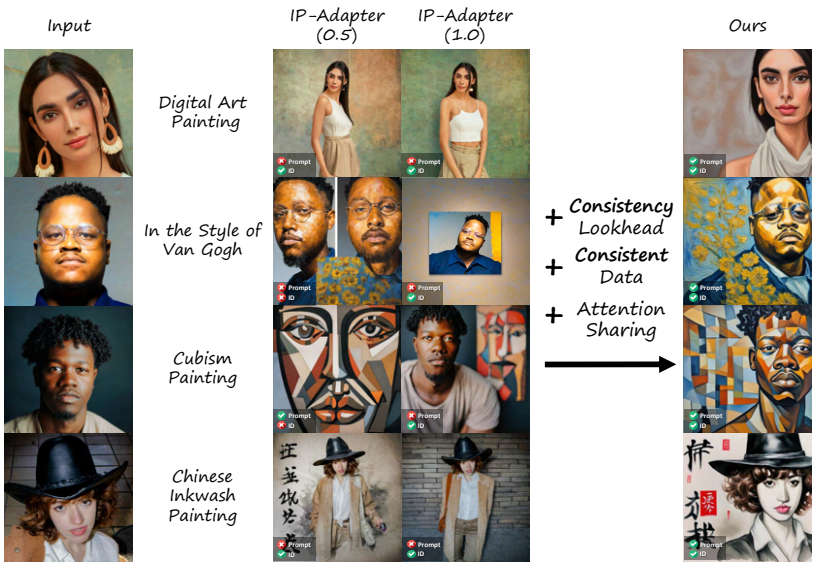

Fig. 1: We introduce a novel LCM-based lookhead mechanism to apply image-space losses to personalization encoder training. These are coupled with consistent data generation and attention sharing techniques to tune existing backbones and improve identity preservation and prompt alignment.

Abstract. Recent advancements in diffusion models have introduced fast sampling methods that can effectively produce high-quality images in just one or a few denoising steps. Interestingly, when these are distilled from existing diffusion models, they often maintain alignment with the original model, retaining similar outputs for similar prompts and seeds. These properties present opportunities to leverage fast sampling methods as a shortcut-mechanism, using them to create a preview of denoised outputs through which we can backpropagate image-space losses. In this work, we explore the potential of using such shortcut-mechanisms to guide the personalization of text-to-image models to specific facial identities. We focus on encoder-based personalization approaches, and demonstrate that by augmenting their training with a lookahead identity loss, we can achieve higher identity fidelity, without sacrificing layout diversity or prompt alignment. Code at `https://lcm-lookahead.github.io/`.

1 Introduction

Text-to-image personalization [18,55] methods enable users to tailor pretrained generative models to their own, *personal* data. Commonly, such methods focus on human data [19,67,70,81,77,56], where users aim to create novel images of specific individuals which were unseen by the pretrained model. Early works proposed to tackle this task by teaching a model new words that describe the user-provided subjects. They do so by optimizing novel word-embeddings [18,68], or by fine-tuning the generative model itself [55,57]. However, such approaches require significant per-subject optimization, leading to lengthy personalization times and large compute requirements. More recent lines of work propose to personalize the model using an encoder – a neural network trained to condition the generative model on user-provided images [19,73,61,79]. While these methods can enable inference-time personalization, they often struggle to maintain a subject's identity, or face difficulties in adapting it to novel styles.

One manner in which encoder-based methods try to bridge these gaps is by leveraging a pretrained face recognition network as a feature extraction backbone. The intuition here is that the pretraining objective of such networks drives them to encode fine identity details which can later be exploited by the encoder. However, this approach overlooks the training loss itself, which is still driven by only an L2 noise-prediction objective. In the realm of Generative Adversarial Networks (GANs, [21]), it was shown that inversion [1,76] can be significantly improved by incorporating additional losses that better align with human perception [53,45], *e.g.*, an identity loss. Applying such image-space losses to the personalization process could be beneficial here as well. However, the diffusion training process works on noisy image samples from intermediate diffusion time steps, and it is not clear how these should be passed into a perceptual loss which expects clean, realistic images.

Here, we present a method for tackling this question by building on recent advancements in fast sampling methods [63,59,22], and specifically latent consistency models (LCM, [41,42]). To do so, we leverage an intriguing property of generative models: fine-tuning alignment, where a child model fine-tuned from a pretrained parent tends to preserve the semantics of the parent's latent space [75,20]. In fig. 2, we show a particular manifestation of this alignment in LCM models, where we compare a single-step LCM output to the DDPM single-step approximations at intermediate steps of the denoising process. The LCM results are not only sharper, but they maintain a high degree of similarity with the final DDPM prediction. This property also holds for personalized models (right, dog). We find this alignment holds particularly well in consistency models, likely by virtue of the consistency training loss itself.

We propose to leverage this alignment during the training of personalization encoders, where we can create a high-quality preview of the denoiser's final output by doing a single LCM step on the noisy latents, guided by the same personalization encoder. This preview can then be used to calculate image-space losses, such as those derived from an identity detection network. Under this approach, the LCM-model provides a "shortcut" through which gradients can

Fig. 2: LCM output alignment. We first denoise an image partway using DDPM [25] sampling with a baseline SDXL model [47]. We then complete sampling in two manners: (top) By performing a single LCM step. (bottom) By approximating the clean image using DDPM. Even at early steps, LCM outputs provide a good approximation of the final DDPM prediction. This also holds for personalized models (*e.g.*, LoRA trained on the DreamBooth [55] dog, right).

back-propagate to earlier diffusion timesteps, without relying on low-quality approximations. In practice, we find that naïve applications of the shortcut loss can break the alignment between the baseline and LCM models. Hence, we investigate mechanisms for preserving alignment, and show that these can improve downstream performance.

In addition to the shortcut mechanism, we further explore an additional architectural modification inspired by recent video models and image editing works [74,11,33,50,10,43,24]. There, it has been shown that extending the self-attention mechanism, such that a generated image can also observe self-attention keys and values of a real, source image, can allow for zero-shot appearance transfer from the source to the new image [2,65]. Hence, we propose to augment the encoder with an additional path where the input image is noised, passed through a copy of the diffusion U-Net, and its self-attention keys and values are extracted and appended to the forward pass of the newly generated image.

Finally, we train our encoder by leveraging existing backbones [51,79] and tune it on newly generated data. To create our dataset, we leverage SDXL Turbo [59], a model distilled for single-step sampling from the vanilla SDXL using score distillation sampling [48] and adversarial training [21]. Notably, we observe that this distillation process causes a significant collapse of diversity, such that a sufficiently detailed prompt will generate the same identity regardless of the input seed. We leverage this property to generate a dataset of images with repeated identities and differing styles. Importantly, the use of generated data allows us to avoid the need to collect sensitive personal data, and ensures that our dataset contains proper representation for minority classes that are typically under-represented in real datasets.

We validate our approach through comparisons to a range of recent baselines, and through a large set of ablation scenarios. These demonstrate that our approach can achieve higher identity fidelity and prompt alignment, and highlight the benefit of integrating image-space losses during model tuning.

2 Related work

Text-to-image Personalization. In the task of text-to-image personalization [18,55], the goal is to adapt a pretrained model to better represent a user-given concept which was unseen during the original training. Initial efforts show that this can be achieved by either optimizing a new text-embedding [18] or by further tuning the denoising network itself [55]. This has spurred a variety of different approaches that try to extend the optimized embedding space [4,68,17], restrict the tuned part of the denoising network [57,35,64,8,23], or optimize for specific target prompts [7]. A joint limitation of all these different approaches is that they require a per-subject optimization process which can be time-consuming.

Encoder-Based Personalization. To overcome the limitations of optimization-based approaches, more recent work proposes that pretrained encoders can be used to initialize the optimization process, significantly shortening tuning times [19,56,6,36]. Others tried to completely avoid any optimization, at the cost of accuracy [73,12,61,79,30].

Recently the problem of face personalization gained focus, with a range of works specifically tackling this domain with the goal of improving identity preservation across prompts [19,67,70,81,77,56]. In PhotoMaker [37] the features of a CLIP-model [51] are modulated using a dedicated ID-oriented dataset. Face0 [67] proposed to replace the CLIP encoder with an identity recognition network. IP-Adapter [79] introduced similar ideas in later variants, with identity network features being used either instead of, or on top of the CLIP image embedding. FaceStudio [78] follows a similar mechanism but also uses a prior model to better adhere to the original prompt. The concurrent InsantID paper builds on IP-Adapter and extends it with a landmark-conditioned ControlNet [82] to preserve pose and simplify identity preservation.

Common to all these methods is that their training signal is based solely on the standard diffusion training loss. This is in contrast with the common practice in GAN-Based encoder methods [53,5,3,66,71,16], which demonstrated improved results by integrating perceptual losses, such as an identity loss. This gap can be attributed to the challenge of applying pixel-based losses during diffusion training. The concurrent PortraitBooth [46] proposes to apply an identity loss only to images sampled with low noise levels, where one can well approximate the clean image. However, this limits the effectiveness of the loss.

Our work is a similar tuning-free face-identity encoder, but we propose to integrate an identity loss into to the model training by using an LCM-based shortcut mechanism.

Fast Diffusion Sampling. Standard diffusion models [25] and their text-to-image variants [54,44,52] are trained to denoise an image over a $1,000$ steps. This span can be significantly shortened by leveraging ordinary differential equation solvers to navigate the diffusion flow in fewer steps [62,39,40,38,31].

More recently, several methods were proposed to distill diffusion models into versions that can be sampled from in fewer steps [58]. The idea there is to teach

the network to predict points farther along the diffusion flow, either by directly predicting a baseline model's output after multiple steps [58], through consistency modeling techniques [63,42,41,34], by leveraging adverserial training [21] and score sampling [48] methods [59] or by matching distributions between two diffusion models [80]. Of these, LCM-Lora [42] has shown that the consistency modeling approach can be applied through a low-rank adaptation [27] of the model weights.

Our work leverages such models, and LCM-LoRA in particular, in order to introduce image-space losses into personalization encoder training.

3 Preliminaries

3.1 Text-to-image Personalization

Text-to-image personalization methods introduce new, user-provided concepts to a pretrained T2I diffusion model, typically by tuning novel word-embeddings, parameters of the denoising network itself, or mixture of both. Such optimization-based methods use a small (typically 3-5) image set depicting the concept, and tune the relevant parameters using the standard diffusion training loss [25]:

$$L_{Diffusion} := \mathbb{E}_{z,y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, y)\|_2^2\right], \tag{1}$$

where ϵ is an unscaled noise sample, ϵ_θ is the denoising network, t is the time step, z_t is an image or latent noised to time t, and y is some conditioning prompt containing a placeholder token which is used to describe the new concept.

To speed up the personalization process, prior work suggested the use of encoders — neural networks trained to take images of the subject, and map them to some conditioning code that can guide the pretrained diffusion network to generate an image of the subject. A common approach [73,30] popularized by IP-Adapter [79] is to do so by extending the diffusion denoiser with additional cross-attention layers, which receive tokens derived from some external encoder module (e.g. a frozen CLIP-backbone [51] with a small projection head). The novel cross-attention heads and the encoder module are then trained using the loss of , where images are drawn from large-scale datasets depicting an array of subjects, commonly from a single domain (e.g. human faces).

In this case, the diffusion loss can be rewritten as:

$$L_{Diffusion} := \mathbb{E}_{z_r,y,I_c,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_{r,t}, t, y, E(I_c))\|_2^2\right], \tag{2}$$

where I_c is a conditioning image sampled from the training set, z_r is the image latent of a reconstruction target showing the same subject as I_c, and $E(I_c)$ is a conditioning code derived from the encoder.

Our work builds on a pretrained IP-Adapter, augments it with additional self-attention-based features, and tunes it to improve identity and prompt alignment.

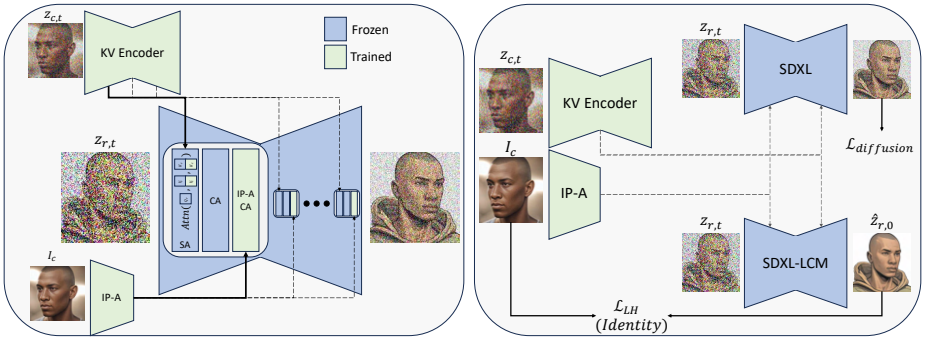

Fig. 3: **(left) Encoder architecture:** Our encoder has two branches: One is the standard IP-Adapter [79] that provides conditions through a new cross-attention head. The second branch is a copy of the SDXL U-net from which we extract self-attention keys and values, which we concatenate with those of the main denoising branch. **(right) Training setup:** The two encoder paths are provided with a conditioning image (and its noisy latent), and their outputs are used to condition the denoising of a *different* image of the same subject. We denoise the image with both the baseline SDXL [47] model and an LCM-model [42]. The baseline model's output is used for calculating the standard diffusion loss (eq. (2)). The LCM output is used to calculate the lookhead identity loss (eq. (4)). We portray latents as images for visual clarity.

4 Method

Our goal is to improve encoder-based methods so that we can concurrently achieve identity preservation, and prompt alignment. For identity preservation, we introduce an LCM-based identity lookahead loss, and a self-attention-sharing based architectural modification for existing encoder pipelines. To improve prompt alignment, we utilize a synthetic dataset which contains consistent characters generated in an array of prompts and styles. fig. 3 provides a high-level outline of our encoder architecture (left) and training process (right). In the following section, we provide additional details on each of these components.

4.1 LCM-Lookahead loss

We begin with the goal of achieving high identity fidelity. Here, we would like to draw on core ideas from existing GAN inversion literature, and particularly the use of identity networks as a loss during encoder training [53,45]. In the GAN-based literature, applying such a loss is trivial since the GAN can produce a clean image in a single forward pass. In the case of diffusion models, the standard training process involves sampling random intermediate diffusion time steps, and performing a single denoising step. These single-step predictions are typically noisy or blurry, and feeding them into a downstream image embedding network has previously been shown to be sub-optimal [15,69]. These prior

works showed that one can improve the quality of guidance by training feature extractors on noisy images, or by performing multiple forward steps to create a clean image, and back-propagating through the entire diffusion chain. However, the first approach is expensive, and the second is impractical in training scenarios: both because such training is typically already memory constrained, and because end-to-end sampling increases iteration times by a factor of ~ 100, making training infeasible. Instead, we propose to utilize a pretrained LCM-LoRA model [42], tuned from the same baseline SDXL [47] backbone, to create higher-quality previews of the fully-denoised images using a single diffusion step. This preview can be fed into the downstream feature extractor, in our case a face recognition network [14], and gradients can be backpropagated to the encoder through this LCM-path. We focus on the LoRA variant of the LCM model as it best preserves the alignment with the baseline model.

More concretely, let $(I_c,\ I_r)$ be a conditioning image and reconstruction-target pair, let $z_{r,t}$ be the image latent z_r, noised to an intermediary time step t, and let ϵ_{LCM} be the LCM denoising network. The preview image is given by:

$$\hat{z}_{r,0} = \frac{1}{\sqrt{\tilde{\alpha}_t}} \left(z_{r,t} - \sqrt{\tilde{\beta}_t} \cdot \epsilon_{LCM}(z_{r,t}, y, t, E(I_c)) \right), \tag{3}$$

where y and $E(I_c)$ are the prompt and encoder conditioning codes respectively, α_t and β_t are defined by the diffusion schedule. The lookahead loss is then:

$$\mathcal{L}_{LH} = \mathcal{D}\left(D_{VAE}(\hat{z}_{r,0}), I_c\right), \tag{4}$$

where $\mathcal{D}$ is some image-space distance metric (*e.g.* an identity loss) and D_{VAE} is the VAE decoder which maps the latents back to image space.

4.2 Maintaining alignment

In initial experiments, we found that applying the lookahead loss of eq. (4) can improve identities over short-training runs. However, over time, this loss causes the LCM pathway to focus solely on the loss-metric at the expense of its prior output, breaking the alignment with the baseline model. We investigated a series of options for improving alignment preservation, including the use of existing distribution matching options like Score Distillation Sampling [48], or appending the standard Consistency Model loss [62] to the LCM-path. Full details and results from this investigation are presented in the supplementary. In practice, the best downstream performance was achieved using a model interpolation approach. There, for half of our training iterations, we randomly re-scale the LoRA component of the LCM-LoRA using $\alpha_{LoRA} \in [0.1, 1.0]$. We hypothesize that applying the loss through the continuously interpolated model makes it more difficult for the encoder to converge to a solution which works differently for the LCM- and non-LCM paths. Moreover, this serves as a form of augmentation which makes adversarial solutions less likely. Lower values of α_{LoRA} can still act as a preview for intermediate (but not overly noisy) outputs, which can still provide a signal through the image-space model.

Finally, similarly to Wallace *et al.* [69], we find it useful to focus on early (noisy) diffusion time steps. Here, we do so by applying annealing to the training time step sampling. We use the importance-weighting function of Huang *et al.* [29]: $f(t) = \frac{1}{T}(1 - \alpha \cos \frac{\pi t}{T})$, where $f(t)$ is the probability to sample time step t, T is the total number of diffusion time steps (*i.e.* $1,000$) and α is a hyperparameter which we empirically set to 0.2.

4.3 Extended self-attention features

As a second component for improving identity fidelity, we propose to leverage recent ideas in video-based modeling [74,11] and appearance transfer [2,24,10]. In these works, it was shown that expanding the self-attention mechanism such that a generated image can attend to the keys and values derived from a source image, can lead to a significant increase in visual similarity between the generated image and the source. Here, we use a similar idea to transfer identity features from the conditioning image to the generated output. These are applied on top of the baseline encoder from which we start.

Specifically, we create a copy of the denoising U-Net which we call a "KV encoder". We pass a noisy version of our conditioning image through this U-Net, and cache the self-attention keys and values derived from this pass. Then, when performing the diffusion denoising pass, we append these keys and values to those derived from the denoised image at each self-attention layer: $K^l := K^l_{z_{r,t}} \odot K^l_{z_{c,t}}$, $V^{'l}_{z_{r,t}} := V^l \odot V^l_{z_{c,t}}$, where l is the layer index and $z_{r,t}, z_{c,t}$ subscripts denote attention features coming from the reconstruction target latent and the conditioning image latent respectively. This mechanism is illustrated in fig. 3.

Directly applying this approach with a pretrained U-net as our encoder can lead to excessive appearance transfer and loss of editing. Hence, we do not keep the encoder frozen, but rather tune it using LoRA [27]. As our training set (detailed below) contains target images which differ in style from the source, this draws the network towards discarding appearance properties that are related to the style and not to the content, greatly improving prompt alignment. We note that a similar attention-expansion idea was used for personalization in the concurrent work of Purushwalkam *et al.* [49]. However, their target images are all photo-realistic, and their results do indeed suffer from reduced editability.

Finally, this pathway requires us to slightly modify the standard classifier-free guidance [26] equation. See the supplementary for more details.

4.4 Consistent data generation

Having introduced components to improve identity preservation, we now turn to improving prompt alignment. Here, we hypothesize that the limited editability in current encoder-based methods is largely grounded in their training set, which typically focuses on reconstructing real images. Moreover, existing large-scale sets are either closed and proprietary or use the withdrawn LAION dataset [60]. They also commonly contain biases, which may result in models whose performance deteriorates on minority classes.

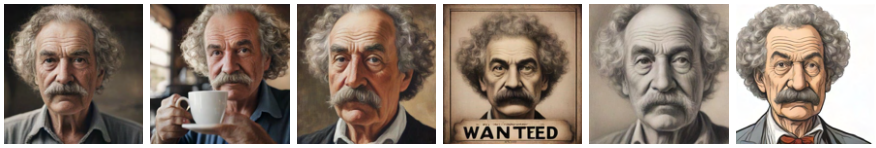

Fig. 4: **Consistent Data.** Consistent data generated using SDXL-Turbo with the description "old man with curly hair and a moustache" incorporated into different prompts (e.g. "as an oil painting", "as a wanted poster")

Instead of relying on such data, we propose to generate a novel consistent dataset in which we generated the same *synthetic* subjects across a wide range of prompts. These can then serve as training data for our encoder, where the encoder itself is provided with one image of a given identity, and the denoising goal considers another image portraying the same subject. Such cross-image training has already been shown to be beneficial with real data [30,37]. Here, the use of the generated data allows us to take the idea a step further, and ensure our reconstruction targets also contain stylized images. These can in turn prevent the encoder from focusing on photo-realism.

To create our data, we investigate multiple approaches for consistent generation, including: (1) generating celebrities, which the model is already familiar with and can consistently generate across various prompts, (2) ConsiStory [65], a recently introduced approach that aligns identities through attention feature sharing, and (3) using SDXL-Turbo [59].

While SDXL-Turbo is designed for fast sampling and not for consistent image generation, its adversarial training leads to mode-collapse. We find that, as a consequence, conditioning it on sufficiently detailed subject prompts will lead to a fixed identity across seeds and styles, as shown in fig. 4. We find this approach to achieve the best trade-off between generation time, identity consistency, and ability to change styles across the generated images. Hence, we use it to generate $500k$ images spanning roughly $100k$ identities. In practice, we trained for less than an epoch ($40k$ identities in total). Further experiments and comparisons on consistent data generation can be found in the supplementary.

4.5 Implementation Details

We initialize our IP-Adapter backbone using the CLIP-based Face-ID model, and tune the same parameters as the original IP Adapter using a learning rate of $1e-5$. The encoder and denoiser U-Nets are initialized from a pretrained SDXL model. We tune the encoder U-Net using LoRA with rank 4 and a learning rate of $5e-6$. The decoder is kept frozen. For our LCM-Lookahead we use TinyVAE [9] to decode the latents. This lighter model reduces our memory consumption and also improves gradient flow during backpropagation. We tune the models over $5,000$ iterations with a total batch size of 8 split across 2 NVIDIA A100 GPUs.

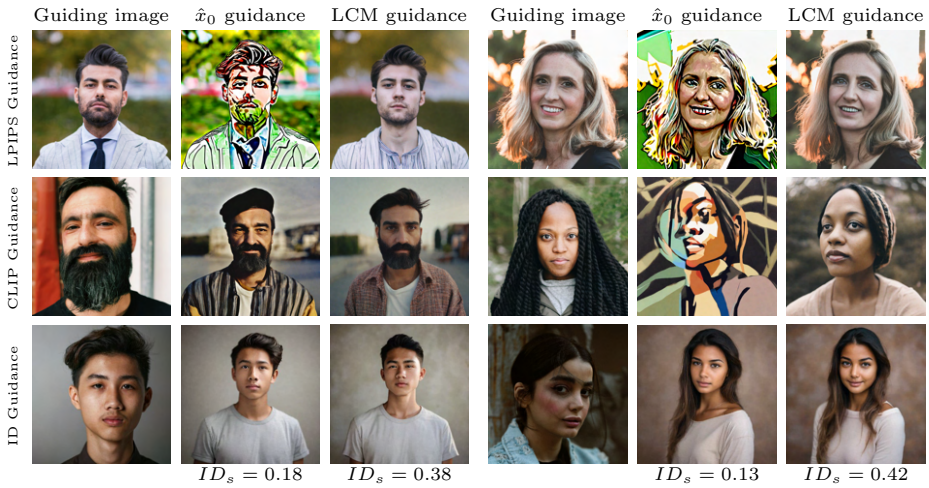

Fig. 5: **LCM-Lookahead Guidance.** Results of classifier guidance when using different classifiers on top of our LCM-Lookahead, or standard $\hat{x}_0$ approximation. Each classifier preserves different attributes of the guiding image. $\hat{x}_0$ guidance may result in reduced quality or visible artifacts. Identity similarity values (↑, measured using [28]) are shown at the bottom.

5 Experiments

5.1 Classifer guidance with LCM-Lookahead

We begin with an exploration of the LCM-Lookahead mechanism using a toy experiment, where we investigate its application to classifier guidance. Although this training-free method is less effective compared to encoder-based approaches, it serves as a simple use-case in which we can analyze the lookahead mechanism and discern its potential. Specifically, we follow [69] and apply repeated guidance iterations on an early diffusion time step ($t = 44$ out of 50 DDIM [62] steps). At each iteration, we denoise the current latents using SDXL-LCM, decode them to an image, apply our pixel-space loss, and backpropagate to modify the latents. Final results are generated by continuing unguided-DDIM sampling for the remaining diffusion steps. fig. 5 shows several outputs of such guidance, using a perceptual LPIPS loss [83], CLIP loss [51] and an identity loss [14]. Each loss preserves different attributes of the guiding image. LPIPS preserves the semantic structure of the image, CLIP preserves semantic attributes such as facial hair, and the identity loss explicitly improves facial similarity. When applying the same guidance to the single-step DDIM-approximations ($\hat{x}_0$ in fig. 5) we observe artifacts or reduced performance, attributed to the fact that $\hat{x}_0$ is typically blurry and discolored in early timesteps. The identity loss is particularly robust to blur, but its performance still improves by using a lookhead loss.

Our encoder focuses on the lookahead identity loss to improve identity preservation during training.

Fig. 6: **Qualitative results.** Our method personalizes a model to specific face identities at inference time, enabling both photo realistic and stylized prompts.

5.2 Encoder evaluation

fig. 6 shows a set of images synthesized using our encoder, across a range of identities and prompts. Our method can produce both photo-realistic results and stylized outputs, while largely preserving the subject's identity.

To better gauge the quality of our results, we evaluate our method against a set of prior and concurrent works on personalized face generation. Specifically, we consider the three leading tuning-free approaches which have public SDXL implementations: IP-Adapter [79], InstantID [70], and PhotoMaker [37]. For IP-Adapter, we use the FaceID model that serves as our backbone, and evaluate it using two adapter-scale settings: The ''official" value (1.0), and the one commonly used by the community (0.5) which significantly improves alignment with the textual prompts but harms identity.

Qualitative Comparison. We first conduct a qualitative comparison. Prior art often shows results on well-known celebrities. However, we find that it is excessively easy to overfit on such identities (indeed, SDXL already contains tokens describing them, relegating an encoder's job to simply finding these tokens).

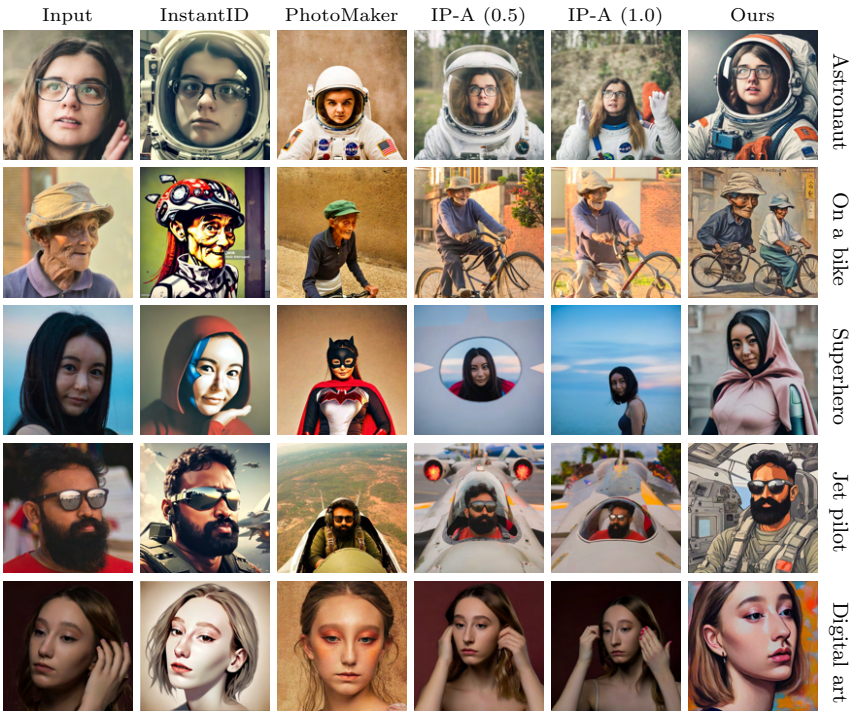

Fig. 7: Comparisons against prior and concurrent face-personalization encoders. IP-A (1.0) and (0.5) represent the IP-Adapter results with a scale of 1.0 and 0.5, respectively. Notably, IP-A (1.0) serves as the backbone which we fine-tune.

Moreover, some recent papers test on identities contained in their training set (*e.g.* LAION-Faces contains images of famous researchers such as Yann Lecun). To paint a full picture, we provide such comparisons in the supplementary. Here, we instead collect a small set of 50 images with permissive licenses which were uploaded to `https://unsplash.com/` over the period of Feb 19th - March 4th (2024). Our assumption is that these portray novel individuals which are less likely to exist in any prior training set, and thus offer the 'cleanest' benchmark.

Comparison results are shown in fig. 7. Notably, our method outperforms our IP-Adapter backbone in editability, while matching or exceeding it on identity preservation. InstantID achieves comparable results. However, it heavily restricts the target's pose to that of the conditioning image. Moreover, InstantID was trained on significantly more data and compute (60 million images and 48 GPUs over an undisclosed time-frame). We are hopeful that our approach could also be applied on top of an InstantID backbone and lead to improved results.

In the supplementary, we further investigate the extent of diverse styles that can be achieved by our approach. While it does not yet reach the expressiveness offered by recent optimization based approaches, we observe that our method can handle varied styles including non photo-realistic rendering or mosaicing.

Quantitative Comparison. Next, we move to quantitative evaluations. Here, we follow prior work [19] and compare the baselines across two metrics - identity similarity and prompt alignment. Our prompts include photo-realistic reconstructions, but also stylization and material change which prior art often struggles with. To measure identity similarity we use CurricularFace [28], which differs from both our loss network and from the backbones used to extract features for the baselines. Text alignment is measured using CLIP similarity (using the ViT-B/16 version which differs from the IP-Adapter backbone). We report both metrics using two sets: (1) $5,000$ identities randomly sampled from the FFHQ dataset [32] and (2) The 50 unsplash identities. The results are shown in table 1. Our method is situated on the pareto front, providing good identity preservation and high editability. Note in particular that we outperform the backbone IP-Adapter in its typically used setup ($\alpha = 0.5$) on both metrics.

Since identity metrics are highly sensitive to both the success of editing (*i.e.* stylized images will have lower identity scores than failed edits), and to poses (which InstantID copies), we also verify our results using a user-study. There, for each question we showed users a reference image and a prompt pair, and the outputs of two models conditioned on this pair (ours, and a random baseline). We asked users to select the image that better preserves the reference identity and better aligns with the prompt. In total, we collected 460 responses from 43 different users. In table 3 we report the percentage of users that preferred our method over each baseline. The results largely align with the automatic metrics, showing that our approach is preferred to the backbone on which we build. InstantID still outperforms all methods, primarily on account of its much improved editability, but the margin is less severe.

Table 1: Quantitative comparisons

	FFHQ-5000		Unsplash-50	
	ID ↑	CLIP-T ↑	ID ↑	CLIP-T ↑
Ours	0.345	26.33	0.308	26.79
IP-A (0.5)	0.268	25.82	0.250	26.36
IP-A (1.0)	0.368	21.39	0.387	22.06
PhotoMaker	0.344	26.69	0.218	27.19
InstantID	0.631	28.58	0.612	29.06

Table 2: Ablation study

	FFHQ-5000		Unsplash-50	
	ID ↑	CLIP-T ↑	ID ↑	CLIP-T ↑
Baseline IPA	0.368	21.39	0.387	22.06
+ Our Data	0.220	28.14	0.205	28.25
w/ ID-Loss x0	0.282	27.50	0.272	27.74
w/ ID-Loss LCM	0.301	27.31	0.281	27.74
+ KV Injection	0.345	26.33	0.308	26.79

Table 3: User study results. For each matchup, we report the fraction of users who prefered our method, and the fraction that preferred the baseline.

	PhotoMaker	IP-A (1.0)	IP-A (0.5)	InstantID
Ours	71.18%	82.25%	57.32%	44.06%
Baseline	28.82%	17.75%	42.68%	55.94%

5.3 Ablation

We conduct an ablation study to evaluate the contribution of our suggested components, using the following setups: (1) The baseline IP-Adapter tuned on

our data, without extended attention or identity losses. (2) The setup of (1) + an identity loss derived through standard $x0$ approximations. (3) The setup of (1) + an identity loss derived through an LCM shortcut. (4) Our full model (the setup of (3) + attention injection branch). For comparison, we again provide the backbone IP Adapter results.Quantitative results are provided in table 2. The use of our novel dataset greatly improves prompt alignment, at the cost of some identity fidelity. We attribute this in part to our smaller training scale, and to the fact that the identity score is impacted by stylization. Adding an identity loss significantly increases identity preservation, while passing this loss through the LCM shortcut provides a noticeably larger increase. Finally, injecting attention features leads to even better identity alignment, but at the cost of editability.

Our results demonstrate that appending losses through an LCM-shortcut mechanism can provide improvements over the direct approximation approach. All-in-all, the combination of our components leads to a high level of both identity preservation and prompt alignment.

6 Limitations and ethic concerns

While our model can improve existing encoders, it is not free of limitations. First, like prior tuning-free encoders, it falls short of the quality of optimization-based methods. This is particularly noticeable when working with inputs that are noticeably out-of-domain compared to standard face imagery (see supplementary).

Secondly, our model may still suffer from biases inherent in both the backbone that we built on, as well as the diffusion model itself. Hence, it may serve to amplify social biases. Moreover, facial editing and generation software can be used to spread disinformation or defame individuals. Existing detection tools can help mitigate such risks [72,13], and we hope that these continue to improve.

7 Conclusions

In this work, we presented LCM-Lookahead, a novel mechanism for applying image-space losses to diffusion training using a fast-sampling-based shortcut mechanism. We then build on top of this mechanism to provide better identity signal to a personalization encoder, leading to improved identity preservation. Our work further explores the shortcomings of current personalization encoders and proposes two additional techniques to further improve their results. First, we show that consistent data generation methods can greatly impact prompt-alignment quality, and that SDXL-Turbo can serve to create such data. Indeed, fine-tuning on such data can even restore editability to encoders which have overfit on the photo-realistic domain. Finally, we show that the common self-attention key-value injection mechanism can also be applied to encoder-based personalization, improving the faithfulness of the generated results. We hope that both our LCM-Lookahead and our improved training scheme will serve to further push the boundaries of text-to-image personalization.

Acknowledgements

This work was partially supported by ISF (grants 1337/22, 2492/20 and 3441/21).

References

1. State-of-the-art in the architecture, methods and applications of stylegan. In: Computer Graphics Forum. vol. 41, pp. 591–611. Wiley Online Library (2022)
2. Alaluf, Y., Garibi, D., Patashnik, O., Averbuch-Elor, H., Cohen-Or, D.: Cross-image attention for zero-shot appearance transfer (2023)
3. Alaluf, Y., Patashnik, O., Cohen-Or, D.: Restyle: A residual-based stylegan encoder via iterative refinement. arXiv preprint arXiv:2104.02699 (2021)
4. Alaluf, Y., Richardson, E., Metzer, G., Cohen-Or, D.: A neural space-time representation for text-to-image personalization. ACM Transactions on Graphics (TOG) **42**(6), 1–10 (2023)
5. Alaluf, Y., Tov, O., Mokady, R., Gal, R., Bermano, A.H.: Hyperstyle: Stylegan inversion with hypernetworks for real image editing (2021)
6. Arar, M., Gal, R., Atzmon, Y., Chechik, G., Cohen-Or, D., Shamir, A., H. Bermano, A.: Domain-agnostic tuning-encoder for fast personalization of text-to-image models. In: SIGGRAPH Asia 2023 Conference Papers. pp. 1–10 (2023)
7. Arar, M., Voynov, A., Hertz, A., Avrahami, O., Fruchter, S., Pritch, Y., Cohen-Or, D., Shamir, A.: Palp: Prompt aligned personalization of text-to-image models. arXiv preprint arXiv:2401.06105 (2024)
8. Avrahami, O., Aberman, K., Fried, O., Cohen-Or, D., Lischinski, D.: Break-a-scene: Extracting multiple concepts from a single image. In: SIGGRAPH Asia 2023 Conference Papers. SA '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3610548.3618154, https://doi.org/10.1145/3610548.3618154
9. Bohan, O.B.: Tiny autoencoder for stable diffusion. https://github.com/madebyollin/taesd (2023)
10. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 22560–22570 (October 2023)
11. Ceylan, D., Huang, C.H.P., Mitra, N.J.: Pix2video: Video editing using image diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23206–23217 (2023)
12. Chen, W., Hu, H., LI, Y., Ruiz, N., Jia, X., Chang, M.W., Cohen, W.W.: Subject-driven text-to-image generation via apprenticeship learning. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=wv3bHyQbX7
13. Corvi, R., Cozzolino, D., Zingarini, G., Poggi, G., Nagano, K., Verdoliva, L.: On the detection of synthetic images generated by diffusion models. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (2023). https://doi.org/10.1109/ICASSP49357.2023.10095167
14. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition (2019)
15. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems **34**, 8780–8794 (2021)

16. Dinh, T.M., Tran, A.T., Nguyen, R., Hua, B.S.: Hyperinverter: Improving stylegan inversion via hypernetwork. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11389–11398 (2022)
17. Dong, Z., Wei, P., Lin, L.: Dreamartist: Towards controllable one-shot text-to-image generation via contrastive prompt-tuning. arXiv preprint arXiv:2211.11337 (2022)
18. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion (2022). https://doi.org/10.48550/ARXIV.2208.01618, https://arxiv.org/abs/2208.01618
19. Gal, R., Arar, M., Atzmon, Y., Bermano, A.H., Chechik, G., Cohen-Or, D.: Encoder-based domain tuning for fast personalization of text-to-image models. ACM Transactions on Graphics (TOG) **42**(4), 1–13 (2023)
20. Gal, R., Patashnik, O., Maron, H., Chechik, G., Cohen-Or, D.: Stylegan-nada: Clip-guided domain adaptation of image generators. arXiv preprint arXiv:2108.00946 (2021)
21. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
22. Gu, J., Zhai, S., Zhang, Y., Liu, L., Susskind, J.M.: Boot: Data-free distillation of denoising diffusion models with bootstrapping. In: ICML 2023 Workshop on Structured Probabilistic Inference {\&} Generative Modeling (2023)
23. Han, L., Li, Y., Zhang, H., Milanfar, P., Metaxas, D., Yang, F.: Svdiff: Compact parameter space for diffusion fine-tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 7323–7334 (October 2023)
24. Hertz, A., Voynov, A., Fruchter, S., Cohen-Or, D.: Style aligned image generation via shared attention (2023)
25. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
26. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications (2021)
27. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Chen, W.: Lora: Low-rank adaptation of large language models. ArXiv **abs/2106.09685** (2021)
28. Huang, Y., Wang, Y., Tai, Y., Liu, X., Shen, P., Li, S., Li, J., Huang, F.: Curricularface: adaptive curriculum learning loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5901–5910 (2020)
29. Huang, Z., Wu, T., Jiang, Y., Chan, K.C., Liu, Z.: Reversion: Diffusion-based relation inversion from images. arXiv preprint arXiv:2303.13495 (2023)
30. Jia, X., Zhao, Y., Chan, K.C., Li, Y., Zhang, H., Gong, B., Hou, T., Wang, H., Su, Y.C.: Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. arXiv preprint arXiv:2304.02642 (2023)
31. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. Advances in Neural Information Processing Systems **35**, 26565–26577 (2022)
32. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
33. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., Shi, H.: Text2video-zero: Text-to-image diffusion models are zero-

shot video generators. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15954–15964 (October 2023)
34. Kim, D., Lai, C.H., Liao, W.H., Murata, N., Takida, Y., Uesaka, T., He, Y., Mitsufuji, Y., Ermon, S.: Consistency trajectory models: Learning probability flow ODE trajectory of diffusion. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=ymjI8feDTD
35. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. arXiv (2022)
36. Li, D., Li, J., Hoi, S.C.: Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. arXiv preprint arXiv:2305.14720 (2023)
37. Li, Z., Cao, M., Wang, X., Qi, Z., Cheng, M.M., Shan, Y.: Photomaker: Customizing realistic human photos via stacked id embedding. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
38. Liu, L., Ren, Y., Lin, Z., Zhao, Z.: Pseudo numerical methods for diffusion models on manifolds. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=PlKWVd2yBkY
39. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: DPM-solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In: Oh, A.H., Agarwal, A., Belgrave, D., Cho, K. (eds.) Advances in Neural Information Processing Systems (2022), https://openreview.net/forum?id=2uAaGwlP_V
40. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: DPM-solver++: Fast solver for guided sampling of diffusion probabilistic models (2023), https://openreview.net/forum?id=4vGwQqviud5
41. Luo, S., Tan, Y., Huang, L., Li, J., Zhao, H.: Latent consistency models: Synthesizing high-resolution images with few-step inference (2023)
42. Luo, S., Tan, Y., Patil, S., Gu, D., von Platen, P., Passos, A., Huang, L., Li, J., Zhao, H.: Lcm-lora: A universal stable-diffusion acceleration module (2023)
43. Mou, C., Wang, X., Song, J., Shan, Y., Zhang, J.: Dragondiffusion: Enabling drag-style manipulation on diffusion models. arXiv preprint arXiv:2307.02421 (2023)
44. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
45. Nitzan, Y., Bermano, A., Li, Y., Cohen-Or, D.: Face identity disentanglement via latent space mapping. ACM Transactions on Graphics (TOG) **39**(6), 1–14 (2020)
46. Peng, X., Zhu, J., Jiang, B., Tai, Y., Luo, D., Zhang, J., Lin, W., Jin, T., Wang, C., Ji, R.: Portraitbooth: A versatile portrait model for fast identity-preserved personalization. arXiv preprint arXiv:2312.06354 (2023)
47. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: Improving latent diffusion models for high-resolution image synthesis. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=di52zR8xgf
48. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=FjNys5c7VyY
49. Purushwalkam, S., Gokul, A., Joty, S., Naik, N.: Bootpig: Bootstrapping zero-shot personalized image generation capabilities in pretrained diffusion models (2024)
50. QI, C., Cun, X., Zhang, Y., Lei, C., Wang, X., Shan, Y., Chen, Q.: Fatezero: Fusing attentions for zero-shot text-based video editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15932–15942 (October 2023)

51. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
52. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
53. Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., Cohen-Or, D.: Encoding in style: a stylegan encoder for image-to-image translation. arXiv preprint arXiv:2008.00951 (2020)
54. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2021)
55. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation (2022)
56. Ruiz, N., Li, Y., Jampani, V., Wei, W., Hou, T., Pritch, Y., Wadhwa, N., Rubinstein, M., Aberman, K.: Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models (2023)
57. Ryu, S.: Low-rank adaptation for fast text-to-image diffusion fine-tuning. `https://github.com/cloneofsimo/lora` (2023)
58. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. In: International Conference on Learning Representations (2022), `https://openreview.net/forum?id=TIdIXIpzhoI`
59. Sauer, A., Lorenz, D., Blattmann, A., Rombach, R.: Adversarial diffusion distillation (2023)
60. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 (2021)
61. Shi, J., Xiong, W., Lin, Z., Jung, H.J.: Instantbooth: Personalized text-to-image generation without test-time finetuning. arXiv preprint arXiv:2304.03411 (2023)
62. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2020)
63. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models (2023)
64. Tewel, Y., Gal, R., Chechik, G., Atzmon, Y.: Key-locked rank one editing for text-to-image personalization. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–11 (2023)
65. Tewel, Y., Kaduri, O., Gal, R., Kasten, Y., Wolf, L., Chechik, G., Atzmon, Y.: Training-free consistent text-to-image generation (2024)
66. Tov, O., Alaluf, Y., Nitzan, Y., Patashnik, O., Cohen-Or, D.: Designing an encoder for stylegan image manipulation. arXiv preprint arXiv:2102.02766 (2021)
67. Valevski, D., Lumen, D., Matias, Y., Leviathan, Y.: Face0: Instantaneously conditioning a text-to-image model on a face. In: SIGGRAPH Asia 2023 Conference Papers. SA '23, Association for Computing Machinery, New York, NY, USA (2023). `https://doi.org/10.1145/3610548.3618249`, `https://doi.org/10.1145/3610548.3618249`
68. Voynov, A., Chu, Q., Cohen-Or, D., Aberman, K.: $p+$: Extended textual conditioning in text-to-image generation. arXiv preprint arXiv:2303.09522 (2023)
69. Wallace, B., Gokul, A., Ermon, S., Naik, N.: End-to-end diffusion latent optimization improves classifier guidance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 7280–7290 (October 2023)

70. Wang, Q., Bai, X., Wang, H., Qin, Z., Chen, A.: Instantid: Zero-shot identity-preserving generation in seconds. arXiv preprint arXiv:2401.07519 (2024)
71. Wang, T., Zhang, Y., Fan, Y., Wang, J., Chen, Q.: High-fidelity gan inversion for image attribute editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
72. Wang, Z., Bao, J., Zhou, W., Wang, W., Hu, H., Chen, H., Li, H.: Dire for diffusion-generated image detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 22445–22455 (October 2023)
73. Wei, Y., Zhang, Y., Ji, Z., Bai, J., Zhang, L., Zuo, W.: Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15943–15953 (October 2023)
74. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7623–7633 (2023)
75. Wu, Z., Nitzan, Y., Shechtman, E., Lischinski, D.: Stylealign: Analysis and applications of aligned stylegan models (2021)
76. Xia, W., Zhang, Y., Yang, Y., Xue, J.H., Zhou, B., Yang, M.H.: Gan inversion: A survey (2021)
77. Xiao, G., Yin, T., Freeman, W.T., Durand, F., Han, S.: Fastcomposer: Tuning-free multi-subject image generation with localized attention. arXiv (2023)
78. Yan, Y., Zhang, C., Wang, R., Cheng, P., Yu, G., Fu, B.: Facestudio: Put your face everywhere in seconds (2023)
79. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
80. Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T., Park, T.: One-step diffusion with distribution matching distillation. CVPR (2024)
81. Yuan, G., Cun, X., Zhang, Y., Li, M., Qi, C., Wang, X., Shan, Y., Zheng, H.: Inserting anybody in diffusion models via celeb basis. arXiv preprint arXiv:2306.00926 (2023)
82. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
83. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)