Attribute-Preserving Pseudo-Labeling for Diffusion-Based Face Swapping

Supplementary Material

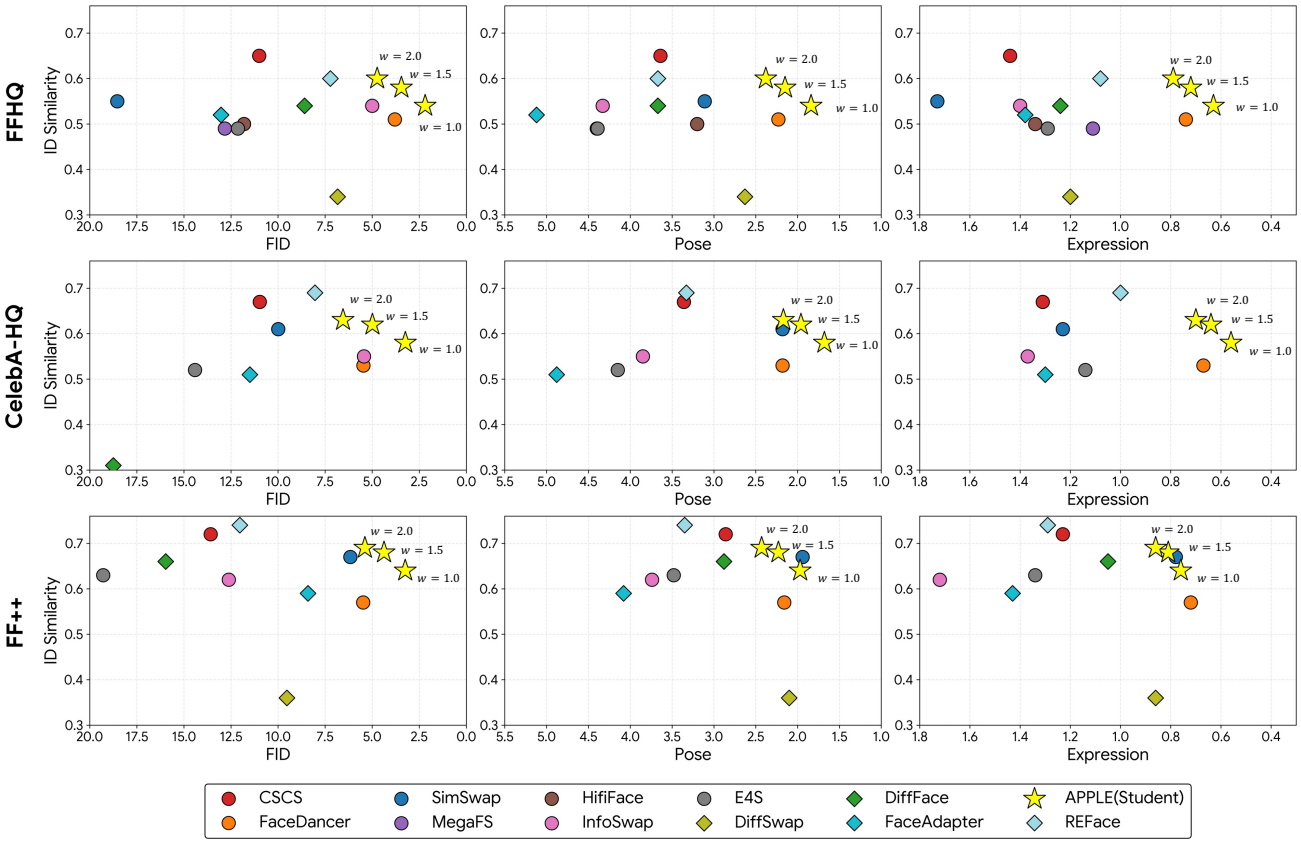

Figure 8. **Pareto frontier visualizations across datasets.** We visualize ID Sim. vs. FID, Pose, and Expression reported in Tabs. 4 to 7. Ideal face-swapping models should occupy the **upper-right region**—high identity similarity with low FID and pose/expression error. APPLE(Student) *consistently* lies on or near the Pareto frontier across all datasets, demonstrating **robust attribute preservation**. For APPLE, we present results under varying guidance scale w. Note that some baselines (REFace, CSCS, InfoSwap) are trained on CelebA-HQ.

A. Experimental Details

A.1. Additional implementation details

We use FLUX.1-Krea [dev] [3] as the base diffusion backbone for all experiments, employing PuLID [15] as the identity encoder and OminiControl [47] as the attribute conditioning branch with a LoRA [20] rank of 64. The model is trained on the VGGFace2-HQ [6] dataset and evaluated on FFHQ [23]. Models are trained on images filtered by an AES [39] ≥ 5.1 threshold to ensure high-quality faces. Following REFace [54], the source image is masked before being fed into the identity encoder. The overall objective combines identity, flow-matching losses. Identity losses are applied to late 35% of the diffusion timesteps for teacher model and late 50% for student model. Teacher model is trained for 15K iterations without the identity loss and for 50K iterations with it enabled. Student model is resumed from teacher model and trained for 15K steps with pseudo-triplet generated by teacher model. Experiments are mainly conducted on four NVIDIA A6000 GPUs with a batch size of 1 per GPU, gradient accumulation of 4, resulting in an effective batch size of 16. AdamW is used as optimizer with learning rate 1e-4 and weight decay of 1e-3 is used. Inference and inversion step of diffusion model is fixed to 28. Training and inference are conducted at a resolution of 512x512.

B. Additional Experiments

B.1. Evaluation on additional datasets with diverse guidance scales

Guidance scale ablation on FFHQ. Since our model $v_t(\cdot)$ accepts source identity feature $\mathbf{id}_{\text{src}}$ and target attribute feature $\mathbf{att}_{\text{tgt}}$ as conditions, classifier-free guidance (CFG) with scale w can be applied to the identity condition to strengthen identity transferability:

$$\begin{aligned}\tilde{v}_t = v_t(z_t, \varnothing, \mathbf{att}_{\text{tgt}}) \\ + w\Big(v_t(z_t, \mathbf{id}_{\text{src}}, \mathbf{att}_{\text{tgt}}) - v_t(z_t, \varnothing, \mathbf{att}_{\text{tgt}})\Big).\end{aligned} \tag{9}$$

We report results of FFHQ across varying guidance scales w in Tab. 5 and Fig. 9. Starting from the default setting $w = 1.0$, increasing w improves identity transferability at the cost of moderate degradation in attribute preservation, demonstrating inference-time control over the trade-off. Thus, user can flexibly adjust w to prioritize either identity transfer or attribute preservation based on specific application needs.

Note that while other diffusion-based baselines (DiffFace, FaceAdapter, REFace) are evaluated using their respective tuned hyperparameter w, **we report our main results without hyperparameter tuning** ($w = 1.0$).

Table 5. **Ablation study of guidance scale w on FFHQ.** Unless otherwise specified, all results in the main paper use $w = 1.0$ by default. Increasing the guidance scale enhances identity similarity and retrieval performance, albeit with a moderate trade-off in attribute preservation metrics, such as pose and expression.

Model	FID↓	ID Sim.↑	ID Retrieval↑ Top-1 / Top-5	Pose↓	Expr.↓
APPLE (Student), $w = 1.0$	2.19	0.54	90.40 / 96.50	1.84	0.63
APPLE (Student), $w = 1.5$	3.45	0.58	94.60 / 98.30	2.15	0.72
APPLE (Student), $w = 2.0$	4.74	0.60	96.00 / 98.40	2.38	0.79

Table 6. **Quantitative results on CelebA-HQ.** Note that some baselines (REFace, CSCS, InfoSwap) are trained on CelebA-HQ, while APPLE is trained solely on VGGFace2-HQ.

Model	FID↓	ID Sim.↑	ID Retrieval↑ Top-1 / Top-5	Pose↓	Expr.↓
SimSwap	9.99	0.61	97.70 / 99.00	2.18	1.23
CSCS	10.97	0.67	97.70 / 99.50	3.36	1.31
FaceDancer	5.47	0.53	92.60 / 98.50	2.18	0.67
InfoSwap	5.44	0.55	93.10 / 97.50	3.85	1.37
E4S	14.41	0.52	83.50 / 92.30	4.15	1.14
DiffSwap	20.08	0.23	18.94 / 35.07	7.54	2.08
DiffFace	18.75	0.31	46.30 / 47.30	10.84	1.88
FaceAdapter	11.50	0.51	85.10 / 90.80	4.88	1.30
REFace	8.04	0.69	99.10 / 99.90	3.33	1.00
APPLE (Student), $w = 1.0$	3.24	0.58	95.80 / 98.60	1.68	0.56
APPLE (Student), $w = 1.5$	4.99	0.62	97.00 / 99.30	1.96	0.64
APPLE (Student), $w = 2.0$	6.55	0.63	97.30 / 99.50	2.17	0.70

Table 7. **Quantitative results on FaceForensics++.**

Model	FID↓	ID Sim.↑	ID Retrieval↑ Top-1 / Top-5	Pose↓	Expr.↓
SimSwap	6.16	0.67	97.89 / 99.00	1.94	0.78
CSCS	13.58	0.72	94.70 / 97.50	2.86	1.23
FaceDancer	5.48	0.57	90.77 / 98.40	2.16	0.72
InfoSwap	12.62	0.62	97.20 / 98.90	3.74	1.72
E4S	19.29	0.63	94.30 / 97.50	3.48	1.34
DiffSwap	9.53	0.36	10.10 / 57.10	2.10	0.86
DiffFace	15.97	0.66	98.08 / 99.04	2.88	1.05
FaceAdapter	8.62	0.59	89.46 / 91.61	4.13	1.44
REFace	12.03	0.74	98.60 / 99.10	3.35	1.29
APPLE (Student), $w = 1.0$	3.25	0.64	97.69 / 98.90	1.97	0.76
APPLE (Student), $w = 1.5$	4.37	0.68	98.29 / 98.90	2.23	0.81
APPLE (Student), $w = 2.0$	5.39	0.69	98.49 / 99.30	2.43	0.86

Extended evaluation on CelebA-HQ and FaceForensics++. To address the concern that evaluation on FFHQ alone may not fully reflect generalization, we further evaluate APPLE on CelebA-HQ and FaceForensics++ (FF++). Compared to FFHQ, these datasets present broader variations in image quality, appearance, and real-world artifacts, providing a more challenging benchmark for jointly assessing identity transfer and attribute preservation.

As shown in Tabs. 6 and 7 and visualized via Pareto frontiers in Fig. 8, APPLE consistently remains on or near the Pareto frontier (upper-right) across all datasets. This performance is particularly notable as it is achieved through training solely on VGGFace2-HQ, whereas several competing methods benefit from dataset-specific training on CelebA-HQ. Furthermore, APPLE demonstrates superior robust-

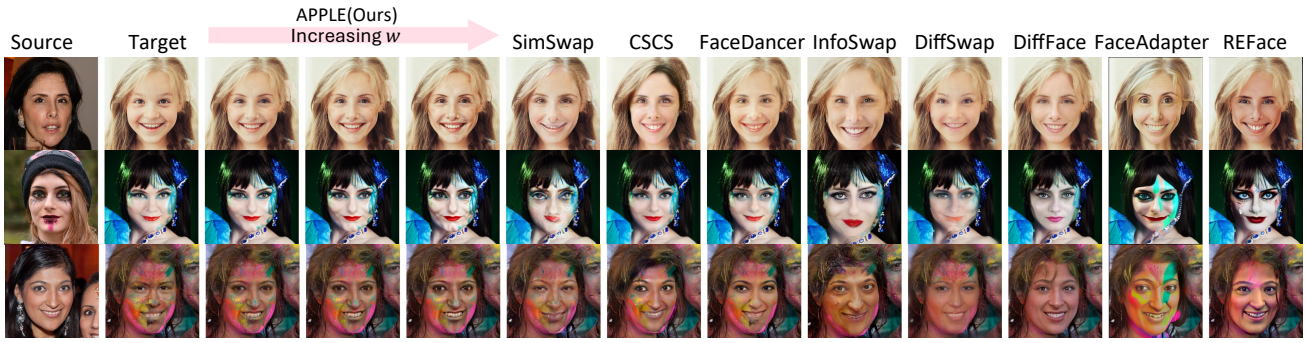

Figure 9. **Ablation study of guidance scale w on FFHQ.** Guidance scale w enables inference-time control of the identity-attribute trade-off. From our default $w = 1.0$, increasing w boosts identity transferability at the cost of moderate attribute preservation degradation.

ness. While baselines such as REFace and CSCS show severe FID degradation on low-quality datasets like FF++, our method maintains high-quality synthesis and identity consistency.

B.2. Additional qualitative results on FFHQ

We present additional qualitative results of APPLE (Student) in Fig. 14–17.

C. Discussion and Analysis

C.1. Detailed motivation and ablation study of conditional deblurring formulation

In Sec. 4.3, we introduce conditional deblurring as a replacement for the conventional masking strategy used in diffusion-based face swapping. This design choice arises from a fundamental constraint of face swapping task: ground-truth face-swapped pairs do not exist in practice.

Self-supervised training in GAN-based methods. GAN-based face swapping models [5, 25, 37] sidestep this inherent limitation by self-supervised training with pairs of images from different identities. Given a source image I^A_{src} of identity A and a target image I^B_{tgt} of identity B, the generator is trained with two explicit objectives applied to the generated result I_{result}. An identity loss encourages I_{result} to share identity with I^A_{src}, typically by maximizing the cosine similarity between features extracted from a pretrained identity encoder (e.g., ArcFace [7]) applied to both images. An attribute-related loss encourages I_{result} to preserve appearance attributes of I^B_{tgt}, commonly implemented by aligning attribute-related features from a custom attribute extractor [25], a GAN discriminator [5], or empirically identified attribute-sensitive layers of the identity encoder itself [37]. Because a GAN produces a clean image in a single forward pass, these feature-based losses can be applied directly to the final output, enabling self-supervised learning without real swapped pairs.

Why self-supervised paradigm of GAN is inadequate for diffusion models? This self-supervised paradigm widely used in GAN methods cannot be directly applied to diffusion-based face swapping. Identity and attribute-related features are typically extracted from clean images. During training, the diffusion model predicts z_0 from noisy image z_t at various noise levels. At high noise levels (early timesteps), the predicted z_0 is far from clean, making feature-based identity and attribute-related losses unreliable or undefined, which leads to training infeasible. While such losses could in principle be applied at low noise levels, this would require excluding high noise levels (early timesteps) from training, severely limiting the model's ability to denoise from initial gaussian noise.

The proxy formulation in diffusion-based methods. To address this, existing diffusion-based face swapping approaches adopt inpainting as a proxy training task . Instead of using images from different identities, they use same-identity pairs: a source image I^A_{src} and a target image I^A_{tgt} of the same person A. The target image is degraded in a way that suppresses its identity information, while the source image provides identity features from the same person. The diffusion model is then trained to reconstruct the original clean target I^A_{tgt} from its degraded version, conditioned on features from I^A_{src}. At inference time, replacing the source with an image from a different identity B, while keeping the target from identity A, yields a face-swapped result.

In this proxy formulation, the degradation applied to the target image is the mechanism determining how much identity information is removed and how much attribute information remains. Prior works [16, 54, 57] typically employ full-face masking, replacing the facial region with black image to ensure strong identity suppression. However, such masking discards most attribute cues including lighting, skin tone, makeup, gaze, and expression—which forces the diffusion model to 'imagine' these details during generation. This leads to severe attribute misalignment at inference time, even when identity transfer succeeds.

Key design questions. Once we recognize that masking is an unsuitable degradation strategy, adopting an alternative

degradation approach requires addressing two key considerations. First, what degradation operator should be applied to the target face in order to reliably suppress identity information. Second, how strong should this degradation be so that identity cues are removed while attribute cues such as lighting, pose, expression, makeup, and accessories remain exploitable by the model. The degradation design must balance two competing objectives. If the degradation is too weak, identity information from the target is still visible, and the diffusion model can solve the proxy task by reconstructing the target identity, rather than by learning to rely on the source identity. This results in poor identity transfer when the source and target identities differ at inference time. If the degradation is too strong, most appearance cues are destroyed and the model observes little about the target's attributes, which leads it to hallucinate lighting, expression, or makeup in a target-agnostic manner. In that case, even if identity transfer is successful, attribute preservation deteriorates.

Ablation study design. To validate our design choices and systematically compare alternatives, we conduct ablation experiments across two dimensions: (1) degradation type and (2) degradation strength. Natural candidates for degradation type include: (1) (fully) masking (2) downsampling–upsampling, (3) Gaussian blurring, and (4) None, as shown in Fig. 10. For downsampling–upsampling and Gaussian blurring, we explore multiple degradation strengths calibrated so that the resulting degraded faces appear qualitatively comparable in terms of visual severity. All teacher models are trained under a unified protocol for fair comparison: we use a two-phase schedule (training without identity loss followed by training with identity loss), fixing the global step to 5K for each phase, with an effective batch size of 8. In our notation, `Downsample-{N}` refers to downsampling the target face to $N \times N$ resolution followed by upsampling back to the original size. `GaussianBlur-{R}` denotes Gaussian blur applied using `PIL.ImageFilter.GaussianBlur` with `radius=R`. We report quantitative results in Tab. 8 and qualitative results in Fig. 11.

The results reveal a clear trend. As the degradation strength increases, identity leakage from the target is reduced, which helps the model rely more consistently on the source identity. This yields higher identity similarity and stronger identity retrieval performance at inference time. In contrast, using a clean, non-degraded target as the conditional input leads to extremely low identity similarity, confirming that explicit identity suppression is necessary.

Increasing degradation strength introduces a moderate drop in pose and expression performance, reflecting the inherent trade-off between identity suppression and attribute preservation. Nonetheless, these metrics remain consistently higher than those obtained with masking-based baselines. At first glance, pose and expression scores may appear only marginally different across degradation types. However, it is important to note that these geometric cues are already provided to the model through landmark conditioning. Qualitative results in Fig. 11 further demonstrate that both downsampling–upsampling and Gaussian blur preserve a wider range of target attributes (particularly lighting) far more faithfully than masking. This observation is reinforced by FID, which captures overall visual realism: masking-based baselines exhibit substantially worse FID scores because they must hallucinate attributes that are otherwise preserved when using downsampling–upsampling or Gaussian blur degradations.

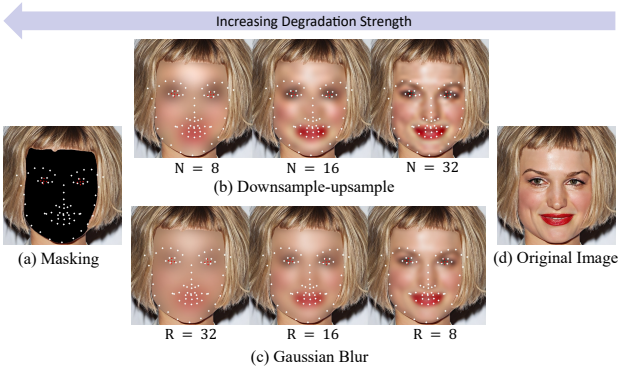

Figure 10. **Example of considered degradation types and strengths for ablation study of conditional deblurring formulation.**

We also find that, for a matched degradation level, the choice between downsampling–upsampling and Gaussian blur has a smaller impact than the overall strength of degradation itself. Both operators achieve similar identity suppression when calibrated to comparable visual degradation.

Additionally, since some diffusion-based face swapping methods [16, 54] inject target CLIP features as an auxiliary attribute cue, we also evaluate a masking variant conditioned on target CLIP features. However, even with CLIP conditioning, masking fails to adequately preserve target attributes, as shown in Fig. 11.

In practice, we adopt the downsampling–upsampling strategy at 8×8 resolution as our default. This choice offers a favorable balance between identity suppression and attribute preservation while being simple, stable, and computationally efficient to implement.

C.2. Design choices for identity loss

As discussed in Sec. C.1, applying identity loss directly in diffusion models is nontrivial because they do not produce clean images in a single forward pass. Standard practice in face swapping is to use a pretrained identity embedding encoder such as ArcFace [7] to enforce similarity between the source identity embedding and the embedding of the

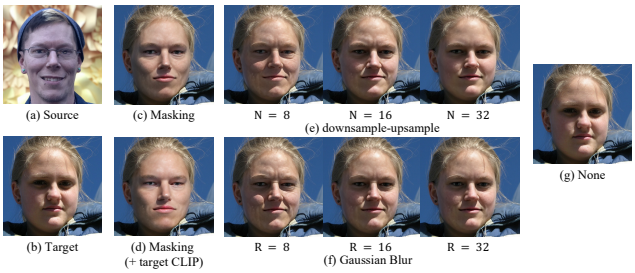

Figure 11. **Qualitative results of ablation study on conditional deblurring formulation.** 'Masking' fails to preserve target attributes (e.g., lighting) due to excessive degradation, even when target CLIP [34] features are provided. Without degradation ('None'), the model simply copies the target image, failing to transfer identity. Moderate degradation ('Downsample-Upsample' or 'Gaussian Blur') successfully alters identity while preserving target attributes. Note that stronger degradation improves identity transferability at the cost of precise attribute preservation.

Table 8. **Ablation results of conditional deblurring formulation.** We compare multiple degradation types and strengths used for conditional deblurring. Both downsample–upsample and Gaussian blur exhibit a consistent trend: stronger degradation increases identity suppression, improving identity transfer while weakening attribute preservation. Nevertheless, either degradation method achieves a substantially better identity–attribute balance than masking-based degradation or no degradation at all.

Degradation type & strength	FID↓	ID Sim.↑	ID Retrieval↑	Pose↓	Expr.↓
Masking	7.93	0.51	87.40 / 95.90	2.59	0.88
Masking (+ target CLIP [34])	7.87	0.42	68.10 / 83.60	2.47	0.83
Downsample-8	3.86	0.51	87.40 / 94.20	2.42	0.78
Downsample-16	2.68	0.47	80.00 / 90.70	2.04	0.71
Downsample-32	2.16	0.44	69.50 / 85.40	1.71	0.64
GaussianBlur-32	3.60	0.48	83.70 / 91.70	2.28	0.76
GaussianBlur-16	2.62	0.46	79.80 / 90.50	1.95	0.70
GaussianBlur-8	2.08	0.45	74.50 / 87.90	1.72	0.65
None	0.19	0.07	0.10 / 0.50	0.27	0.18

generated result. However, these pretrained encoders are typically trained on clean images and are not designed to handle the noisy intermediate predictions produced during diffusion training.

Existing approaches to identity loss in diffusion models. Prior diffusion-based face swapping methods [54, 57] employ various strategies to enable identity supervision. DiffSwap [57] introduces midpoint estimation, in which the model takes two denoising steps from the current noisy latent to obtain a cleaner prediction z_0, computes the identity loss on this prediction, and backpropagates the gradient. REFace [54] applies identity loss to the final output z_0 obtained after running N DDIM steps from an initial noise. While these approaches enable identity supervision throughout training, they incur substantial memory overhead due to the additional denoising steps required at each training iteration. This makes them difficult to scale to large modern models such as FLUX.1 [3].

Table 9. **Ablation study of identity loss application across different timestep ranges.** Expanding the range of timesteps at which identity loss is applied improves identity transferability but reduces attribute preservation.

Method	FID↓	ID Sim.↑	ID Retrieval↑	Pose↓	Expr.↓
Teacher (30%)	4.03	0.47	81.70 / 92.50	2.57	0.79
Teacher (35%)	4.12	0.48	83.50 / 94.20	2.55	0.80
Teacher (40%)	6.26	0.57	94.60 / 97.90	2.88	0.89
Teacher (50%)	7.01	0.58	94.60 / 98.00	3.03	0.93
Student (35%)	1.92	0.49	82.80 / 92.00	1.81	0.61
Student (50%)	2.47	0.53	87.50 / 95.60	2.07	0.65
Student (75%)	3.91	0.58	94.00 / 98.10	2.52	0.74

Our approach. Following PortraitBooth [31], we adopt a simpler strategy: applying identity loss only at low noise levels (late timesteps), where the predicted z_0 is sufficiently clean for reliable feature extraction. The key hyperparameter in this approach is the threshold that defines the low noise level. We conduct an ablation study on both teacher and student models to tune this hyperparameter, fixing the training steps to 5K for all configurations. Note that for the teacher model, we initialize from a checkpoint pretrained on FFHQ [23] without identity loss, which differs from the teacher setting used in the main paper. The student model configuration remains identical.

For notation, we use `model(num%)`, where `num` denotes the percentage of late timesteps (low noise levels) at which identity loss is applied. For example, Teacher (30%) indicates that the teacher model was trained with identity loss applied to the latest 30% of timesteps. As expected, there is a trade-off between identity transferability and attribute preservation. Expanding the range of timesteps at which identity loss is applied improves identity similarity but introduces a moderate drop in attribute-related metrics such as pose and expression accuracy. We select a threshold of 35% for the teacher model and 50% for the student model, balancing identity transfer and attribute preservation.

C.3. Further boosting GAN via training with our occlusion-augmented pseudo-triplet

Although diffusion models offer superior generative fidelity, GAN-based face swapping remains widely used in practical applications due to its substantially faster inference speed, as shown in Tab. 10. Improving GAN performance therefore remains an important and complementary research direction.

Beyond improving diffusion-based models within the teacher–student framework, our attribute-aware pseudo triplets can serve as strong supervision for GAN-based face swapping models also. While some previous works [22, 55] rely on proxy data generated by older face-swapping models [13] or reenactment systems [50] to compensate for the lack of real supervision, such proxy pairs frequently suf-

Figure 12. **Example of applying occlusion augmentation in pseudo-triplets.**

Table 10. **Comparison of inference speed between GAN and Diffusion model.** Speeds are evaluated on RTX A6000.

Model	Second/Sample ↓
GAN (CSCS-O)	0.011
Diffusion (APPLE)	19.858

fer from noise-like artifacts, inconsistent lighting or gaze, and weak attribute alignment, which ultimately reduces their usefulness as training signals. In contrast, our pseudo triplets preserve target attributes with substantially higher fidelity, providing both qualitatively and quantitatively superior supervision compared to GAN-generated proxies.

Table 11. **Quantitative results of CSCS-O on FFHQ, which is trained with occlusion-augmented pseudo-triplets generated by APPLE (Teacher).** Original CSCS [22] reports abnormally high identity similarity and retrieval scores due to its copy–paste behavior, which comes at the cost of substantially degraded FID, pose, and expression performance. CSCS-O, which is trained with our occlusion-augmented pseudo-triplets, exhibits markedly stronger robustness to occlusion while maintaining high identity transferability, attribute preservation, and visual fidelity. This demonstrates that our pseudo-triplets provide effective training signals not only for diffusion models but also for GAN-based face swapping.

Model	FID↓	ID Sim.↑	ID Retrieval↑	Pose↓	Expr.↓
SimSwap [5]	18.54	0.55	94.10/ 99.00	3.11	1.73
MegaFS [58]	12.83	0.49	79.6 / 86.3	4.40	1.11
FaceDancer [37]	3.80	0.51	89.70 / 96.50	2.23	0.74
HiFiFace [49]	11.81	0.50	85.40 / 93.40	3.20	1.34
InfoSwap [13]	5.00	0.54	91.40 / 97.50	4.33	1.40
E4S [28]	12.13	0.49	78.30 / 87.80	4.39	1.29
DiffSwap [57]	6.84	0.34	41.92 / 63.09	2.63	1.20
DiffFace [24]	8.59	0.54	90.70 / 95.90	3.67	1.24
FaceAdapter [16]	13.03	0.52	87.00 / 93.20	5.12	1.38
REFace [54]	7.22	0.60	97.60 / 99.40	3.67	1.08
CSCS [22]	11.00	0.65	99.00 / 99.50	3.64	1.44
CSCS-O (w/o LPIPS)	5.98	0.52	90.20 / 97.10	2.04	1.23
CSCS-O (w/ LPIPS)	4.55	0.54	90.40 / 96.40	2.12	1.30

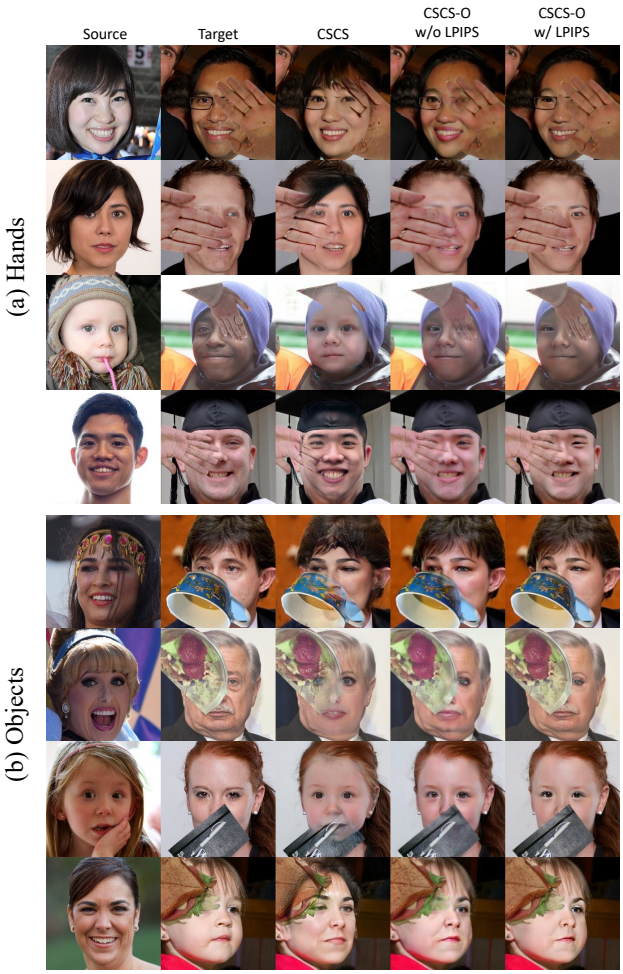

Figure 13. **Results of CSCS trained with our occlusion-augmented pseudo-triplets.** CSCS-O, which is variant of CSCS [22] trained with our occlusion-augmented pseudo-triplets, exhibits greatly improved robustness to occlusions compared to baseline (CSCS), especially when trained with LPIPS [56] loss.

In addition, our pseudo triplets can be further extended with occlusion augmentation, enabling direct supervision for training models to be robust under real-world occlusions such as hands, accessories, or objects covering the face. As show in Fig. 12, we overlay hand and random object occlusions on the target images in our pseudo-triplets and use these occluded pairs as supervision, following [48]. A key challenge here is that the identity loss encourages the generator to reconstruct the source identity across the entire facial region, which causes the model to overwrite occluding objects with hallucinated face components. To prevent this undesired behavior, we additionally employ an LPIPS [56] loss between the generated image and the occluded pseudo target. This perceptual loss encourages the generator to respect the structure and texture of the occlusion, preventing it from being incorrectly "corrected" to a facial region. As

Table 12. **Training cost breakdown.** Cost is calculated assuming 4×RTX A6000 GPUs. The full training pipeline takes approximately 13 days, which is reasonable even compared to single-stage methods. For example, REFace reports 18 days of training cost on 2×A100 GPUs.

Stage	Time (days)
Teacher training (65K steps)	4.3
Pseudo-triplet generation (50K images)	5.78
Student training (15K steps)	2.4
Total	**12.78**

a result, LPIPS stabilizes fine-tuning under occlusion, enabling the model to maintain identity similarity where appropriate while faithfully preserving the occluding objects.

We evaluate the effectiveness of our occlusion-augmented pseudo-triplet by fine-tuning a representative GAN-based model, CSCS [22], and report quantitative results in Tab. 11 and qualitative results in Fig. 13. We denote the model trained with our occlusion-augmented pseudo-triplets as CSCS-O. Qualitatively, CSCS-O demonstrates markedly improved robustness to occlusions compared to models trained with conventional proxy data, particularly in settings where LPIPS loss is used. Quantitatively, it yields substantial gains in FID, pose, and expression consistency while maintaining strong identity transferability.

C.4. Training cost breakdown

We report the training cost breakdown of our teacher-student framework in Tab. 12. The teacher model is trained for 15K iterations without identity loss and 50K iterations with identity loss, while the student model is trained for 15K iterations with pseudo-triplet generated by the teacher model. The overall training time is approximately 13 days on 4×NVIDIA A6000 GPUs.

D. Limitations and future directions

While our teacher–student framework generally performs robustly, it still relies on several external modules such as 3DMM landmarks, gaze landmarks, and segmentation maps, when constructing pseudo-triplets for student training. Errors introduced by these modules may propagate into the pseudo-triplet, which may influence the downstream student training. Reducing this dependency, or designing architectures that are more resilient to errors of external module, is an important direction for future work. Additionally, improving the inference efficiency of diffusion models and extending the framework to video face swapping present promising avenues for further research.

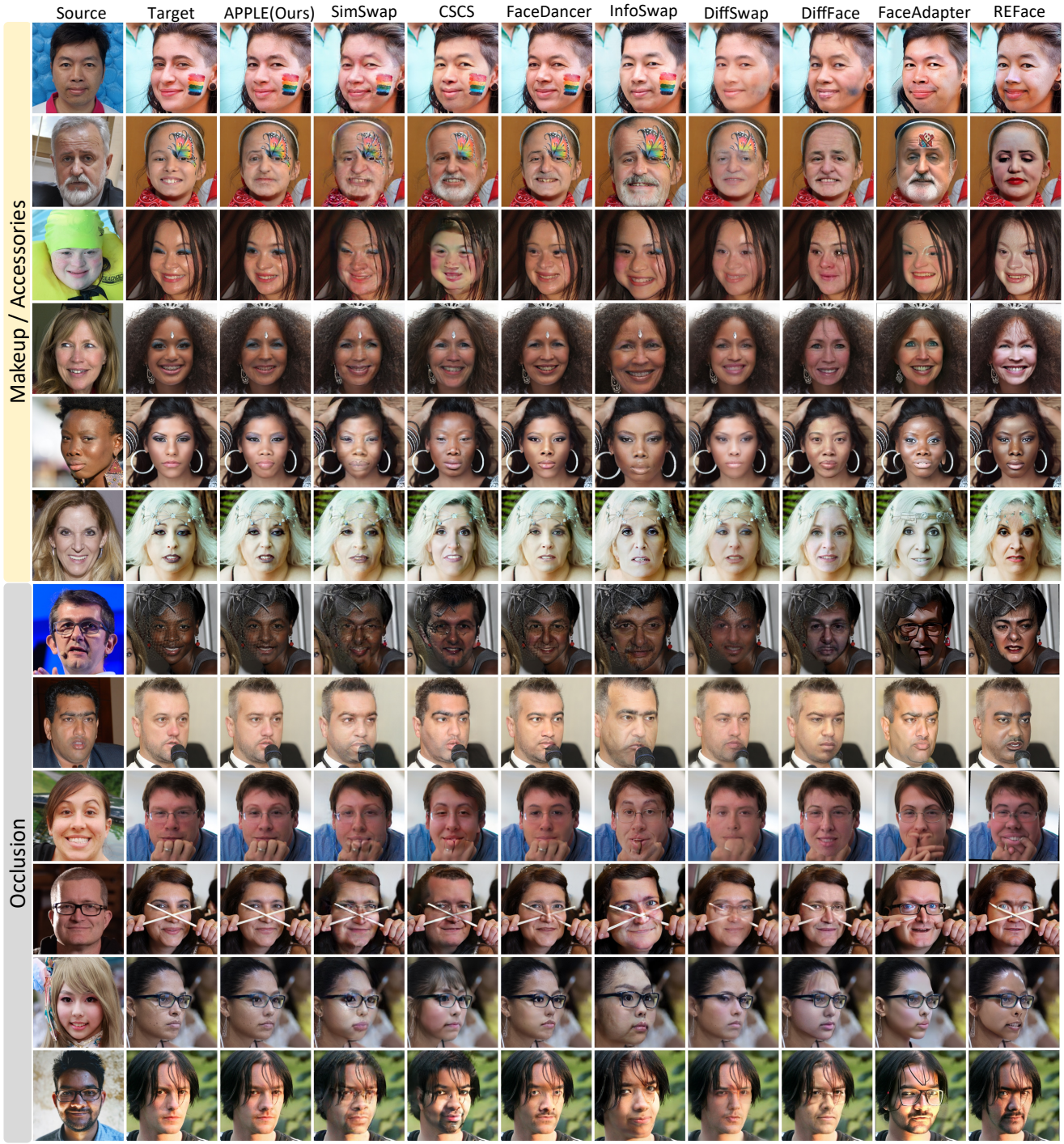

Figure 14. **Qualitative results.** Face swapping results on the FFHQ dataset [23] compared with existing baselines. APPLE effectively preserves the target image's attributes while faithfully transferring the source identity.

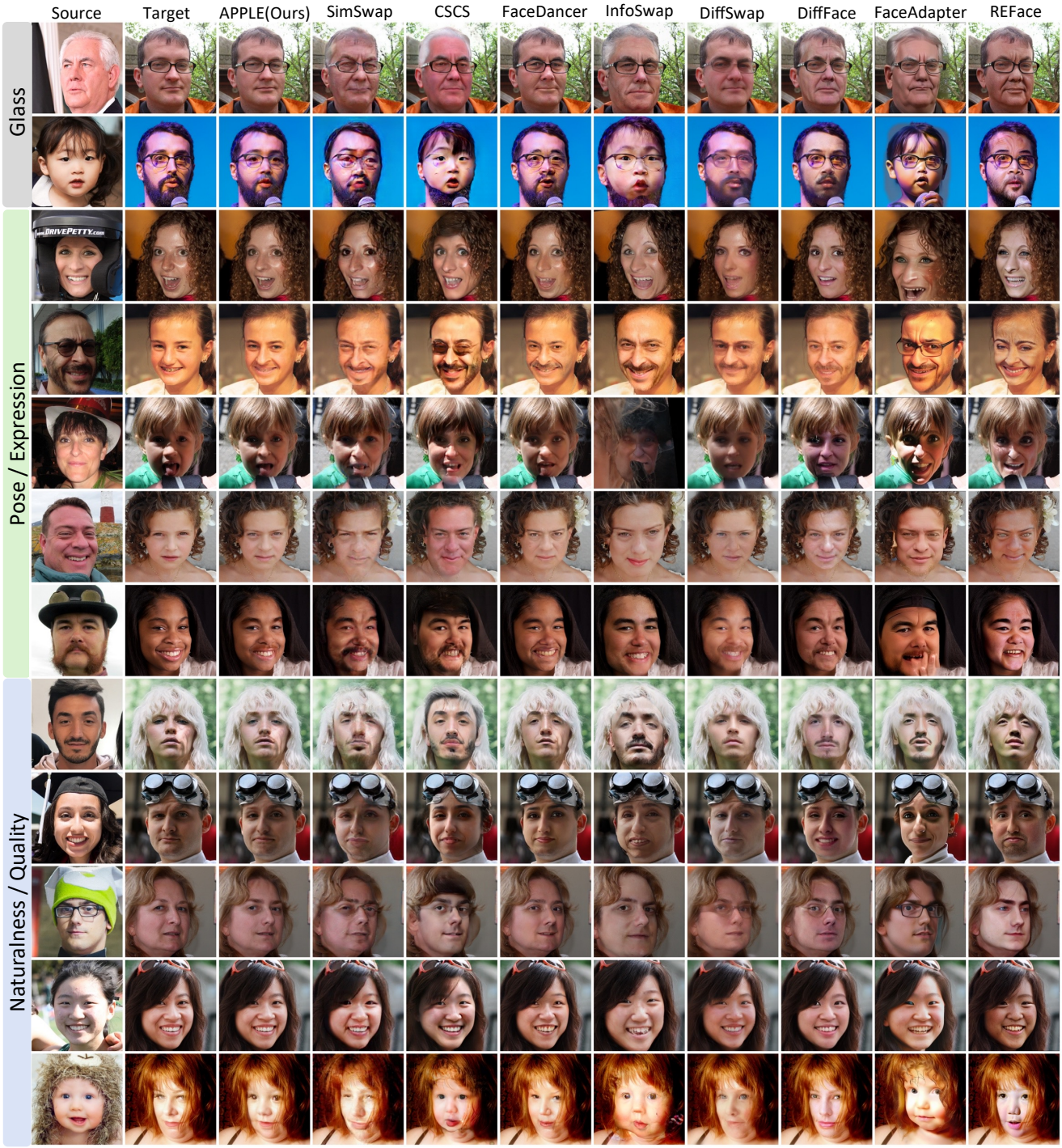

Figure 15. **Qualitative results.** Face swapping results on the FFHQ dataset [23] compared with existing baselines. APPLE effectively preserves the target image's attributes while faithfully transferring the source identity.

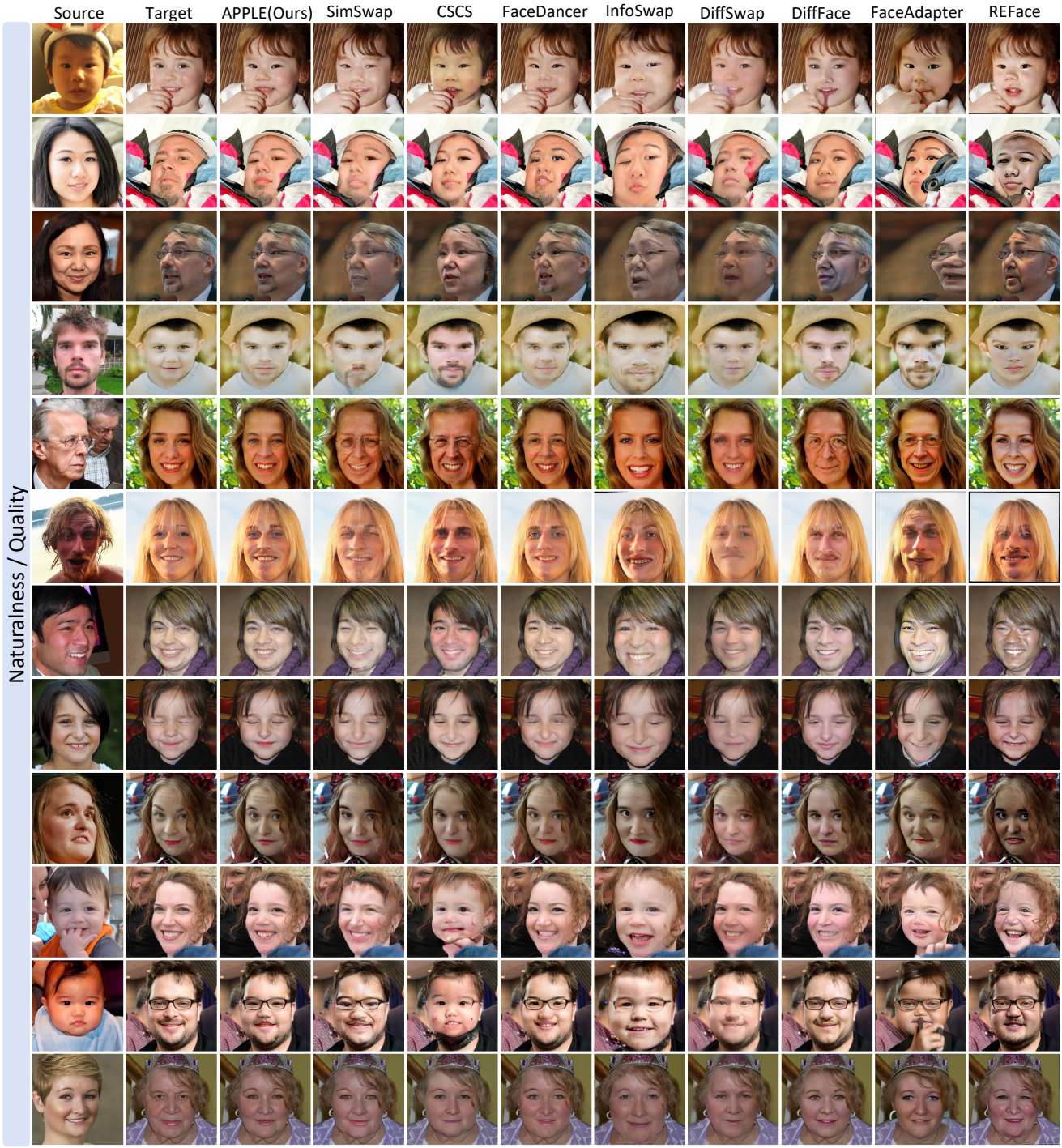

Figure 16. **Qualitative results.** Face swapping results on the FFHQ dataset [23] compared with existing baselines. APPLE effectively preserves the target image's attributes while faithfully transferring the source identity.

Figure 17. **Qualitative results.** Face swapping results on the FFHQ dataset [23] compared with existing baselines. APPLE effectively preserves the target image's attributes while faithfully transferring the source identity.

References

[1] Artsiom Ablavatski, Andrey Vakunov, Ivan Grishchenko, Karthik Raveendran, and Matsvei Zhdanovich. Real-time pupil tracking from monocular video for digital puppetry. *arXiv preprint arXiv:2006.11341*, 2020. 5

[2] Donghoon Ahn, Jiwon Kang, Sanghyun Lee, Jaewon Min, Minjae Kim, Wooseok Jang, Hyoungwon Cho, Sayak Paul, SeonHwa Kim, Eunju Cha, et al. A noise is worth diffusion guidance. *arXiv preprint arXiv:2412.03895*, 2024. 5

[3] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 1, 6, 2, 5

[4] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 157–164. 2023. 2, 5

[5] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2003–2011, 2020. 1, 2, 7, 3, 6

[6] Xuanhong Chen, Bingbing Ni, Yutian Liu, Naiyuan Liu, Zhilin Zeng, and Hang Wang. Simswap++: Towards faster and high-quality identity swapping. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1):576–592, 2023. 2, 6, 8

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 6, 3, 4

[8] Yingying Deng, Xiangyu He, Changwang Mei, Peisong Wang, and Fan Tang. Fireflow: Fast inversion of rectified flow for image semantic editing. In *Forty-second International Conference on Machine Learning*. 4, 5

[9] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 6

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural information processing systems*, 34:8780–8794, 2021. 2

[11] Bardia Doosti, Shujon Naha, Majid Mirbagheri, and David J Crandall. Hope-net: A graph-based model for hand-object pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6608–6617, 2020. 6

[12] 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. 1

[13] Gege Gao, Huaibo Huang, Chaoyou Fu, Zhaoyang Li, and Ran He. Information bottleneck disentanglement for identity swapping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3404–3413, 2021. 1, 2, 7, 5, 6

[14] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 1, 2

[15] 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. 6, 2

[16] Yue Han, Junwei Zhu, Keke He, Xu Chen, Yanhao Ge, Wei Li, Xiangtai Li, Jiangning Zhang, Chengjie Wang, and Yong Liu. Face-adapter for pre-trained diffusion models with fine-grained id and attribute control. In *European Conference on Computer Vision*, pages 20–36. Springer, 2024. 2, 3, 4, 7, 6

[17] 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. 4, 5

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 1, 2, 3

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[20] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 2

[21] Ziyao Huang, Fan Tang, Yong Zhang, Juan Cao, Chengyu Li, Sheng Tang, Jintao Li, and Tong-Yee Lee. Identity-preserving face swapping via dual surrogate generative models. *ACM Transactions on Graphics*, 43(5):1–19, 2024. 6

[22] Ziyao Huang, Fan Tang, Yong Zhang, Juan Cao, Chengyu Li, Sheng Tang, Jintao Li, and Tong-Yee Lee. Identity-preserving face swapping via dual surrogate generative models. *ACM Transactions on Graphics*, 43(5):1–19, 2024. 2, 7, 5, 6

[23] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2, 6, 5, 8, 9, 10, 11

[24] Kihong Kim, Yunho Kim, Seokju Cho, Junyoung Seo, Jisu Nam, Kychul Lee, Seungryong Kim, and KwangHee Lee. Diffface: Diffusion-based face swapping with facial guidance. *Pattern Recognition*, 163:111451, 2025. 3, 6, 7

[25] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*, 2019. 2, 3

[26] Ruibin Li, Ruihuang Li, Song Guo, and Lei Zhang. Source prompt disentangled inversion for boosting image editability with diffusion models. In *European Conference on Computer Vision*, pages 404–421. Springer, 2024. 5

[27] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 3

[28] Zhian Liu, Ming Li, Yibing Zhang, Chen Wang, Qi Zhang, Jun Wang, and Yongwei Nie. Fine-grained face swapping via regional gan inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8578–8587, 2023. 2, 6, 7

[29] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6038–6047, 2023. 4, 5

[30] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7184–7193, 2019. 2

[31] Xu Peng, Junwei Zhu, Boyuan Jiang, Ying Tai, Donghao Luo, Jiangning Zhang, Wei Lin, Taisong Jin, Chengjie Wang, and Rongrong Ji. Portraitbooth: A versatile portrait model for fast identity-preserved personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27080–27090, 2024. 5

[32] 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 (Cornell University)*, 2023. 2

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 5

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 4

[36] 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*, pages 10684–10695, 2022. 1, 2

[37] Felix Rosberg, Eren Erdal Aksoy, Fernando Alonso-Fernandez, and Cristofer Englund. Facedancer: Pose-and occlusion-aware high fidelity face swapping. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3454–3463, 2023. 2, 3, 6, 7, 8

[38] Litu Rout, Yujia Chen, Nataniel Ruiz, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Semantic image inversion and editing using rectified stochastic differential equations. In *The Thirteenth International Conference on Learning Representations*. 4, 5

[39] Christoph Schuhmann. Improved aesthetic predictor. https://github.com/christophschuhmann/improved-aesthetic-predictor, 2022. 6, 2

[40] Hao Shao, Shulun Wang, Yang Zhou, Guanglu Song, Dailan He, Zhuofan Zong, Shuo Qin, Yu Liu, and Hongsheng Li. Vividface: A robust and high-fidelity video face swapping framework. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 1

[41] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. *arXiv (Cornell University)*, 2015. 2

[42] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *ArXiv*, abs/2010.02502, 2020. 4

[43] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Neural Information Processing Systems*, pages 11895–11907, 2019. 2

[44] 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. 3

[45] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv (Cornell University)*, 2020. 2

[46] Łukasz Staniszewski, Łukasz Kuciński, and Kamil Deja. There and back again: On the relation between noise and image inversions in diffusion models. *arXiv preprint arXiv:2410.23530*, 2024. 5

[47] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14940–14950, 2025. 6, 2

[48] Kenny TR Voo, Liming Jiang, and Chen Change Loy. Delving into high-quality synthetic face occlusion segmentation datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4711–4720, 2022. 5, 6

[49] Yuhan Wang, Xu Chen, Junwei Zhu, Wenqing Chu, Ying Tai, Chengjie Wang, Jilin Li, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Hififace: 3d shape and semantic prior guided high fidelity face swapping. *arXiv preprint arXiv:2106.09965*, 2021. 1, 2, 7, 6

[50] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. In *International Conference on Learning Representations*, 2022. 5

[51] Chenxi Xie, Minghan Li, Shuai Li, Yuhui Wu, Qiaosi Yi, and Lei Zhang. Dnaedit: Direct noise alignment for text-guided rectified flow editing. *arXiv preprint arXiv:2506.01430*, 2025. 4, 5

[52] Fulong Ye, Miao Hua, Pengze Zhang, Xinghui Li, Qichao Sun, Songtao Zhao, Qian He, and Xinglong Wu. Dreamid: High-fidelity and fast diffusion-based face swapping via triplet id group learning. *arXiv preprint arXiv:2504.14509*, 2025. 1, 3, 4, 6, 8

[53] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation

network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018. 5

[54] Jongmin Yu, Hyeontaek Oh, Zhongtian Sun, Younkwan Lee, and Jinhong Yang. Real-time, high-fidelity face identity swapping with a vision foundation model. *IEEE Access*, 2025. 2, 3, 4, 6, 7, 5

[55] Ge Yuan, Maomao Li, Yong Zhang, and Huicheng Zheng. Reliableswap: Boosting general face swapping via reliable supervision. *arXiv preprint arXiv:2306.05356*, 2023. 1, 2, 5

[56] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[57] Wenliang Zhao, Yongming Rao, Weikang Shi, Zuyan Liu, Jie Zhou, and Jiwen Lu. Diffswap: High-fidelity and controllable face swapping via 3d-aware masked diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8568–8577, 2023. 2, 3, 4, 6, 7, 5

[58] Yuhao Zhu, Qi Li, Jian Wang, Cheng-Zhong Xu, and Zhenan Sun. One shot face swapping on megapixels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4834–4844, 2021. 1, 7, 6