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

Jiwon Kang*[1] Yeji Choi*[1] JoungBin Lee[1] Wooseok Jang[1]
Jinhyeok Choi[1] Taekeun Kang[2] Yongjae Park[2] Myungin Kim[2] Seungryong Kim[1]
[1]KAIST AI [2]SAMSUNG
https://cvlab-kaist.github.io/APPLE

Figure 1. **APPLE** (**A**ttribute-**P**reserving **P**seudo-**L**ab**e**ling) successfully transfers the identity of a source (**top left**) onto a target (**bottom left**) while accurately preserving target attributes (e.g. pose, expression, skin tone, lighting) across ethnicity, input variations, and gender.

Abstract

Face swapping aims to transfer the identity of a source face onto a target face while preserving target-specific attributes such as pose, expression, lighting, skin tone, and makeup. However, since real ground truth for face swapping is unavailable, achieving both accurate identity transfer and high-quality attribute preservation remains challenging. Recent diffusion-based approaches attempt to improve visual fidelity through conditional inpainting on masked target images, but the masked condition removes crucial appearance cues, resulting in plausible yet misaligned attributes. To address this limitation, we propose ***APPLE*** *(****A****ttribute-****P****reserving* ***P****seudo-****L****ab****e****ling), a fully diffusion-based teacher–student framework for attribute-preserving face swapping. Our approach introduces a teacher design to produce pseudo-labels aligned with the target attributes through (1) a conditional deblurring formulation that improves the preservation of global attributes such as skin tone and illumination, and (2) an attribute-aware inversion scheme that further enhances fine-grained attribute preservation such as makeup. APPLE conditions the student on clean pseudo-labels rather than degraded masked inputs, enabling more faithful attribute preservation. As a result, APPLE achieves state-of-the-art performance in attribute preservation while maintaining competitive identity transferability.*

1. Introduction

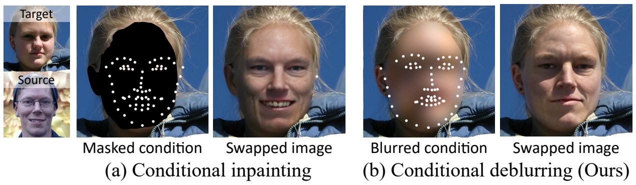
Face swapping aims to replace the identity of a person in a target image with that of a source image while faithfully preserving target-specific attributes such as pose, expression, skin tone, lighting, gaze, makeup, and accessories. It is widely applied in digital content creation, privacy protection, and film production, emphasizing the importance of generating high-quality outputs. However, since real face swapping data do not exist, supervised training of face swapping models is fundamentally infeasible, making it challenging to ensure accurate identity transfer and consistent attribute preservation at high fidelity.

Early face swapping approaches [5, 13, 49, 55, 58] primarily adopt GAN-based methods [14], leveraging the generative capability of GANs to perform identity replacement through facial editing. However, these models typically rely on two conflicting objectives: an identity loss that enforces source identity transfer, and reconstruction or attribute-related losses that encourage target attribute preservation. Such indirect and competing supervision makes GAN training unstable and requires extensive hyperparameter tuning, often resulting in copy-and-paste-like artifacts and visually unnatural outputs [40, 52].

Recently, the strong generative priors of diffusion models [3, 12, 18, 36] have driven significant progress in various image synthesis tasks, providing high-quality synthesis, precise conditional control, and reliable training stability. Motivated by these advantages, diffusion-based face-

---

*Equal contribution.



(a) Conditional inpainting (b) Conditional deblurring (Ours)

Figure 2. **Comparison of conditioning methods.** Compared to conditional inpainting widely used in existing works [16, 54, 57], the proposed conditional deblurring strategy achieves largely improved attribute (e.g., lighting) preservation of targets.

swapping methods [16, 54, 57] formulate the task as conditional inpainting. Under this proxy training objective, the model is trained to reconstruct the original clean target image from its masked version, given a source image of the same identity but with different attributes. However, this masked conditioning removes crucial cues from the target, such as lighting, skin tone, makeup, and subtle expression dynamics. Consequently, even with auxiliary attribute information like 3DMM [4] landmarks or CLIP [33] features, existing models often fail to preserve target-specific attributes during face swapping.

To address the issue of misaligned attributes, we propose **APPLE** (**A**ttribute-**P**reserving **P**seudo-**L**ab**e**ling), a fully diffusion-based teacher–student framework for attribute-preserving face swapping. The core idea is to train a teacher model to generate high-quality pseudo-labels aligned with the target's visual attributes and use them as conditioning inputs for the student. Unlike conventional inpainting-based methods that rely on masked conditions, the student is conditioned on clean, unmasked pseudo-labels that provide richer attribute cues and is supervised under a direct image-editing objective to reconstruct the original target image, thereby achieving high-fidelity attribute preservation.

Importantly, the effectiveness of this training scheme depends on the teacher's ability to generate pseudo-labels that maintain attribute consistency with the target image. Without the consistency, the student receives conflicting signals (e.g., in pose or lighting), which degrades attribute preservation. To this end, APPLE first replaces the conventional conditional inpainting objective of the teacher diffusion model with a conditional deblurring formulation, leading to improved preservation of attributes such as skin tone and illumination. Second, to further enhance fine-grained attribute preservation during pseudo-label generation, we introduce an attribute-aware inversion scheme that intentionally exploits the reduced editability of inverted noise to anchor fine-grained details. Together, these designs enable the teacher to produce attribute-aligned pseudo-labels, allowing the student to learn under a direct editing objective and achieve reliable identity transfer with improved preservation of fine-grained target attributes.

In summary, APPLE achieves state-of-the-art face-swapping performance in terms of attribute preservation while maintaining competitive identity transferability, yielding more coherent and photorealistic images than previous methods. Our main contributions are as follows:

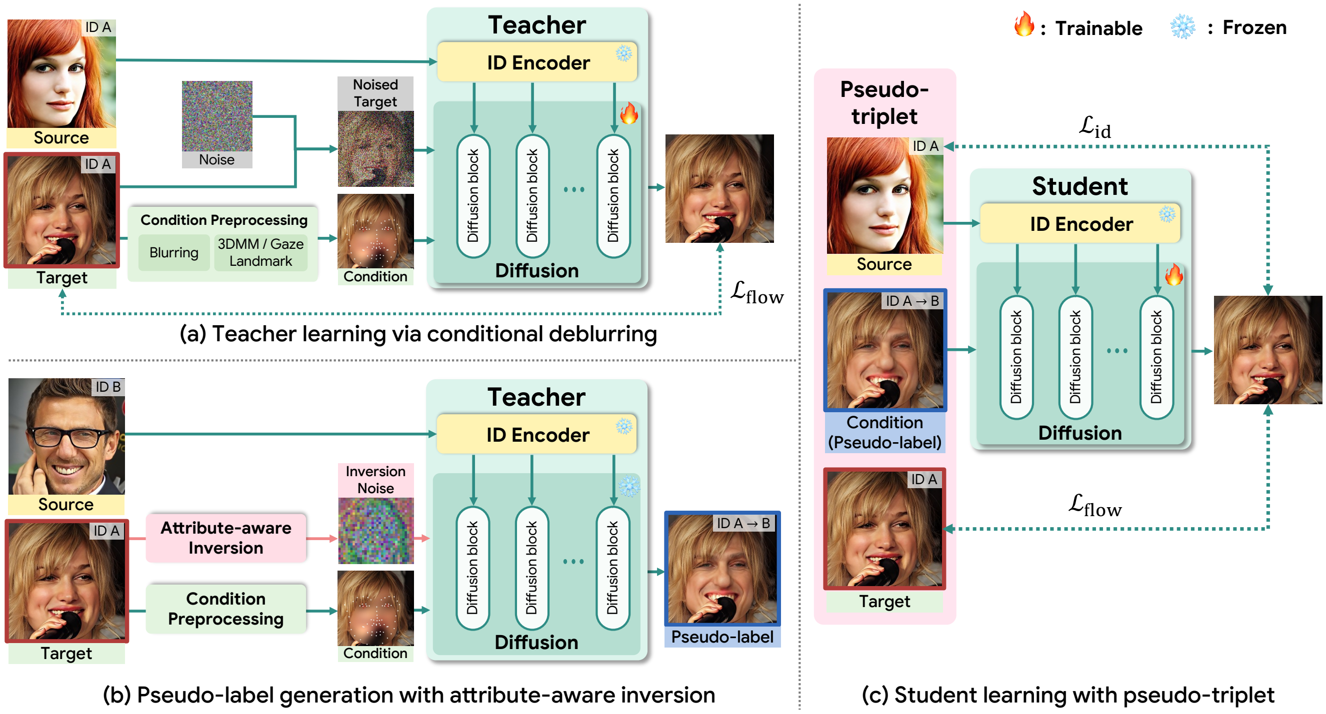
- We propose **APPLE**, a diffusion-based teacher–student framework that leverages improved pseudo-label quality as the key to achieving superior attribute preservation.
- We enhance the attribute preservation of the teacher's output by replacing the conventional conditional inpainting formulation with a *conditional deblurring* objective during training and introducing an *attribute-aware inversion* scheme during inference.
- We demonstrate that by training on these high-fidelity pseudo-labels, the student model achieves state-of-the-art attribute-preservation while maintaining high identity similarity.
- Our framework offers high practical value for real-world deployment, as the student model requires no complex preprocessing for attribute conditioning at inference.

2. Related Work

GAN-based models. GANs [14] play a dominant role in early face-swapping research [5, 13, 25, 30]. FSGAN [30] introduces an identity-agnostic framework that fuses source identity with target attributes through adaptive blending. SimSwap [5] proposes an ID-injection module that embeds identity information and applies a weak feature-matching loss to preserve target attributes, while SimSwap++ [6] further enhances the model's efficiency. To improve robustness under pose and expression variations, HiFiFace [49] incorporates 3D priors (e.g., 3DMM [4]) and face-recognition constraints to achieve geometrically consistent face generation. In parallel, StyleGAN-based models [23] are widely adopted for their high-fidelity generative capability. Representative works, such as FaceDancer [37] and E4S [28], further enhance visual realism by employing adaptive feature fusion attention for hierarchical feature integration and region-based inversion within the StyleGAN latent space. On another branch, to compensate for the absence of ground-truth supervision, recent methods such as CSCS [22] and ReliableSwap [55] generate pseudo pairs using pre-trained GAN models to augment training data and improve identity consistency. However, such pseudo samples often suffer from attribute misalignment and visual artifacts inherited from the GAN generator, limiting their effectiveness as reliable supervision.

Overall, GAN-based approaches still rely on complex loss balancing and extensive hyperparameter tuning, often failing to reproduce fine-grained details and exhibiting local artifacts under extreme pose or expression variations, which limit both naturalness and photorealism.

Diffusion-based models. Diffusion models [10, 18, 32, 36, 41–43, 45] have recently emerged as a powerful alternative to GANs, providing superior image quality and



(a) Teacher learning via conditional deblurring

(b) Pseudo-label generation with attribute-aware inversion

(c) Student learning with pseudo-triplet

Figure 3. **Overall architecture of the proposed method.** We propose **APPLE**, a diffusion-based teacher-student framework that focuses on improving attribute preservation. (a) To improve target-attribute preservation, we propose training a teacher with conditional deblurring rather than conditional inpainting widely used in existing works [16, 54, 57]. (b) When constructing a pseudo-label with a teacher, we propose *attribute-aware inversion* which further improves fine-grained attribute preservation in inference time. Note that inversion noise cannot be used during training due to its non-Gaussian property. (c) The student model is trained using attribute-aligned pseudo-labels generated by the teacher. By leveraging high-fidelity attribute conditioning—clean images rather than degraded inputs—the student model eventually outperforms the teacher, achieving state-of-the-art attribute-preservation while maintaining high identity similarity. Note that the noised target input to the diffusion model is omitted in (c).

semantic controllability. Following this trend, various studies have adopted diffusion-based frameworks for face-swapping tasks. DiffFace [24] first introduces an identity-conditional DDPM [18] that incorporates multiple facial guidance signals, including semantic parsing and gaze direction. However, these cues are only applied during training and are not utilized at inference time, resulting in noticeable noise artifacts, particularly around the eyes. To alleviate this issue, subsequent approaches, such as FaceAdapter [16], DiffSwap [57], and ReFace [54], reformulate the task as conditional inpainting, where the target facial region is masked and structural priors such as facial landmarks are used as conditioning signals. Although this strategy helps prevent identity leakage from the target, masking out the entire face inevitably removes critical visual cues such as illumination, makeup, and accessories, making it difficult to preserve the target's fine-grained attributes. More recently, DreamID [52] addresses this issue by constructing pseudo datasets using a GAN-based face-swapping model [37] and training a diffusion network for identity transfer. While effective, it offers limited exploration into how to build high-quality pseudo triplets, particularly regarding attribute preservation. In contrast, our framework progressively mitigates these limitations through a teacher–student design. Crucially, we focus on improving the diffusion teacher itself so that it can generate high-fidelity, attribute preserving pseudo triplets, achieving significantly better performance than simply relying on an off-the-shelf face-swapping model.

3. Preliminaries

Rectified flow. Diffusion models [19, 44] learn to generate data by gradually denoising samples drawn from a Gaussian prior through a stochastic process. More recently, *rectified flow* [27] reformulated this process as a deterministic flow, simplifying sampling while maintaining distributional expressiveness. Specifically, it defines a linear interpolation between a noise sample $\epsilon \sim \mathcal{N}(0, 1)$ and a real sample $x_0 \sim p_0(x)$ at a timestep $t \in [0, 1]$ as:

$$z_t = (1 - t)\, x_0 + t\, \epsilon. \tag{1}$$

The model learns to predict the velocity $v_t(x)$ that transports x_0 toward ϵ (or vice versa) along this linear path using a conditional flow-matching objective:

$$\mathcal{L}_{\text{flow}} = \mathbb{E}_{t, x_0 \sim p_0(x)} \left[\|(\epsilon - x_0) - v_t(z_t)\|^2 \right]. \tag{2}$$

For the rectified flow models, intermediate prediction $\hat{x}_0$ is obtained by following equation:

$$\hat{x}_0(z_t) = z_t - t\, v_t(z_t). \tag{3}$$

Note that we omit the VAE encoder–decoder [35] for notational simplicity.

Diffusion inversion. Inversion maps a real image back to its underlying Gaussian noise representation. When performed correctly, it yields an initial noise vector that can faithfully reconstruct the reference image, providing a reliable starting point for subsequent editing or manipulation. For rectified flow models, this can be accomplished similarly to DDIM inversion [42]. The latent trajectory is estimated by iteratively adding noise according to

$$z_{t+\Delta t} = z_t + \Delta t \cdot v_t(z_t), \tag{4}$$

where $v_t(\cdot)$ denotes the learned velocity or score field. By anchoring the generative path to real observations, inversion provides a crucial foundation for a wide range of image editing tasks [8, 17, 29, 38, 51].

4. Methodology

4.1. Overview

Sec. 4.2 formalizes the conventional conditional diffusion formulation for face swapping and defines the identity and attribute conditioning that our models follow. Sec. 4.3 presents our attribute-preserving diffusion teacher, trained with conditional deblurring. Sec. 4.4 introduces attribute-aware inversion scheme to improve fine-grained attribute preservation. Sec. 4.5 then describes how a student diffusion model is trained with pseudo-label in a direct image-editing setting, leading to improved attribute fidelity while maintaining identity consistency.

4.2. Problem formulation

Given a source and target image pair $\{I_{\text{src}}, I_{\text{tgt}}\}$, the face-swapping model aims to generate a swapped image $\hat{I}$ that reflects the identity of I_{src} and the attributes of I_{tgt}, ensuring both identity fidelity and attribute consistency.

Recent diffusion-based face swapping methods [16, 52, 57] commonly follow a conditional diffusion framework in which identity and attribute information are injected into the model through separate conditioning pathways. To formalize this conventional design and establish a unified notation, we define two independent conditioning functions, each responsible for identity and attribute representation:

$$\mathbf{id}_{\text{src}} = \mathcal{F}_{\text{id}}(I_{\text{src}}), \qquad \mathbf{att}_{\text{tgt}} = \mathcal{F}_{\text{att}}(I_{\text{tgt}}), \tag{5}$$

where $\mathcal{F}_{\text{id}}(\cdot)$ extracts identity-related features (e.g., identity embeddings from a face recognition network), and $\mathcal{F}_{\text{att}}(\cdot)$ encodes attribute-related representations from the target (e.g., structural cues such as pose, expression, or lighting).

During training, source-target image pair shares the same identity. Formally, the diffusion process for the face-swapping task is parameterized by a velocity field $v_t(\cdot)$ trained under a flow-matching objective that drives the model to reconstruct the target image conditioned on the source identity and target attributes:

$$\mathcal{L}_{\text{flow}} = \mathbb{E}_{t,\, I,\, \epsilon} \left[\|(\epsilon - I_{\text{tgt}}) - v_t(z_t,\, \mathbf{id}_{\text{src}},\, \mathbf{att}_{\text{tgt}})\|_2^2 \right], \tag{6}$$

where $\epsilon \sim \mathcal{N}(0, I)$ represents Gaussian noise, and z_t is the interpolated latent between I_{tgt} and ϵ at timestep t.

Following prior diffusion-based face swapping methods, we employ an identity loss to encourage the swapped output to match the source identity. The identity loss is defined as:

$$\mathcal{L}_{\text{id}} = 1 - \cos\big(\mathcal{F}_{\text{id}}\big(\hat{x}_0(z_t)\big), \mathcal{F}_{\text{id}}\big(I_{\text{src}}\big)\big), \tag{7}$$

The overall training objective for the diffusion model combines the flow-matching loss with the identity loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{flow}} + \lambda_{\text{id}}\, \mathcal{L}_{\text{id}}. \tag{8}$$

During inference, face swapping is achieved by using a source-target image pair with different identities.

4.3. Teacher learning via conditional deblurring

In our framework, the attributes include pose, facial expression, gaze direction, skin tone, makeup style, illumination, accessories, and background appearance. To ensure faithful preservation of these attributes, our framework introduces a carefully designed conditioning mechanism, enabling the teacher to preserve the aforementioned attributes while maintaining identity transferability.

Conditional deblurring. Most diffusion-based face-swapping models [16, 54, 57] formulate the task as conditional *inpainting*, where the facial region of the target image is masked and supervised to reconstruct target with source identity. Motivation for masking is to suppress identity information from the target image. Without masking, the model tends to reconstruct the target identity rather than learning to transfer the source identity. While masking effectively prevents identity leakage, it also removes important attribute cues—such as lighting, skin tone, makeup, and accessories—that are crucial for faithful attribute preservation. As a result, inpainting-based diffusion models often fail to reproduce fine-grained target styles and yield inconsistent attribute appearance in the swapped face.

To overcome this limitation, we reformulate the training objective as a conditional *deblurring* task. Instead of masking the entire facial region, we replace it with a blurred version of the target image, which removes high-frequency identity details while retaining low-frequency but informative attribute cues, including pose, lighting, and expression. This reformulation allows the model to exploit the target's contextual attribute information more effectively during training, leading to superior preservation of lighting, skin tone, and structural attributes while maintaining strong identity transferability, as shown in Fig. 2

Specifically, we downsample the target image to a resolution of 8 × 8 and then upsample it back to the original size, effectively eliminating fine-grained identity details while preserving low-frequency appearance such as color tone and lighting. A facial mask extracted from a face parsing model [48] is then used to apply the blurring only to the facial region, ensuring that background and contextual information remain unaffected.

Enriching semantic conditions. To maintain structural consistency and realism, we enrich the semantic conditions by incorporating multi-level structural cues beyond standard 3DMM [4] landmarks. While gaze is a critical factor, existing diffusion-based models often struggle to preserve gaze alignment or introduce artifacts when using gaze loss as sampling guidance. To address these limitations, we overlay eye landmarks derived from a gaze estimator [1] and glass segmentation masks obtained via face parsing [53] onto the blurred target condition. This integrated approach provides the diffusion model with both coarse attribute context and fine structural details, enabling the faithful preservation of pose, gaze, and accessories during face swapping.

4.4. Pseudo-label generation with inversion

Although the proposed deblurring strategy and semantic condition used in training the teacher model effectively improve attribute preservation, there remains room to enhance further the reconstruction of fine-grained details such as makeup and accessories. This is because the model is required to implicitly infer high-frequency details that are not explicitly present in the blurred inputs.

Attribute-aware inversion. To address this issue, we draw inspiration from inversion-based editing techniques [8, 17, 29, 38, 51], which enable more precise control over fine-grained attributes by aligning the generation process with latent representations. Recent studies have shown that noise obtained through diffusion inversion deviates from the ideal Gaussian prior [2, 46] and retains residual semantic information from the input, such as image structure or prompt-related signals [26, 46]. While prior work seeks to suppress this residual information due to its negative impact on editability [46], we instead exploit it. In face swapping, where the goal is to modify identity while preserving target-specific attributes, retaining such information in the noise is beneficial rather than detrimental.

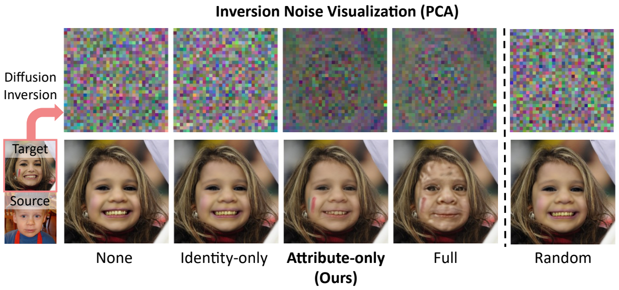


Figure 4. **Comparison of conditioning configuration for inversion.** (Top) PCA visualization of target-inverted noise and random Gaussian noise. When attribute-only conditioning is used, inverted noise encodes more semantic information compared to the others. (Bottom) Results of face swapping when each inverted noise is used. Using attribute-only conditioned noise yields the most makeup preserved results without introducing artifacts.

Motivated by this insight, we propose *attribute-aware inversion* which inverts the target image under attribute-only conditioning, yielding noise representations that are enriched with target attributes but free from target identity bias. Using these attribute-aware noise at inference time enables the model to better preserve fine-grained appearance characteristics that are often lost due to blurring, while maintaining strong identity transferability.

To validate our choice, we conduct an empirical analysis comparing four conditioning configurations that can be used in inversion, formally defined as

$$(\mathcal{F}_{\text{id}}(I), \mathcal{F}_{\text{att}}(I)), (\varnothing, \mathcal{F}_{\text{att}}(I)), (\mathcal{F}_{\text{id}}(I), \varnothing), (\varnothing, \varnothing)$$

where each element corresponds to full conditioning, attribute-only conditioning, identity-only conditioning, and non-conditioning.

We first analyze how each configuration affects the inverted noise by performing PCA on the resulting noise representations. The non-conditioned and identity-only configurations produce noise distributions with little semantic structure, similar to random noise. In contrast, configurations that incorporate attribute conditioning yield noise patterns that exhibit clear facial semantics, indicating that attribute information is effectively retained during inversion.

We additionally compare the swapped outputs obtained from each inversion configuration. Both attribute-only and full conditioning preserve fine-grained attributes such as makeup more reliably than the other variants. However, full conditioning often produces artifacts in the swapped images. We attribute this to residual identity information embedded in the full-conditioned inverted noise, which restricts editability and interferes with identity replacement. In contrast, attribute-only conditioning $(\varnothing, \mathcal{F}_{\text{att}}(I))$ avoids identity-related biases while still embedding meaningful attribute cues. This allows the model to preserve target attributes better during swapping, without introducing artifacts. Based on these observations, we adopt attribute-only conditioning as our inversion strategy.

4.5. Student learning with pseudo-triplet

We employ the teacher model to synthesize a pseudo-label with the attribute-aware inversion scheme. Specifically, by performing face swapping on a target image I^{A}_{tgt} of identity A with another subject B via the teacher model, we obtain a pseudo-label $\hat{I}^{A\to B}_{\text{tgt}}$ that is used to construct the pseudo-triplet $(I^{A}_{\text{src}}, \hat{I}^{A\to B}_{\text{tgt}}, I^{A}_{\text{tgt}})$. The triplet is then used to train the student under an explicit editing objective. Specifically, the student takes an identity feature from source image I^{A}_{src} and an attribute feature from pseudo-label $I^{A\to B}_{\text{tgt}}$ as input and learns to reconstruct the original target I^{A}_{tgt}.

By designing the student model to accept pseudo-labels in their raw image format as attribute conditioning inputs, we achieve two significant benefits. First, during the training phase, the model leverages these high-fidelity, unmasked inputs instead of the degraded images used in conventional approaches. This allows the student model to learn superior attribute-preservation more effectively. Second, this design removes the need for auxiliary networks or complex preprocessing pipelines for attribute conditioning, thereby improving practicality in real-world deployment.

5. Experiments

5.1. Experimental setups

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 LoRA rank of 64. The model is trained on the VGGFace2-HQ dataset [6] which is filtered with AES [39] to ensure high-quality faces, with threshold 5.1. Following REFace [54], the source image is masked before being fed into the identity encoder. Teacher model is trained for 15K iterations without the identity loss and an additional 50K iterations with it enabled. Student model is trained for 15K iterations, resuming from teacher. Experiments are conducted on four NVIDIA A6000 GPUs with a batch size of 1 per GPU, gradient accumulation of 4, resulting in effective batch size 16.

Evaluation protocol. Following previous studies [21, 52], we select 1,000 source faces and 1,000 target faces from the FFHQ dataset [23], and generated 1,000 corresponding face-swapped results. To assess image fidelity, we compute Frechet inception distance(FID) between the face-swapped images and the real FFHQ images. For pose and expression consistency, we employ HopeNet [11] and Deep3DFaceRecon [9], respectively, and measure the L2 distance between each target image and its corresponding swapped image. To evaluate identity preservation, ArcFace [7] is to extract identity embeddings, and the cosine distance is calculated between the embeddings of the swapped and source faces. To compute ID Retrieval, we find the most similar source faces based on cosine similarity, and report Top-1 and Top-5 accuracy. All experiments, including training and inference, are performed on images of resolution 512 × 512.

Table 1. **Ablation of teacher model.** Applying the proposed components sequentially leads to gradual performance improvements. In particular, the attribute-aware inversion module substantially improves FID, pose, and expression metrics while maintaining identity similarity at a comparable level.

Model	**FID↓**	**ID Sim.↑**	**ID Retrieval↑ Top-1 / Top-5**	**Pose↓**	**Expr.↓**
Inpainting (Baseline)	11.00	0.54	92.80 / 96.90	3.37	1.01
Deblurring	4.20	0.53	89.50 / 96.20	2.58	0.79
Deblurring + Inv.	**3.68**	0.54	90.40 / 96.70	**2.07**	**0.70**

Table 2. **Ablation of conditioning configurations for inversion.** For inversion, using attribute-only condition improves FID, pose and expression effectively while not sacrificing identity transferability. Applying the other conditions does not robustly improve FID, pose or expression.

Condition config.	**FID↓**	**ID Sim.↑**	**ID Retrieval↑ Top-1 / Top-5**	**Pose↓**	**Expr.↓**
Baseline	4.20	0.53	89.50 / 96.20	2.58	0.79
+ None	6.20	0.52	87.90 / 95.80	**2.03**	0.74
+ Identity-only	10.02	0.53	90.50 / 95.90	2.57	0.83
+ Attribute-only	**3.68**	0.54	90.40 / 96.70	2.07	**0.70**
+ Full	10.51	0.53	89.10 / 93.80	3.13	0.99

Table 3. **Ablation of pseudo-label quality.** We compare the performance of student models trained on pseudo datasets generated by FaceDancer [37] and by our teacher model. The model trained on our pseudo-triplet achieves superior attribute-preserving results indicated by lower pose and expression consistency, while maintaining comparable identity similarity. Results indicate that our teacher model serves as a more effective teacher.

Teacher model	**FID↓**	**ID Sim.↑**	**ID Retrieval↑**	**Pose↓**	**Expr.↓**
FaceDancer [37]	2.47	0.53	87.50 / 95.60	2.07	0.65
APPLE (Teacher)	**1.98**	0.53	88.60 / 95.80	**1.77**	**0.62**

5.2. Comparison with other models

We compare our method both diffusion-based and GAN-based approaches. Across all experiments, our model delivers substantially stronger attribute preservation while maintaining identity similarity competitive with the best existing methods. It is worth noting that several baselines, including DiffSwap [57], DiffFace [24], and E4S [28], are trained directly on FFHQ and therefore benefit from favorable domain alignment with the evaluation benchmark.

Qualitative comparison. We present qualitative results in Fig. 5. Our model maintains subtle cues such as tongue-out *expressions*, fine-grained *makeup* details, and *accessory* (e.g. glass), while still performing reliable identity transfer. Competing diffusion-based methods often fail in these scenarios by either erasing attributes or inventing new ones.

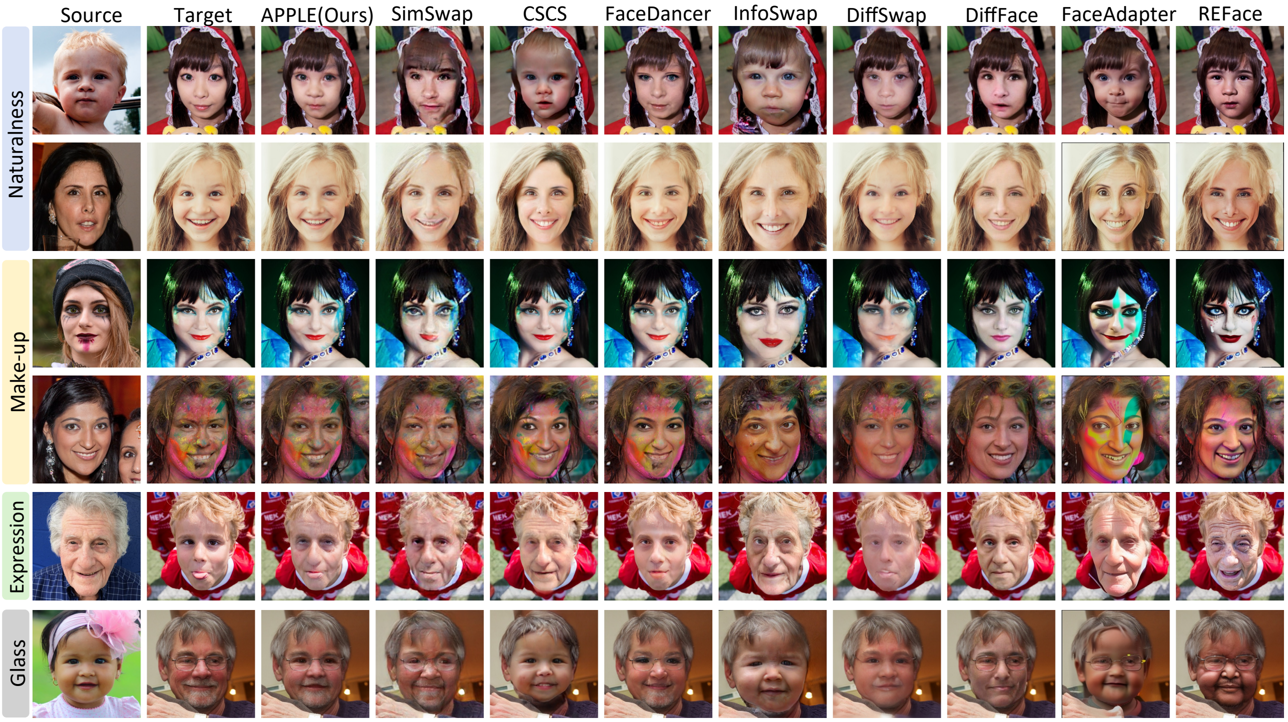


Figure 5. **Qualitative results on FFHQ.** Compared to existing baselines, APPLE effectively preserves the target image's attributes while faithfully transferring the source identity.

Table 4. **Quantitative results on FFHQ.** While maintaining identity transferability, APPLE achieves superior attribute preserving performance, which is evidenced by lowest pose and expression error.

Model	**FID↓**	**ID Sim.↑**	**ID Retrieval↑ Top-1 / Top-5**	**Pose↓**	**Expr.↓**
SimSwap [5]	18.54	0.55	94.10 / 99.00	3.11	1.73
CSCS [22]	11.00	0.65	99.00 / 99.50	3.64	1.44
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
APPLE (Teacher)	3.68	0.54	90.40 / 96.70	2.07	0.70
APPLE (Student)	**2.18**	0.54	90.50 / 97.00	**1.85**	**0.64**

GAN-based models preserve attributes better than existing diffusion-based models because their generators observe the full target image during self-supervised training. However, their outputs frequently contain artifacts, color inconsistencies, and unnatural textures due to the inherent limitations of adversarial training. In contrast, our method produces clean and realistic results while retaining the attribute fidelity even better than GAN based models.

Quantitative comparison. The quantitative results reported in Tab. 4 follow the similar trend observed in the qualitative results. Our method achieves the lowest FID among all baselines, indicating cleaner and more realistic synthesis, and shows clear improvements in pose and expression consistency while maintaining competitive identity similarity. Although CSCS [22] and REFace [54] report notably high identity similarity, this comes with a substantial drop in attribute preservation, evidenced by the copy-paste–like artifacts visible in their qualitative outputs in Fig. 5. Identity transfer and attribute preservation inherently trade off against each other, so a well-designed method should balance the two. However, they are strongly biased toward identity matching, leading to poor attribute fidelity, which is an undesirable behavior for face swapping. In contrast, APPLE achieves a more balanced trade-off, delivering reliable identity transfer while consistently preserving target attributes across diverse scenarios.

5.3. Ablation studies

In this section, we analyze the effectiveness of our teacher-student framework through component-wise ablations.

Inpainting versus deblurring. We compare the conventional inpainting with our deblurring formulation. The qualitative and quantitative results are presented in Fig. 6 and Tab. 2. Qualitative results show that transition from inpainting to deblurring substantially improves attribute preservation. Deblurring produces noticeably more consistent skin tone, lighting, pose, and expression, since the model is no

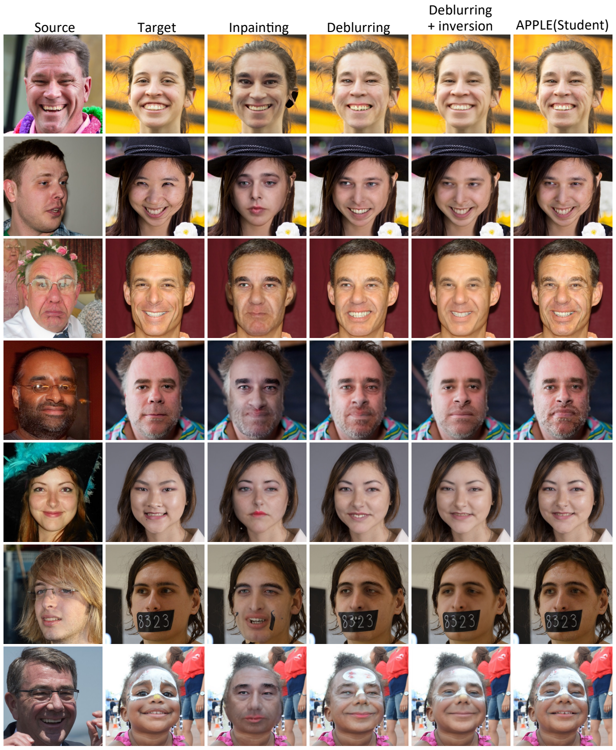


Figure 6. **Qualitative ablation results of proposed method.** The inpainting method fails to preserve the target image's attributes. Applying deblurring (4th column) and attribute-aware inversion (5th column) progressively better maintain the target attributes. Overall quality, including attribute-preserving is maximized on student (6th column).

longer required to synthesize these cues. These qualitative gains are reflected in quantitative metrics, which show huge improvements in pose and expression consistency.

Effect of attribute-aware inversion. We additionally study the impact of attribute-aware inversion qualitatively and quantitatively, presenting results presented in Fig. 6 and Tab. 2. When proposed attribute-aware inversion is used, the model better reconstructs fine-grained attributes such as makeup, jewelry, and accessories, which are difficult to recover from blurred observations alone. Quantitative results confirm that attribute-aware inversion further improves pose and expression without sacrificing identity transferability.

Various inversion strategies. We conduct a quantitative ablation on the conditioning configurations used during the inversion process. Specifically, we evaluate the result of teacher model using four types of inversion conditions and report the results in Tab. 2. The results show that using attribute-only condition $(\varnothing, \mathcal{F}_{\text{att}}(I))$ yields the robust performance, whereas the other variants introduce noticeable side effects such as increased FID, pose or expression.

Quality of pseudo-label. We further evaluate the quality of our pseudo-triplet by comparing student models trained on different pseudo-triplets. Since DreamID [52] adopts FaceDancer [37] as a teacher due to its strong attribute preservation, we compare a student trained on FaceDancer-generated pseudo-triplet with one trained on pseudo-triplet produced by our teacher. Both 50K pseudo-triplets are generated from VGGFace2-HQ [6], and all models are evaluated at 5K training steps. While the FaceDancer-based student achieves slightly higher identity similarity, the student trained on our pseudo-triplet performs better on pose, expression, and fine-grained attribute metrics. Qualitative results in Fig. 7 further show that our pseudo-triplet leads to more accurate preserving of attributes such as makeup, gaze, accessories etc. This demonstrates that pseudo-triplet generated by our teacher provides stronger supervision for training high-quality face-swapping models.

Figure 7. **Qualitative ablation results of pseudo-label quality.** We compare two student models trained on pseudo-triplet generated by FaceDancer [37] and APPLE (Teacher). Results demonstrate that student trained by APPLE (Teacher) more faithfully preserves the target attributes.

6. Conclusion

In this work, we introduced **APPLE**, a diffusion-based teacher–student framework that tackles the challenge of attribute preservation in face swapping. We designed a training and inference strategy that produces attribute-preserving pseudo-labels, enabling the student to learn in a direct image-editing setting with a clean attribute condition image. This allows the student to retain fine-grained details more faithfully than its teacher and existing diffusion- and GAN-based methods, while maintaining competitive identity transferability. Our results demonstrated that the proposed teacher–student design provides an effective path toward high-fidelity, attribute-preserving diffusion face swapping without sacrificing identity transferability.

Acknowledgements

This research was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2019-II190075, RS-2024-00509279, RS-2025-II212068, RS-2023-00227592, RS-2025-02214479, RS-2024-00457882, RS-2025-25441838, RS-2025-25441838, RS-2025-02214479, RS-2025-02217259) and the Culture, Sports, and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism (RS-2024-00345025, RS-2024-00333068, RS-2023-00222280, RS-2023-00266509), and National Research Foundation of Korea (RS-2024-00346597).

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 vi-*

sion 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