

FASTFACE: TRAINING-FREE IDENTITY PRESERVATION TUNING IN DISTILLED DIFFUSION VIA GUIDANCE AND ATTENTION

Anonymous authors

Paper under double-blind review

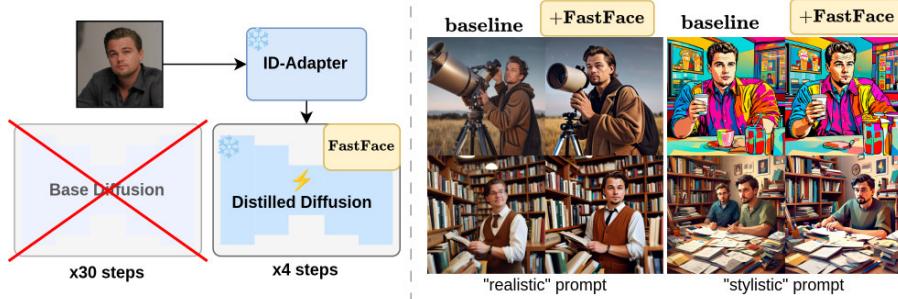


Figure 1: FastFace method framework: on the left - high-level idea of pipeline, enabling few-step id-preserving generation, on the right - effect of FastFace components on realistic and stylistic generations

ABSTRACT

The recent proliferation of identity-preserving (ID) adapters has significantly advanced personalized generation with diffusion models. However, these adapters are predominantly co-trained with base diffusion models, inheriting their critical drawback: slow, multi-step inference. This work addresses the challenge of adapting pre-trained ID adapters to much faster distilled diffusion models without requiring any further training. We introduce FastFace, a universal framework that achieves this via two key mechanisms: (1) the decomposition and adaptation of classifier-free guidance for few-step stylistic generation, and (2) attention manipulation within decoupled blocks to enhance identity similarity and fidelity. We demonstrate that FastFace generalizes effectively across various distilled models and maintains full compatibility with a wide range of existing ID-preserving methods, enabling high-fidelity personalized image generation at unprecedented speeds.

1 INTRODUCTION

Diffusion models have emerged as a dominant paradigm in generative modeling, achieving state-of-the-art performance in high-fidelity image synthesis, with plethora of models coming out in recent years (Ho et al. (2020), Dhariwal & Nichol (2021), Rombach et al. (2022), Podell et al. (2023), Esser et al. (2024), Labs (2024)). Diffusion distillation aims to speed up inference by reducing number of sampling steps, with a lot of approaches and versions releasing in past several years, such as LCM, Turbo, Lightning, Hyper, and others (Luo et al. (2023), Sauer et al. (2024a), Lin et al. (2024c), Ren et al. (2024), Sauer et al. (2024b)); common results of these distillation are 1) architecture of diffusion model remains the same 2) inference becomes significantly more efficient in terms of number of steps. In parallel, diffusion models have been adapted for task of id-preserving generation, where image with face of a person c_{id} is used as condition, and diffusion can generate images with novel identities without further finetuning (Ye et al. (2023), Li et al. (2024), Wang et al. (2024a), Guo et al. (2024), Jiang et al. (2025)). These methods are commonly trained with base diffusion models and can be denoted as ID-adapters.

Integrating image conditioning with distilled diffusion has recently emerged as a separate problem. Several works (Zhang et al. (2023), Xiao et al. (2023), Parmar et al. (2024)) propose to adapt prior of ControlNet towards new trajectories of distilled models through finetuning. While promising, these works do not propose a universal approach to any model, i.e. for a new model, a completely new algorithmic design is required. In recent work Sun et al. (2024) authors introduce training-free adaptation of image conditioning through classifier guidance, but their approach suffers from requirement for backpropagation during inference.

In contrast to previous work, we aim to develop general, light, training-free mechanisms that can be used in plug-and-play manner to improve quality of id-preserving generation with any distilled diffusion model. We separate identity-preserving generation scenarios and develop two contributions - for stylistic generation we adopt and tune decoupled classifier-free guidance, where conditional noise prediction is splitted into two parallel terms, and for realistic id preservation we introduce attention manipulation in decoupled blocks is transformed via simple analytical functions to be more focused on facial regions during generation. We denote FastFace as joint application of these mechanisms, and it's effect on generated image is visualized in Figure 7 - it achieves superior identity preservation and image fidelity while not losing prompt following.

To demonstrate the effectiveness and generality of FastFace, we conduct a comprehensive evaluation across a range of SDXL-distilled checkpoints and identity adapters. Our empirical results, which demonstrate robust generalization, are enabled by two technical contributions:

Decoupled Classifier-Free Guidance Mechanism: We introduce a guidance strategy that decomposes the network output into semantically interpretable components. This decomposition is specifically tuned to enhance performance in the few-step sampling regime characteristic of distilled models.

Attention Manipulation for Identity Enhancement: We develop an inference-time method to precisely manipulate attention maps within decoupled attention blocks. By strategically reinforcing attention over facial regions, this approach substantially improves identity similarity without any additional training.

We evaluate our method on subset of high-quality OmniContext dataset (Wu et al. (2025)), as well as collected and released DiverseFaces benchmark, which includes a wide variety of nationalities, age groups, and genders and decomposes scenarios of stylistic and realistic generation, allowing to use both for specific ablations and general evaluation. The results confirm the superior performance and broad applicability of FastFace for efficient identity-preserving generation.

2 RELATED WORK

ID-preserving generation methods Identity-preserving generation, as we describe it, is a problem of preserving identity similarity in generation output given an image with the face. A lot of methods came out around this problem, including IpAdapter-FaceID Ye et al. (2023), Photomaker Li et al. (2024), PuLID Guo et al. (2024), InstantID Wang et al. (2024a). They differ in their overall approaches and flexibility, with later methods building on top of FaceID, however, id-adapters trained for new diffusion models frequently rely on conventional FaceID approach and codebase (Team (2024)). Another group of methods such as DreamBooth Ruiz et al. (2023) and similar are also applicable to this problem, however, they are heavily limited due to need for finetuning for each new identity.

Diffusion distillation Diffusion distillation is an approach to accelerate trained diffusion models by training them to sample in few steps while still trying to model original $p_{data}(x)$ as close as possible (Salimans & Ho (2022), Song et al. (2023), Yin et al. (2024)). State of the art approaches such as LCM Luo et al. (2023) and Hyper Ren et al. (2024) remain common for new model releases (Ke et al. (2024), Chen et al. (2024)), but new distillation techniques are actively being developed. In practice, these distilled versions may differ in their inference qualities and sampling procedures, generally applicable in range of 1-8 sampling steps. Application of these distilled models to image conditioned generation and in particular id-preserving generation is at the heart of this work.

Adaptation to new diffusion models Cheap adaptation of pretrained modules for diffusion models to new checkpoints has been explored in Lin et al. (2024a) authors train an adapter module that

acts as a latent projection between the inner-layer connection of the original ControlNet and new diffusion model and they achieve fast generalization. In other work Xu et al. (2024) authors consider a case of efficient adaptation of ControlNet to new conditional domains. In the context of distilled diffusion models similar problems have also been explored with ControlNet: (Xiao et al. (2023), Parmar et al. (2024)), where in both works specific finetuning approaches are proposed either to match distillation objective or enforce cycle-consistency. Limitations of available solutions are either designing finetuning approach per checkpoint or tolerating baseline quality. We show that it is possible to universally boost quality of such adaptation without any additional training.

3 METHOD

3.1 BACKGROUND

We highlight three design choices that are commonly used in construction and training of ID-adapters. Firstly, Eq. 1 describes information flow introduced through additional cross-attention blocks. These additional blocks are called decoupled and introduce new keys and values K' and V' (for details see Ye et al. (2023)). Importantly, a single scalar value λ controls the input of visual information from face, and we will address this in Section 3.3.

$$z_{new} = \text{Attn}(z; Q, K, V) + \lambda \cdot \text{Attn}(z; Q, K', V') \quad (1)$$

Secondly, a commonly known technique classifier-free guidance is adopted for two distinct conditions as described in Eq. 2. x_t is omitted for clarity. It can be seen that scale parameter w impacts both conditioning strength on c_{text} and c_{id} , not allowing any flexibility given two distinct conditions.

$$\hat{\epsilon} = \epsilon(\emptyset, \emptyset) + w \cdot (\epsilon(c_{text}, c_{id}) - \epsilon(\emptyset, \emptyset)) \quad (2)$$

Lastly, we note that training id-preserving adapters trained via diffusion target (Ho et al. (2020)) are proximally trained to reconstruct identity in the image, i.e. maximize similarity between the person in c_{id} and $\hat{x}_0$ given conditional information. However, during inference, this is not strictly the case. Given pretrained ID-adapter we identify two common generation purposes - *stylistic* and *realistic*. By "stylistic" we define generation that implies visual domain shift towards some priory known style, e.g. "pixel art" implies that generated image is expected to follow pixel-like visual appearance; by "realistic" setup we define generation that is not biased to any specific style or biased explicitly towards "realism" as specified in prompt. These cases correspond to different goals - in "stylistic" the user is less interested in facial features similarity and more in style following, while in realism situation is opposite, examples given in Figure 2.



Figure 2: Different cases of user intention during ID-preserving generation: (a) - stylistic, (b) - realistic

In the following section we built components of FastFace on top of described design choices, and then unify them in one framework.

3.2 DECOUPLED CLASSIFIER FREE GUIDANCE

CFG Decomposition Classifier-free guidance decomposition was introduced in Brooks et al. (2023) in the following form: $\hat{\epsilon} = \epsilon(\emptyset, \emptyset) + \alpha \cdot (\epsilon(c_{img}, \emptyset) - \epsilon(\emptyset, \emptyset)) + \beta \cdot (\epsilon(c_{text}, c_{img}) - \epsilon(c_{img}, \emptyset))$. In this equation we split guidance between c_{text} and c_{img} which enables to tune between preservation and editing in instruct editing task. However it remains understudied, since it has

not been widely adopted in later works (Labs et al. (2025), Zhang et al. (2025)) and has not applied before for generation with ID-adapters.

In identity-preserving generation we find same decoupling also works, while instead of c_{img} we use c_{id} with reference identity. We give additional possible derivation and ablations of this expression in in Appendix D. In this setup α corresponds to strength of id conditioning and β corresponds to textual strength conditioning, however it is not yet applicable to distilled models.

DCG - few-step stylistic tuning Simply substituting c_{id} and changing α and β will result in degradation and artifacts for distilled models not suited for guidance, see Appendix D for examples. Therefore we introduce two contributions to make it work with distilled models. Firstly we ablate scheduling regimes of conventional classifier-free guidance for our setup. Contrary to findings in previous works (Wang et al. (2024b), Starodubcev et al. (2024)) we find that scheduling should be applied only to intermediate steps, see Appendix D, and use it to bias model towards prompt following.

Secondly, to further enhance visual quality we apply rescaling to decoupled terms to balance norms of output and predicted noises, and , which is inspired by rescaling trick introduced in Lin et al. (2024b). Overall algorithm of DCG is given in equations below - in second expression σ_i and σ_{ti} correspond to standard deviation of corresponding conditional predictions $\epsilon(c_{id}, \emptyset)$ and $\epsilon(c_{text}, c_{id})$, and this deviations are averaged, last equation introduces interpolation trade-off between stability and quality, scaling hyper-parameter ϕ in practice is fixed.

$$\hat{\epsilon} = \epsilon(\emptyset, \emptyset) + \underbrace{\alpha(t) \cdot (\epsilon(c_{id}, \emptyset) - \epsilon(\emptyset, \emptyset))}_{\text{id guidance}} + \underbrace{\beta(t) \cdot (\epsilon(c_{text}, c_{id}) - \epsilon(c_{id}, \emptyset))}_{\text{text guidance}} \quad (3)$$

$$\epsilon_{\text{rescaled}} = \frac{\sigma_i + \sigma_{ti}}{2\hat{\sigma}} \hat{\epsilon}, \quad (4)$$

$$\epsilon_{\text{DCG}} = \phi \cdot \epsilon_{\text{rescaled}} + (1 - \phi) \epsilon_{\text{dclg}} \quad (5)$$

After applying proposed changes, we can choose coefficients α and β in wider range as demonstrated in Figure 3a. We choose $\alpha(t) = [1.0, 1.5, 1.5, 1.0]$ and $\beta(t) = [1.0, 3.0, 3.0, 1.0]$ and find them applicable to all studied checkpoints with "style" prompts, as it enhances coherence with described style at low-level details additional ablation with stylistic part of DiverseBench is given in Table 1.



Figure 3: DCG visualizations; (a) - ablation of coefficient demonstrate visual trade-off between tuning α and β , (b) - visual result of tuned DCG added to inference across different distilled checkpoints

guidance setup	CLIP ($\uparrow$)	IR ($\uparrow$)	ID ($\uparrow$)
const(baseline)	0.268	0.901	0.258
schedule	<u>0.276</u>	<u>1.278</u>	0.320
schedule + rescale	0.277	1.289	<u>0.318</u>

Table 1: DCG components ablation with stylistic generation

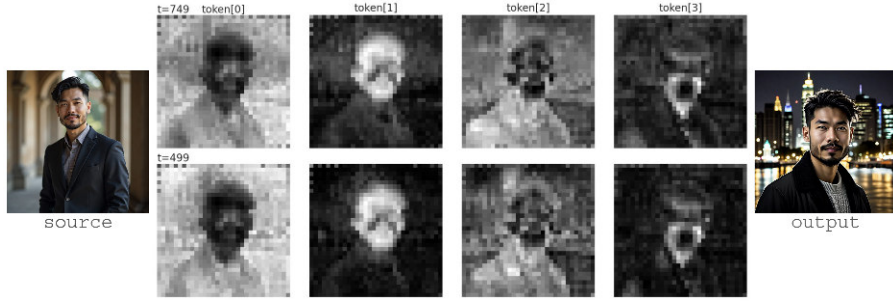


Figure 4: Visualization of attention maps timesteps 749 and 499 in decoupled block of SDXL in relation to generation output

3.3 ATTENTION MANIPULATION

Motivation Attention maps in diffusion models are known to contain a lot of semantic and spatial information, which has been applied in numerous works of image editing (Hertz et al. (2022), Cao et al. (2023), Epstein et al. (2023), Titov et al. (2024)). Nuance of ID-adapters is that they train new cross-attention blocks within UNet to condition on visual information from c_{id} . We inspect these new blocks and visualize attention maps in Figure 4 - it can be seen that they share a lot of information with facial features and position in generated images, while also containing a lot of noisy signal about surrounding context, which can't be removed by changing `ip_adapter_scale` (see Eq. 1). Therefore we opt to work with attention maps directly.

Basic formulation We begin with formulation of general Attention Manipulation (AM) algorithm in Equation 6. Main challenge is to construct such $f(\cdot) : A \mapsto \tilde{A}$, where A is attention map in decoupled blocks, that $\tilde{A}$ would allow achieve properties of 1) increasing face similarity/fidelity without significantly damaging prompt following 2) steering id-preserving generation towards more stable results, which we achieve by *focusing attention on face regions*.

$$\text{softmax}\left(\frac{Q(K')^T}{\sqrt{d}}\right) \rightarrow f(A) \rightarrow \tilde{A}V' \rightarrow z \quad (6)$$

Scale-power transform First transformation is designed via simple composition of scale and power transform applied to attention maps. Detailed ablation of this operations is given in Appendix E, intuitively power transformation applied to values less than 1 shifts everything closer to 0, while scaling linearly enhances attention mainly in meaningful tail of distribution with face region.

$$f_{sp} := (\text{scale} \circ \text{power})(A) = s \cdot A^p \quad (7)$$

Steering scheduled-softmax transform Second transformation is designed in more tricky way to steer generation towards more stable, portrait-like images on average. This purpose is motivated by presence of "failure" cases, where for some reason id-preserving generation deviates towards unrealistic imagery or fails to preserve features in meaningful way, therefore requiring more global transformation, examples are given in Appendix E. It is constructed of following components 1) firstly Equation 9 performs an adaptive distribution shift of values less than $Q_p(A)$ towards 0 and others towards 1, strength of shift is defined by parameter d 2) d is scheduled to large value at first step to influence global structure of the image 3) smooth alignment with original attention statistics inspired by AdaIN Huang & Belongie (2017) is applied - normalizing transformed attention maps, modulate them using μ_A and σ_A of original maps and interpolate between modulated and transformed versions, same operation is also applied to output of attention block. Complete definition of $f_{ss}()$ is given in Equations below.

$$F(A, p) := \text{norm}(A) - Q_p(A) \quad (8)$$

$$\text{SM}(A, d, p) := s \cdot \sigma(\text{norm}(\sigma(-dF(A, p)))) \quad (9)$$

$$f_{ss}(A) := ws \cdot \text{SM}(A, d, p) + (1 - w)\text{AdaIN}(A, s \cdot \text{SM}(A, d, p)) \quad (10)$$

In Equation 9 $\text{norm}(\cdot)$ denotes normalization and $Q_p(\cdot)$ is a p-th quantile function. In Figure 5 we visualize effect transforms have on attention values. We validate that proposed transforms achieve desired effect of enhancing identity similarity independently of face size in Figure 6.

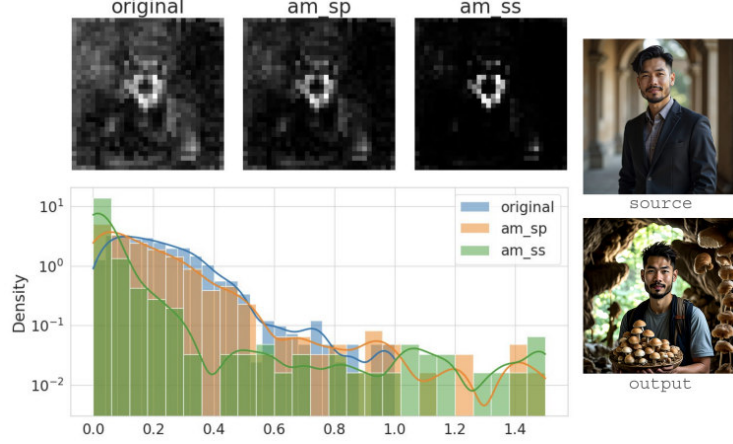


Figure 5: Visualizations of f_{sp} and f_{sm} transforms. At the top - visual result of transformation on the level of attention maps at certain block/step/token, bottom - distribution shift of attention values

We as well provide detailed ablations and visual results for AM setups, as well as sensitivity analysis in Appendix E and G.

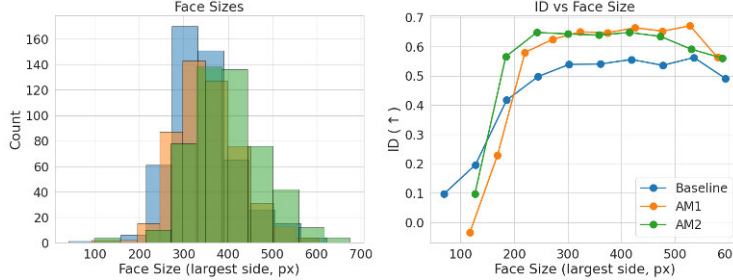


Figure 6: left - histogram of face sizes in baseline and different AM configurations, right - distribution of face sizes values in different setups (AM1 - scale-power, AM2 - scheduled-softmax)

3.4 FULL FRAMEWORK AND EVALUATION

Together, presented mechanisms formulate joint framework of FastFace - through use of DCG and AM, which can be applied together or independently to any few-step sampling models, and are visualized in Figure 7. In further sections we will demonstrate that these mechanisms work well together in general setting of id-preserving generation, as well as their respective setups of stylistic/realistic generations.

Open evaluation In recent works (Ye et al. (2023), Wang et al. (2024a), Guo et al. (2024)) authors rarely provide clarity about data which was used for evaluation, not allowing to fairly compare one method quality to another and understand their strength and weaknesses with respect to realistic or stylized generation with detailed prompts. When evaluating these methods we find that in a lot

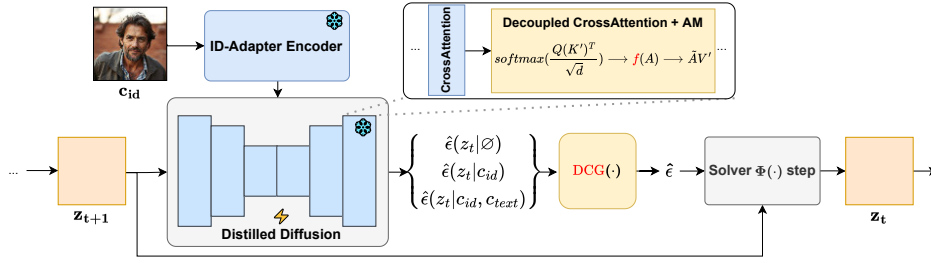


Figure 7: FastFace joint pipeline with proposed mechanisms - decoupled classifier free guidance, expanding on outputs of UNet, and attention manipulation as transform in decoupled blocks

of cases they fail completely in practical setting, see Appendix C. In this work stick to transparent evaluation in terms of identity images and prompts used.

Dataset details We collect a synthetic dataset DiverseBench consisting of 54 high-resolution identity images from several models, ensuring diversity and filtering by mean similarity threshold within identity groups. Prompts are constructed for two settings - 80 for realistic and 40 for stylistic. During evaluation, fixed amount of random pairs without replacement is sampled, covering a wide range of identity and prompts combinations. Additionally we use a subset from OmniContext dataset, which corresponds to single character generation with detailed prompts. Detailed description of collection and processing of DiverseBench is given in Appendix B.

4 EXPERIMENTS

In following experiments we try to answer two questions - "does FastFace generalize across models" and "does FastFace generalize across ID-adapters" without specific tuning of parameters per model/adapter. We do this by comparing against multiple known baseline which are attached to same checkpoints, and support presented metrics with multiple visual examples.

4.1 METRICS

Common metrics Metrics applied in both setups are face-similarity (ID), estimated as cosine distance between embeddings extracted by `buffalo-l` backbone from faces in source and generated images, and CLIP score (CLIP) between generated images and prompt computed with `clip/l-14` to estimate prompt alignment. We use LAION-Aesthetic (AE) reward model, which was trained on LAION subset, to estimate image quality/fidelity of image in both general evaluation and with realistic subset LAION (2022). Additionally we use ImageReward (IR) reward model to measure quality of stylistic images in stylistic ablations - as it was trained on synthetic data and is biased towards colorfulness and details and is more suited for that setting Xu et al. (2023). We account for `face_fail_cnt` (FFC) - an integer metric which value represents amount of cases where no face was detected in generated image - measure of how unstable certain method is.

4.2 RESULTS

Below in Table 2 we present evaluation of FastFace framework applied with FaceID-Plus-v2, as well as other common ID-adapters applied with different checkpoints. A more technical ablation over this and other checkpoints can be found in Appendix H. FastFace allows to achieve superior identity preservation, while achieving CLIP and AE metrics comparable with less constrained methods like PuLID and Photomaker and not requiring additional compute like RectifID.

In Figure 8 we present result of applying FastFace framework for both realistic and stylistic generation results. Beyond main metrics given in, FastFace is able to introduce stability into resulting face generation. We additionally analyze sensitivity of FastFace hyperparameters in Appendix H and show that it is robust across wide range of values.

Table 2: Metric comparison of baseline setup against FastFace setups - FF_{AM1} denotes application of DCG with scale-power transform, FF_{AM2} - DCG with scheduled-softmax transform

Model	Method	DiverseFaces				OmniContext				time(sec)
		ID ($\uparrow$)	CLIP ($\uparrow$)	AE ($\uparrow$)	FFC ($\downarrow$)	ID ($\uparrow$)	CLIP ($\uparrow$)	AE ($\uparrow$)	FFC ($\downarrow$)	
SDXL-Hyper	FaceID	0.536	0.267	5.661	1	0.643	0.268	5.422	1	2.32 $\pm$ 0.03
	FaceID-Portrait	0.354	0.259	5.736	25	0.581	0.251	5.328	0	2.13 $\pm$ 0.01
	FaceID-Plus-v2	0.588	0.259	5.889	0	0.646	0.269	5.665	0	2.41 $\pm$ 0.04
	PuLID	0.238	0.262	5.902	72	0.335	0.277	5.828	0	2.56 $\pm$ 0.38
	Photomaker	0.127	0.267	5.691	80	0.155	0.281	5.658	0	1.50 $\pm$ 1.38
	RectifID*	0.369	0.263	5.322	6	0.283	0.288	5.309	2	16.87 $\pm$ 4.73
	FastFace-AM1 (ours)	0.623	0.260	5.854	0	0.683	0.264	5.591	0	2.43 $\pm$ 0.06
	FastFace-AM2 (ours)	<u>0.590</u>	0.242	5.882	1	0.627	0.260	<u>5.704</u>	0	2.42 $\pm$ 0.05
SDXL-Lightning	FaceID	0.474	0.265	5.545	1	0.593	0.264	5.341	0	
	FaceID-Portrait	0.341	0.256	5.661	18	0.515	0.256	5.366	0	
	FaceID-Plus-v2	0.517	0.259	5.715	0	0.590	0.271	5.511	0	
	PuLID	0.218	0.257	5.752	154	0.295	0.273	5.609	1	
	Photomaker	0.093	0.267	5.649	86	0.141	0.282	5.580	0	
	FastFace-AM1 (ours)	0.568	0.261	<u>5.747</u>	0	0.626	0.267	5.568	0	
	FastFace-AM2 (ours)	<u>0.553</u>	0.236	5.699	2	<u>0.589</u>	0.258	5.477	0	

* evaluated as in original paper with 'perflow-sd15-dreamshaper'



Figure 8: Demonstration of application of framework to real based identities from evaluation set

4.3 BEYOND FACEID-PLUS-V2

To further study generalization of proposed framework, we additionally evaluate it with standard FaceID and also try applying to one of the recent ID-adapters - PuLID. In Table 3 we numerically account for contribution of proposed methods towards better trade offs, and visual examples are given in Fig. 9. Note that PuLID generally finds current evaluation prompts challenging to follow while preserving any ID, but proposed framework still allows to boost methods quality.

5 CONCLUSION

This work presents lightweight and easy-to-implement FastFace framework, which solves problem of adaptation of pretrained id-preserving generation adapter to distilled diffusion model without additional retraining. Included methods are developed for different cases of id-preserving generation - "stylistic", to better match style described in prompt, and "realistic", to enhance identity similarity or fidelity of the image. Presented contributions are evaluated in general, as well in specific scenarios on constructed evaluation dataset for id-preserving generation, showing generally better trade-offs in terms of identity preservation, prompt following and image quality.

6 LIMITATIONS

Although proposed methods show promising results, scope of current work is limited to training-free methods, which are ultimately bottle-necked by distilled diffusion model checkpoint, and generally shows less impressive results in extreme cases such single-step sampling regime. It is a future work matter to address these limitation and adapt id-preserving generation to single-step models.

Model	ID ($\uparrow$)	CLIP ($\uparrow$)	AE ($\uparrow$)
Hyper + FaceID			
base	0.580	0.251	6.047
FF_{AM1}	0.595	0.258	6.125
Lightning + FaceID			
base	0.508	0.246	6.002
FF_{AM1}	0.535	0.258	6.102
Hyper + PuLID			
base	0.179	0.262	6.207
FF_{AM1}	0.227	0.261	6.261
Lightning + PuLID			
base	0.172	0.258	6.096
FF_{AM1}	0.228	0.254	6.254

Table 3: Evaluation results for other id-adapters

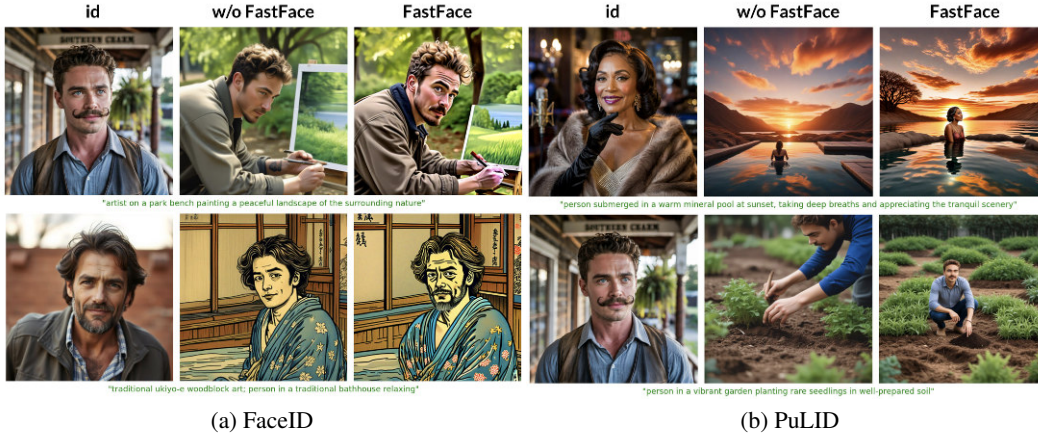


Figure 9: Application of FastFace framework to other ID-adapters

7 REPRODUCIBILITY

This work ensures reproducibility through three primary measures: (1) detailed algorithmic descriptions within the manuscript, (2) release of evaluation datasets, and (3) the use of fixed random seeds to ensure deterministic experimental outcomes.

8 ETHICAL STATEMENT

This work presents methods for generating and manipulating human faces. All facial images used for evaluation in our benchmark, DiverseFaces, are synthetically generated. However, we acknowledge that the ability to realistically modify human likenesses carries inherent risks, including the potential for misuse to create misleading or harmful content. To mitigate this, we have chosen to use only high-quality and diverse data for our benchmark, avoiding the use of real individuals' likenesses without explicit consent. We strongly advocate for the responsible development and use of such technologies, including the implementation of robust safeguards, provenance tracking, and public education to prevent misuse.

REFERENCES

- Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18392–18402, 2023.
- Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 22560–22570, 2023.
- Junsong Chen, Yue Wu, Simian Luo, Enze Xie, Sayak Paul, Ping Luo, Hang Zhao, and Zhenguo Li. Pixart- δ : Fast and controllable image generation with latent consistency models. *arXiv preprint arXiv:2401.05252*, 2024.
- Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.
- Dave Epstein, Allan Jabri, Ben Poole, Alexei Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. *Advances in Neural Information Processing Systems*, 36:16222–16239, 2023.
- Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.
- Zinan Guo, Yanze Wu, Chen Zhuowei, Peng Zhang, Qian He, et al. Pulid: Pure and lightning id customization via contrastive alignment. *Advances in neural information processing systems*, 37:36777–36804, 2024.
- Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020.
- Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 1501–1510, 2017.
- Deep Insight. Insightface: 2d and 3d face analysis project. <https://github.com/deepinsight/insightface>, 2023.
- Liming Jiang, Qing Yan, Yumin Jia, Zichuan Liu, Hao Kang, and Xin Lu. Infinityyou: Flexible photo recrafting while preserving your identity. *arXiv preprint arXiv:2503.16418*, 2025.
- Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024.
- Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.
- LAION. Aesthetic model predictor - GitHub repository. <https://github.com/LAION-AI/aesthetic-predictor>, 2022.

- Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8640–8650, 2024.
- Han Lin, Jaemin Cho, Abhay Zala, and Mohit Bansal. Ctrl-adapter: An efficient and versatile framework for adapting diverse controls to any diffusion model. *arXiv preprint arXiv:2404.09967*, 2024a.
- Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 5404–5411, 2024b.
- Shanchuan Lin, Anran Wang, and Xiao Yang. Sdxl-lightning: Progressive adversarial diffusion distillation. *arXiv preprint arXiv:2402.13929*, 2024c.
- Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.
- Gaurav Parmar, Taesung Park, Srinivasa Narasimhan, and Jun-Yan Zhu. One-step image translation with text-to-image models. *arXiv preprint arXiv:2403.12036*, 2024.
- Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.
- Yuxi Ren, Xin Xia, Yanzuo Lu, Jiacheng Zhang, Jie Wu, Pan Xie, Xing Wang, and Xuefeng Xiao. Hyper-sd: Trajectory segmented consistency model for efficient image synthesis. *arXiv preprint arXiv:2404.13686*, 2024.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.
- Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22500–22510, 2023.
- Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.
- Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024a.
- Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pp. 87–103. Springer, 2024b.
- Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.
- Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. 2023.
- Nikita Starodubcev, Mikhail Khoroshikh, Artem Babenko, and Dmitry Baranchuk. Invertible consistency distillation for text-guided image editing in around 7 steps. *arXiv preprint arXiv:2406.14539*, 2024.
- Zhicheng Sun, Zhenhao Yang, Yang Jin, Haozhe Chi, Kun Xu, Kun Xu, Liwei Chen, Hao Jiang, Yang Song, Kun Gai, and Yadong Mu. Rectifid: Personalizing rectified flow with anchored classifier guidance. In *Advances in Neural Information Processing Systems*, 2024.
- Kolors Team. Kolors: Effective training of diffusion model for photorealistic text-to-image synthesis. *arXiv preprint*, 2024.

- Vadim Titov, Madina Khalmatova, Alexandra Ivanova, Dmitry Vetrov, and Aibek Alanov. Guide-and-rescale: Self-guidance mechanism for effective tuning-free real image editing. In *European Conference on Computer Vision*, pp. 235–251. Springer, 2024.
- Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024a.
- Xi Wang, Nicolas Dufour, Nefeli Andreou, Marie-Paule Cani, Victoria Fernández Abrevaya, David Picard, and Vicky Kalogeiton. Analysis of classifier-free guidance weight schedulers. *arXiv preprint arXiv:2404.13040*, 2024b.
- Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025.
- Jie Xiao, Kai Zhu, Han Zhang, Zhiheng Liu, Yujun Shen, Yu Liu, Xueyang Fu, and Zheng-Jun Zha. Ccm: Adding conditional controls to text-to-image consistency models. *arXiv preprint arXiv:2312.06971*, 2023.
- Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.
- Yifeng Xu, Zhenliang He, Shiguang Shan, and Xilin Chen. Ctrlora: An extensible and efficient framework for controllable image generation. *arXiv preprint arXiv:2410.09400*, 2024.
- Hu Ye, Jun Zhang, Sibio Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.
- Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6613–6623, 2024.
- Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023.
- Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. In-context edit: Enabling instructional image editing with in-context generation in large scale diffusion transformer. *arXiv preprint arXiv:2504.20690*, 2025.

A APPENDIX

B DETAILS OF EVALUATION DATASET

We develop an evaluation dataset consisting of 54 high quality identity images and 120 prompts, which are used as input conditions for generation and further evaluation. Identity images are synthetic images from models such as Flux and Ideogram 3.0 (Labs (2024)), representing different age groups (young, middle age and old), genders and ethnicities, examples are presented in Figure 11. Part of images was also synthesized using id-preserving methods with from real identities, thus avoiding bias towards only synthetic facial features. Additionally, to ensure variance within groups of identities of same gender and age, further cleaning was done by thresholding and replacing identity images with largest mean face similarity to others, i.e. if $\frac{1}{n-1} \sum_{j,j \neq i} \text{sim}(c_i, c_j) > 0.3$ for c_i within group, it was discarded. Prompt description were also synthetically generated using ChatGPT version of November 2024, generally following structure of `style + ';' + 'Person' + location + action`, and then additionally cleared and enriched. Prompts are categorized into two groups - 80 "realistic" prompts and 40 "style" prompts with certain style. Product of id images and prompts from category is considered as evaluation set, resulting in two sets - stylistic with 2160 and realistic with 4320 examples. Schematic depiction of the data collection is visualized in Figure 10.

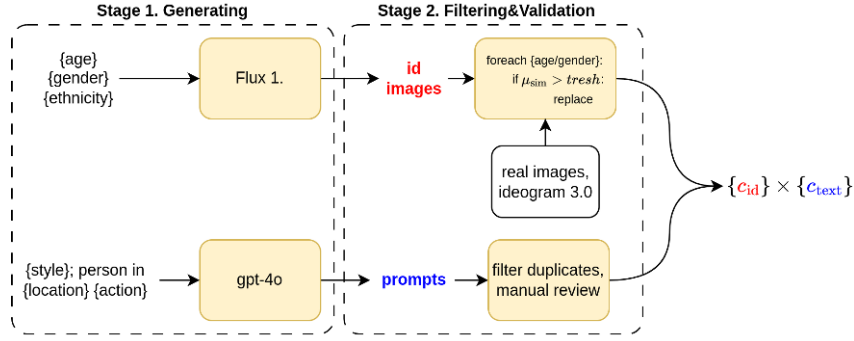


Figure 10: Evaluation dataset preparation pipeline



Figure 11: Evaluation dataset identity samples

C CERTAIN ID-METHODS INFERENCE FAILURE EXAMPLES

Below we provide examples of recent id-preserving generation methods that we found to have limitations in terms of application with our evaluation set.

PuLID In Figure 12 we provide example common failure for PuLID method. From our experiments we find that it is not applicable with prompts that have description of context like location and action, which our evaluation set prompts have. We hypothesize that this effect is rooted in aligned training of PuLID, where inner representations of UNet are regularized to match generation without C_{id} condition - in our experiments we found that in baseline setup FFC metric accounts around for 50% of sampled images failing (meaning around half of images doesn't have any identity detected).

InstantID This method is example of opposite behavior - it's pipeline includes ControlNet-like module that is conditioned on face key-points, which are extracted from source image by standard CV packages (e.g. *insightface* Insight (2023)). However, when tested against multiple different prompts, we observe in Fig. 13 that despite showing state of the art in terms of face preservation, outperforming any other method, it lacks prompt following and variability, not being able to properly follow details regarding background and person body position (additionally it has large bias towards watermark generation with 1:1 resolutions).



Figure 12: Demonstration of common case of failure for PuLID method - method lacks bias to human-centric generation to perform identity preservation, especially with small faces.



Figure 13: Demonstration of common case of failure for InstantID method - generated images are highly constrained and often omit details in the prompt, prompts used for generation: "Person in an ancient library reading", "Person in a futuristic space station repairing equipment", "Person in a high-tech laboratory conducting experiments"

D DCG VARIATIONS AND DERIVATIONS

Preliminary To simplify derivation process let's recall that reverse diffusion process is formulated in terms of score function $\nabla_{x_t} \log p(x_t|y)$ Song et al. (2020), where x_t is noised latent and y is conditional information, in text2image models being prompt. Then classifier guidance can be derived as below, where in Eq. 13 w is added as a hyper-parameter to control conditioning strength.

$$\nabla_{x_t} \log p(x_t|y) = \nabla_{x_t} \log \left(\frac{p(y|x_t)p(x_t)}{p(y)} \right) \quad (11)$$

$$= \nabla_{x_t} \log p(y|x_t) + \nabla_{x_t} \log p(x_t) - \nabla_{x_t} \log p(y) \quad (12)$$

$$\Rightarrow \nabla_{x_t} \log p(x_t) + w \cdot \nabla_{x_t} \log p(y|x_t) \quad (13)$$

Then to arrive to classifier-free guidance (which removes need for learning classifier $f(y|x_t)$ for estimation of $\nabla_{x_t} \log p(y|x_t)$), we rearrange terms in 13 and arrive to following:

$$\nabla_{x_t} \log p(x_t|y) = \nabla_{x_t} \log p(x_t) + w \cdot (\nabla_{x_t} \log p(x_t|y) - \nabla_{x_t} \log p(x_t)) \quad (14)$$

DCG variants Now let's derive possible decoupled classifier-free variants for two conditions, specifically when $y = [c_{text}, c_{id}]$. We note that $\nabla \log p(x_t|c_{text}, c_{id}) - \nabla \log p(x_t)$ from classifier-free guidance corresponds to estimation of $\nabla_{x_t} \log p(c_{id}, c_{text}|x_t)$ score function, which can be expressed in following ways:

$$\nabla_{x_t} \log p(c_{id}, c_{text}|x_t) = \begin{cases} \nabla_{x_t} \log p(c_{id}|x_t, c_{text}) + \nabla_{x_t} \log p(c_{text}|x_t) \\ \nabla_{x_t} \log p(c_{text}|x_t, c_{id}) + \nabla_{x_t} \log p(c_{id}|x_t) \\ \nabla_{x_t} \log p(c_{id}|x_t) + \nabla_{x_t} \log p(c_{text}|x_t) \end{cases} \quad (15)$$

Last expression is possible if we assume that $p(c_{id}, c_{text}) = p(c_{id})p(c_{text})$, which generally is not true, but since in practice choice of prompts and identities for id-preserving generation are not dependent, it can be valid. Finally, reformulating back to noise prediction, we arrive to three possible DCG formulations, where DCG_2 is the one used in main sections of the paper:

$$DCG_1(\hat{\epsilon}) := \epsilon(\emptyset, \emptyset) + \alpha \cdot (\epsilon(c_{text}, \emptyset) - \epsilon(\emptyset, \emptyset)) + \beta \cdot (\epsilon(c_{text}, c_{id}) - \epsilon(c_{text}, \emptyset)) \quad (16)$$

$$DCG_2(\hat{\epsilon}) := \epsilon(\emptyset, \emptyset) + \alpha \cdot (\epsilon(\emptyset, c_{id}) - \epsilon(\emptyset, \emptyset)) + \beta \cdot (\epsilon(c_{text}, c_{id}) - \epsilon(\emptyset, c_{id})) \quad (17)$$

$$DCG_3(\hat{\epsilon}) := \epsilon(\emptyset, \emptyset) + \alpha \cdot (\epsilon(c_{text}, \emptyset) - \epsilon(\emptyset, \emptyset)) + \beta \cdot (\epsilon(\emptyset, c_{id}) - \epsilon(\emptyset, \emptyset)) \quad (18)$$

In practice we find that expression in Eq. 17 works best in terms of semantic changes in the image. While Eq. 18 performs similarly, version in Eq. 16 suffers from image quality degradation and doesn't introduce smooth trade off between identity preservation and stylization, see Figure 14.

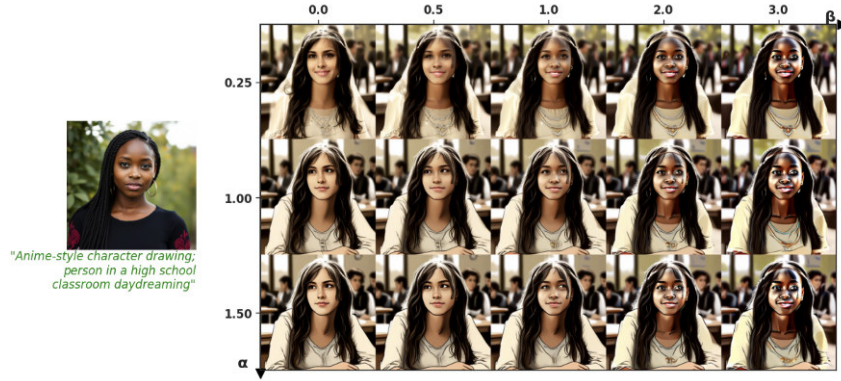


Figure 14: Ablation grid of DCG-1 variant, which can be observed to under perform compared to other decoupled options (same identity used as in Figure 3a)



Figure 15: Scheduling effect on DCG, from right to left - baseline generation, single step alterations of α and β coefficients to high value. In first steps image is completely corrupted, while last step introduces local visual artifacts

In Figure 15 we show that scaling at first and last steps results in significant artifacts, and below in Figure 16 we additionally provide visual examples of rescaling trick contribution in terms of local details in generated images. As it can be seen, rescaling provides additional low-level enhancements of visual images in terms of details coherence.

E AM ANALYSIS AND DETAILS

Scale-power ablation We provide visual ablation why scale-power transformation works in Figure 17. Scaling increases similarity, but alters image background, resulting in prompt following degradation. This is expected, as plugging scaling transform into Eq. 6 instead of $f(\cdot)$ we can see that it is same as increasing λ . When raising attention values to some power, we achieve attention values shifting to 0, which decreases identity preservation, but increases prompt following, especially around face, since attention values in decoupled blocks stop interfering with attention from cross-attention blocks. Combination of transforms results in power transform basically canceling prompt following degradation of scale transform.

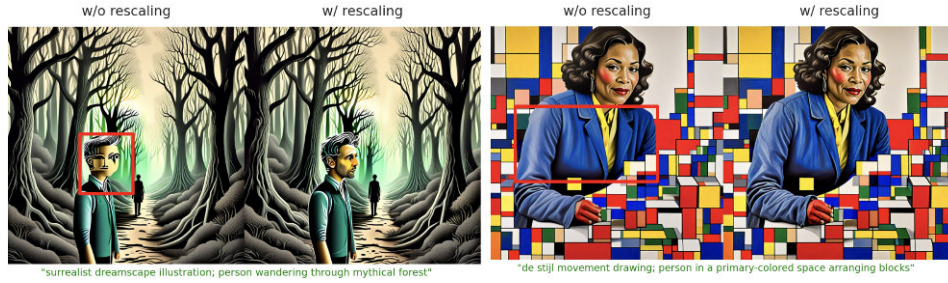


Figure 16: Effect of proposed rescaling on generated images with few-step models - areas with changes are highlighted



Figure 17: Visual ablation of scale-power transform components

Failure cases demonstration In Figure 18 we give examples of id-preserving failures with distilled diffusion model, where instead of expected outcome with human-centric generation method fails to preserve meaningfully align identity and surrounding context, which can result in identity morphing into background, being between multiple humans in image, unrealistic postures and etc. Such cases often can't be fixed by proposed scale-power transform, which serves as motivation for a more control-nature transform that changes structure of images.



Figure 18: Generation examples with distilled model where generated image fails to successfully preserve identity in meaningful way

Scheduled-softmax transform details Beyond details provided in main sections, we also found that attention values for the first token in decoupled CrossAttention (see Fig.4) in FaceID-Plus-v2

are inverted - attention is focused on background across all blocks and timesteps, and it's values histogram has mode closer to 1 value. Therefore, when applying transformation to first token, we first invert it's values, and after transform invert back so that AM transformation has same expected effect across all tokens.

F RESULTS OF DCG IN STYLISTIC SETUP

In Figure 19 we present fronts for DCG in stylistic dataset for Hyper and Lightning. Parameters are specified in main section of the text are shares across all models and also joint application with AM.

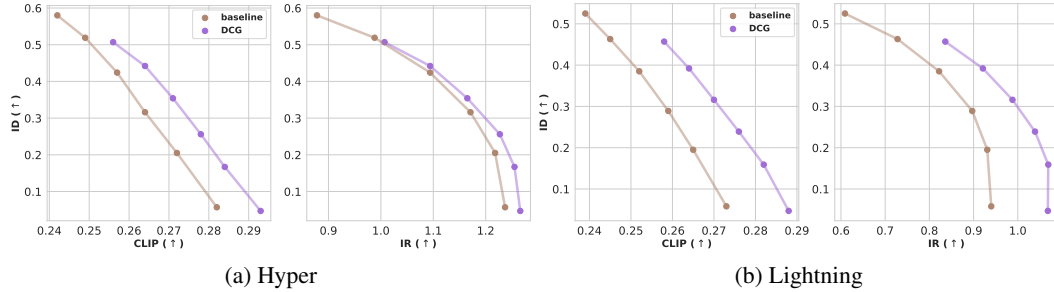


Figure 19: Pareto fronts of Hyper and Lightning with DCG against baseline, stylistic setup, `lora_scale=1`

In Table 4 we report metric comparisons for fixed `ip_adapter_scale=0.8` for all models. We can observe that DCG achieves expected degradation of face similarity, while increasing CLIP, IR and FCS.

Table 4: Ablation of DCG against baseline on stylistic data - DCG increases IR, CLIP score for general image and face area, while also bringing decrease in ID preservation

Model	ID (↑)	CLIP (↑)	IR (↑)	FCS (↑)
Hyper				
base	0.519	0.249	0.988	0.180
DCG	0.442	0.264	1.094	0.184
Lightning				
base	0.463	0.245	0.728	0.175
DCG	0.392	0.264	0.921	0.181
LCM				
base	0.439	0.259	0.540	0.180
DCG	0.336	0.270	0.639	0.181
Turbo				
base	0.310	0.252	0.888	0.165
DCG	0.254	0.277	1.007	0.175

G RESULTS OF AMs IN REALISTIC SETUP

Below we present results in terms of fronts computed on realistic subset and full table computed for fixed `ip_adapter_scale=0.8`. AM1 denotes scale-power transform and AM2 denotes scheduled-softmax transform. In all setups (including joint application with DCG in following sections) all hyper-parameters are fixed across checkpoints and are following:

AM1 - target "up" and "down" unet parts, power strength $p = 1.3$, scale strength $s = 1.45$ in "down" part and $s = 1.55$ in "up" part.

AM2 - target "up" and "down" unet parts, scale strength $s = 1.55$ everywhere except first step; softmax quantile $p = 0.65$ softmax $d = 7.5$ at first step, $d = 5$. at other steps; AdaIN blend coefficient $w = 0.7$.

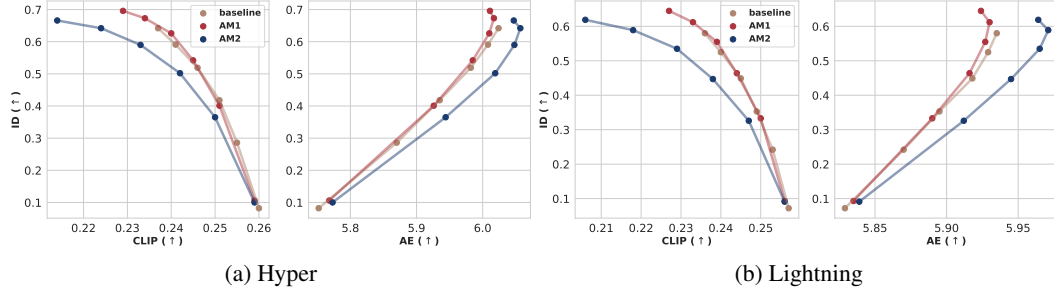


Figure 20: Pareto fronts of Hyper and Lightning with AM mechanisms against baseline, realistic setup

Table 5: Ablation of *AM* transforms against baselines - Both *AM* transformations increase identity preservation, aesthetic scores and stability, while slightly decreasing prompt following, *AM2* results in lower CLIP-score due to larger face bias

Model	lora_scale= 1.0				lora_scale= 0.5			
	ID ↑	CLIP ↑	AE ↑	FFC ↓	ID ↑	CLIP ↑	AE ↑	FFC ↓
Hyper								
base	0.591	0.241	6.008	0	0.408	0.255	6.229	19
AM1	0.673	<u>0.234</u>	<u>6.017</u>	1	0.523	<u>0.247</u>	6.220	10
AM2	<u>0.642</u>	0.224	6.057	0	<u>0.517</u>	0.239	6.265	3
Lightning								
base	0.525	0.240	5.929	0	0.386	0.249	6.079	18
AM1	0.612	<u>0.233</u>	<u>5.930</u>	0	<u>0.494</u>	<u>0.241</u>	6.088	12
AM2	<u>0.589</u>	0.218	5.971	0	0.496	0.231	6.107	1
LCM								
base	0.552	0.235	5.754	3	0.380	0.249	5.927	46
AM1	0.610	<u>0.227</u>	<u>5.783</u>	1	0.477	<u>0.240</u>	5.942	34
AM2	<u>0.597</u>	0.214	5.802	1	<u>0.476</u>	0.231	5.974	18
Turbo								
base	0.349	0.243	5.650	94	0.189	0.250	5.764	116
AM1	0.467	<u>0.235</u>	5.635	57	0.289	<u>0.244</u>	5.769	63
AM2	<u>0.443</u>	0.230	<u>5.647</u>	62	<u>0.283</u>	0.240	5.784	51

In Figures 21 and 22 we demonstrate examples of applying just AM during inference. It can be seen that AM1 enhances identity similarity locally, without disrupting prompt following, while AM2 introduces larger faces and portrait like bias for generated outputs, which also results in lower prompt following.



Figure 21: Application of AM compared to baselines

Figure 22: Application of AM compared to baselines, `lora_scale=0.5`

H ADDITIONAL FASTFACE RESULTS

H.1 PARETO FRONTS

Below in Fig. 23 and 24 we provide Pareto fronts evaluated for FastFace framework on DiverseBench for varying `ip_adapter.scale` $\in \{0.1, 0.35, 0.5, 0.65, 0.8, 0.95\}$. These plots give additional information of scaling behaviors when trying to tune just ip-adapter scale. It can be seen that fronts introduced by FastFace achieve superior trade offs across different models.

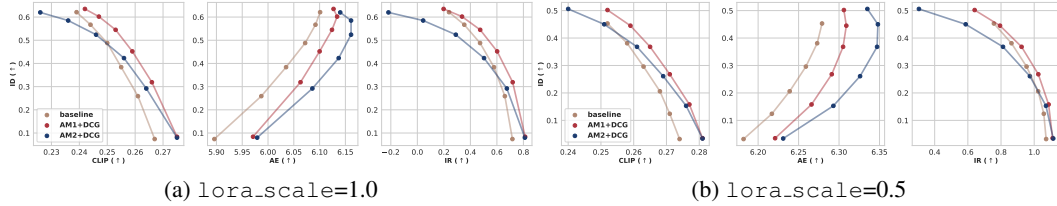
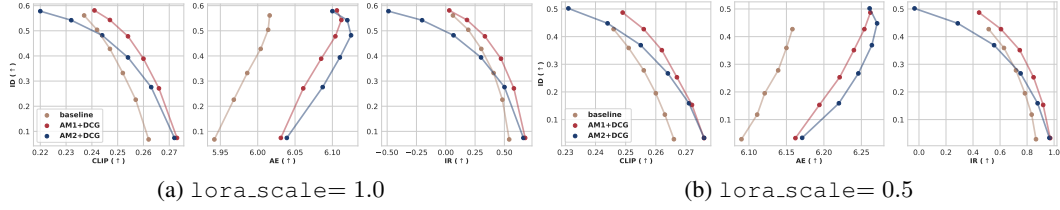


Figure 23: Pareto fronts built for Hyper model metrics with different scales of LoRA

H.2 FASTFACE ABLATION

In Table 6 we report main metric evaluation for fixed value of λ across all models with full and lower LoRA scale, common trick when applying ID-Adapters for more creative generation - as a result FastFace enhances identity similarity, image quality and stability without loss of prompt following.

Figure 24: Lightning fronts for full data setup, different FastFace configurations and `lora_scales`Table 6: Metric comparison of baseline setup against FastFace setups - FF_{AM1} denotes application of DCG with scale-power transform, FF_{AM2} - DCG with scheduled-softmax transform

Model	lora_scale=1.0				lora_scale=0.5			
	ID (↑)	CLIP (↑)	AE (↑)	FFC (↓)	ID (↑)	CLIP (↑)	AE (↑)	FFC (↓)
Hyper								
base	0.567	0.244	6.092	2	0.381	0.258	6.273	88
FF_{AM1}	0.602	0.247	6.134	2	0.445	0.259	6.309	72
FF_{AM2}	0.585	0.236	6.161	0	0.450	0.251	6.348	34
Lightning								
base	0.504	0.242	6.014	4	0.359	0.251	6.150	89
FF_{AM1}	0.543	0.247	6.112	2	0.427	0.256	6.254	66
FF_{AM2}	0.542	0.232	6.120	0	0.448	0.244	6.271	33
LCM								
base	0.515	0.243	5.770	53	0.344	0.258	5.911	288
FF_{AM1}	0.525	0.244	5.796	37	0.383	0.258	5.968	202
FF_{AM2}	0.533	0.229	5.807	20	0.406	0.246	5.979	136
Turbo								
base	0.336	0.246	5.689	161	0.177	0.257	5.791	242
FF_{AM1}	0.416	0.249	5.698	139	0.239	0.262	5.757	431
FF_{AM2}	0.409	0.242	5.707	94	0.244	0.256	5.798	271

Additionally we provide sensitivity analysis of FastFace with respect to AM hyperparameters. It can be seen in the figure below that hyperparameters do not affect quality of the output in random way and can be chosen from wide range, offering optional tuning depending on the task and model.

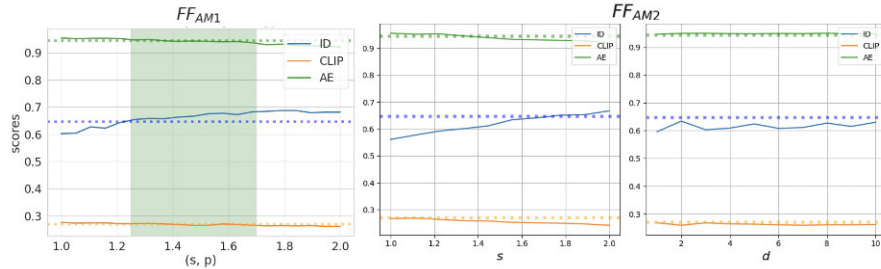
Figure 25: Sensitivity analysis w.r.t. to AM hyperparameters; left - AM_1 , right - AM_2 , AE metric is rescaled to match other metrics range between 0 and 1, dotted lines denote baseline quality



Figure 26: More inference examples of FastFace pipeline with IpAdapter-FaceID-v2 in different setups