arXiv:2511.05575v2 [cs.CV] 24 Aug 2026

DiffSwap++: 3D Latent-Controlled Diffusion for Identity-Preserving Face Swapping

Weston Bondurant Arkaprava Sinha Hieu Le Srijan Das Stephanie Schuckers
The University of North Carolina at Charlotte
wbondura@charlotte.edu

Abstract

*Diffusion-based face swapping has recently demonstrated superior visual quality compared to GAN-based approaches; however, existing methods often suffer from fine-grained artifacts and weak identity preservation due to limited modeling of 3D facial structure. To address these limitations, we propose **DiffSwap++**, a novel face-swapping pipeline that leverages **3D-aware latent representations** to guide the diffusion generation process, enabling improved disentanglement of identity from pose and expression. Our method conditions the diffusion architecture on **source identity embeddings** and **target facial landmarks**, producing high-fidelity swaps that preserve the source identity while maintaining target attributes. Extensive experiments on three public datasets, supported by biometric evaluations and human user studies, demonstrate that DiffSwap++ consistently outperforms prior state-of-the-art methods in identity retention and perceptual realism. Code will be made publicly available at https://github.com/WestonBond/DiffSwapPP*

1. Introduction

Face swapping is the task of transferring the identity from a source face image onto a target face image while preserving the *background*, *pose*, and *expression* of the target. Earlier methods have shown promising results, through the use of GAN-based architectures that integrate identity features from the source image with attribute features (e.g., *pose*, *expression*) from the target image [6, 39, 20, 40]. However, GAN training is inherently unstable, and qualitative evaluations reveal noticeable limitations in the realism and consistency of generated swapped faces[26]. To address the training challenges and generation artifacts associated with GANs, recent approaches have explored the use of diffusion models for face swapping, demonstrating improved fidelity and robustness [16, 43, 1].

While diffusion-based methods generate high-fidelity face swaps through multi-step processes using either DDPM [13] or DDIM [35], they remain prone to fine-grained artifacts. This limitation arises from their strong dependence on training data that sufficiently captures such challenging scenarios. Unlike unconditional models, which generate data without explicit guidance, conditional diffusion models enable control over the generation process by incorporating auxiliary inputs such as text descriptions, class labels, or images [2]. The aforementioned face-swapping approaches such as DiffFace [16], DiffSwap [43], and REFace [1] leverage conditional diffusion frameworks using different identity conditioning strategies including identity embeddings, landmark features, or pretrained CLIP features [27]. However, these methods do not explicitly incorporate the underlying 3D facial structure, which can be critical for achieving robust, controllable, and photorealistic face-swapping results. Although DiffSwap utilizes 3D facial features, their use is limited to incorporating 3D face landmarks at inference and not directly employed to guide the training process.

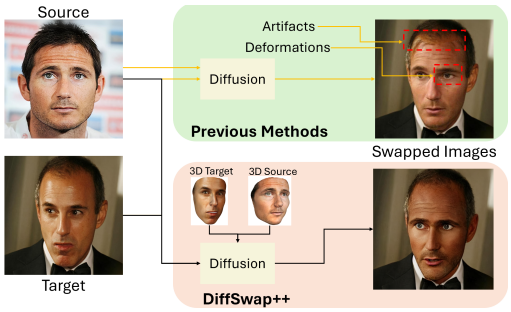


Figure 1. Comparison between **Diffswap++** and previous diffusion methods showcasing the lack of artifacts and deformations in our swapped outputs.

To address the lack of 3D awareness in diffusion-based face-swapping approaches, we propose **DiffSwap++**, a novel pipeline that incorporates 3D facial information during training, yet operates without any 3D features at inference time. By integrating 3D latent features during training, our method achieves enhanced visual realism and improved levels of identity preservation. As shown in figure 1 This 3D supervision enables the model to more effectively disentangle facial identity from pose and expression. Specif-

ically, we leverage a 3D face reconstruction model [3] to project 2D face images into 3D latent embeddings, which are then rendered back into 2D views. These 3D-aware latent features, capturing the structural layout of the face, are used to guide the diffusion model during training. Our proposed diffusion pipeline formulates face swapping as a target inpainting task, conditioned on both facial landmarks and identity embeddings.

Through extensive experiments on the CelebA [19], FFHQ [15], and CelebV-Text [41] datasets, we find that our model consistently outperforms existing methods in preserving the identity of the source face during swapping, while also maintaining comparable performance in retaining the structural attributes of the target face. Additionally, inspired by evaluation practices in biometric recognition research [25, 24], we assess the identity preservation capabilities of our model and several baselines by analyzing the trends between Impostor Attack Presentation Acceptance Rate (IAPAR) and the False Reject Rate (FRR) across varying acceptance thresholds. This biometric-style evaluation provides a more rigorous and quantitative understanding of how effectively each method retains the source identity during face swapping.

Our main contributions are summarized as follows:

- We propose **DiffSwap++**, the first face-swapping framework that leverages 3D facial latent features to guide a diffusion-based generation pipeline. Our method incorporates 3D geometric cues during training while remaining fully 2D at inference.
- We conduct extensive experiments on three publicly available datasets that demonstrate the superior identity preservation and structural consistency of our approach. Additionally, a user study confirms that DiffSwap++ produces more realistic and visually convincing face swaps compared to prior methods.
- We introduce a novel evaluation perspective for face-swapping models by analyzing trends between IAPAR and False Reject Rate FRR, offering a rigorous biometric analysis of identity preservation performance.

2. Related Work

2.1. GAN Face-Swapping Approaches

Generative Adversarial Networks, GANs are a class of deep learning models that can generate realistic images [11]. Earlier approaches to the face-swapping task utilized GANs, training them to swap the attributes of a source face image onto a target face image while preserving the pose and expression of the face in the target image [6, 39, 20, 40]. Despite promising results from many of these GAN based approaches to face-swapping, these models continued to experience issues inherent to GAN models such as a lack training stability and requiring careful hyperparameter tuning. These approaches also largely failed to produce high fidelity results that are often desired when performing face-swapping. More recently diffusion based approaches to face swapping have been introduced to remedy these issues.

2.2. Diffusion Face-Swapping Approaches

Diffusion models offered the training stability and high fidelity that GANs lacked, showing promising improvements on image synthesis [9]. The first face-swapping framework utilizing a diffusion approach came in DiffFace [16]. The DiffFace framework proposes the training of an ID Conditional DDPM [13], the output of which requires multiple stages of guidance at inference and needs to be progressively blended onto the original target face. DiffSwap [43] posed face-swapping as an inpainting task where only the central portion of the face is considered as a candidate for swapping, eliminating the need for any blending at inference. REFace [1] leverages features from an image foundation model (CLIP [27]) in the face-swapping pipeline to provide the model with a semantic understanding of both the source and target face images. These methods lead to face-swaps which either contain artifacts or make only subtle modifications to the target faces and fail to transfer the identity of the source face effectively. These issues can be minimized by utilizing 3D features to disentangle face identity from pose and expression.

2.3. 3D Awareness in Diffusion Models and Face-Swapping Approaches

Previous works have integrated 3D awareness in 2D diffusion processes [36, 34, 5]. MVDiffusion [36] proposes a method for multi-view image generation by integrating camera pose conditioning and enforcing geometry constraints across latent representations. The diffusion process is further guided by cross-view attention, ensuring that image outputs remain consistent with underlying 3D scene structure. 3DFuse [34] conditions a pretrained 2D diffusion model using rendered depth maps from a coarse point cloud generated with a text prompt. This 3D structure is injected during sampling, allowing the 2D images to better reflect realistic shapes and poses. GeNVS [5] utilizes a learned latent 3D feature volume derived from a single source image. This latent space can then be rendered into novel views and the result is used to condition a 2D diffusion model, allowing for 3D-aware novel view synthesis with 2D outputs.

Face-swapping methods have also attempted to implement 3D features. Some GAN based models utilize 3D features in their face-swapping pipelines. HiFiFace [39] generates 3D reconstructions of both the source and target faces, utilizing the shape information provided by these reconstructions for guidance during training. 3dSwap [21]

projects 2D faces into the 3D latent space where face-swapping is performed before the 3D swapped face is then rendered back to 2D. However, this method requires camera position information at inference to perform face swapping so the results of this model are notably absent from our comparison. The only diffusion-based approach to the face-swapping task that is 3D aware is DiffSwap [43]. However, in DiffSwap the use of 3D features is rather limited as 3D face reconstruction is only used to generate landmarks that are provided as input to the UNet at inference. Our proposed method brings 3D awareness to the task of diffusion-based face-swapping in a much more substantial way as we utilize informative 3D latent features that are incorporated as conditioning features during the training stage. DiffSwap++ does not require 3D features at inference.

3. Proposed Method: DiffSwap++

3.1. Preliminaries on Diffusion Models

To mitigate the computational expense of pixel-space generation, DiffSwap++ operates entirely within the semantic latent space of a pretrained Latent Diffusion Model (LDM) [29]. Furthermore, to improve training and sampling efficiency over standard Denoising Diffusion Probabilistic Models (DDPMs) [13], we build upon Denoising Diffusion Implicit Models (DDIMs) [35], which abandon the Markov assumption in favor of a deterministic update to predict the fully denoised representation $\hat{z}_0$ at each step.

3.2. Face-Swapping Task

To formulate the face-swapping task, given a source image x^s and a target image x^t, the goal is to generate a new image in which the facial identity matches that of the source image x^s, while preserving the pose, expression, and background of the target image x^t. In our proposed DiffSwap++, we train a diffusion model for conditional inpainting, where the region consisting of the face of the target image is masked, and the face is reconstructed using conditioning features in the diffusion process. These conditioning features include the identity features of x^s and characteristic features of x^t. Along parallel lines of research, ReFace [1] introduces CLIP [27] features of both x^s and x^t as conditioning. However, these features fail to capture the fine-grained expressions of the source face x^s and the general pose of the target face x^t and hence result in swaps which lack photorealism. To mitigate this, we introduce 3D latents into the diffusion model to guide the generative process in maintaining the pose and expression of the target face x^t while preserving the identity of the source face x^s.

3.3. Conditioning for Diffusion Guidance

For effective face-swapping, it is essential for the generated image to preserve (i) the identity of the source image x^s and (ii) pose and expression of the target image x^t. To obtain the identity features f_{ID} of x^s, latent representations of a pretrained Face Recognition model F_{ID} (e.g. ArcFace [7], CosFace [38]) are projected using an MLP P_{ID}:

$$f_{\text{ID}} = P_{\text{ID}}(F_{\text{ID}}(x^s)). \tag{1}$$

To extract pose and expression features, the facial landmarks obtained from a landmark extractor F_{LM} are projected using another MLP P_{LM} to obtain the landmark feature representation of the target face x^t:

$$f_{LM} = P_{LM}(F_{LM}(x^t)). \tag{2}$$

While these landmark features capture 2D information about the pose and expression of the target face, they are often insufficient in challenging scenarios where the pose or expression deviates significantly from that of the source image. In such cases, diffusion models struggle to generate realistic and coherent face swaps. To address these challenges, DiffSwap++ incorporates 3D latent representations of both the source and target images to more effectively guide the diffusion process, with the extraction of these 3D features depicted in Fig. 2). Our overall conditioning process can be viewed in Fig. 3.

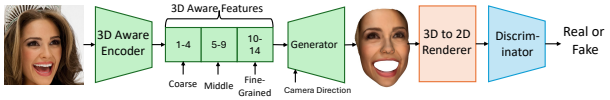


Figure 2. The 3D features are extracted from an encoder trained for 3D face reconstruction and are projector into a compatible feature dimension to be used as a conditioning feature in DiffSwap++.

3D Conditioning Latent. In DiffSwap++, we incorporate 3D latent features as conditioning vectors for cross-attention with the Diffusion UNet. These 3D latents are generated by leveraging the encoder, $\mathcal{E}$, from a pretrained GAN model. During training the encoder is used to project an input image x into latent space $\mathcal{W}$ to get a 3D aware high-dimension intermediate latent vector w_x. Then the pretrained EG3D[3] generator $\mathcal{G}$ is used to generate a triplane 3D representation of x before rendering this representation back into a final 2D reconstruction $x' = \mathcal{G}(w_x, d)$ where d is the input camera direction of x estimated by Deep3d Face Reconstruction[8]. Additionally, the EG3D generator is used to generate another view of the input face $\hat{x} = \mathcal{G}(w_x, \hat{d})$ where $\hat{d}$ is a random camera direction. This new view of the input face is then fed into the encoder to get $w_{\hat{x}}$ and the reconstruction process is repeated once again to produce $\hat{x}' = \mathcal{G}(w_{\hat{x}}, d)$. The overall optimization of this model can be written as: $\min_\theta\{\mathcal{L}(x, x')+\eta\mathcal{L}(x, \hat{x}')+\mathcal{L}(w_x, w_{\hat{x}})\}$. The latent representations generated by this encoder are of a standard dimension for 3D face representations consistent with that of EG3D[3].

The latent space is hierarchically organized by feature frequency, inherited from the underlying 3D generator

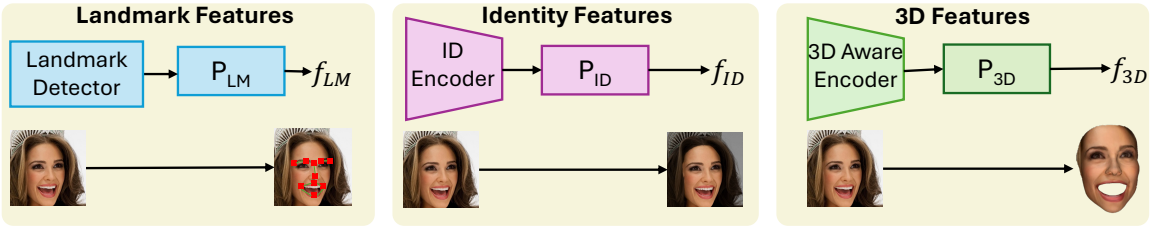


Figure 3. Overview of the condition generation pipeline for **DiffSwap++**. We utilize the standard landmark and identity features alongside our 3D features, novel to diffusion face-swapping.

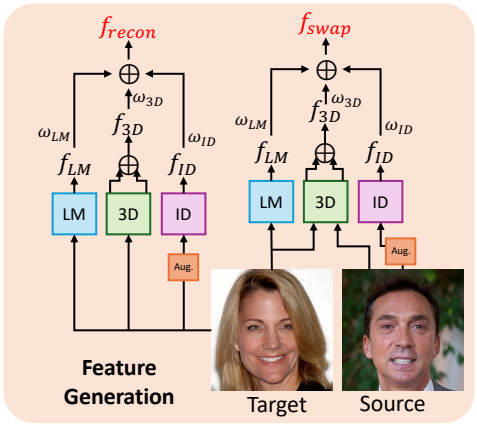


Figure 4. Our proposed **feature generation pipeline**. The obtained features are fed as conditional input to our Diffswap++ diffusion model.

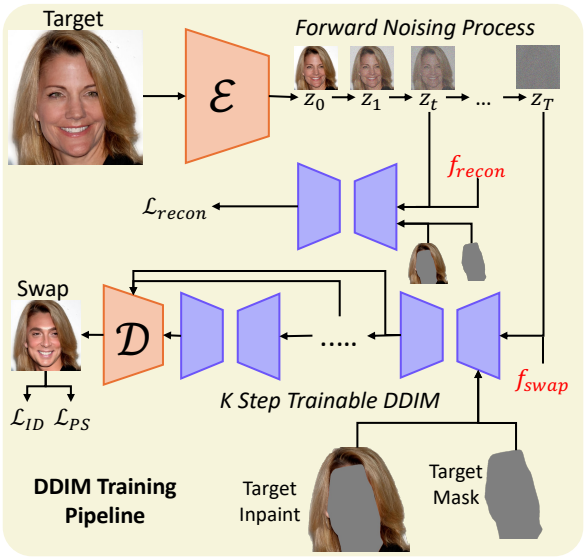


Figure 5. Training pipeline for **DiffSwap++**. This figure highlights our primary DDIM training pipeline where we perform both reconstruction and face-swapping.

structure. We categorize these into three levels of granularity to analyze their specific impacts:

- **Coarse-grained (Layers 1-4):** These layers control low-frequency structural geometry, including head shape, volumetric features, and pose orientation. They provide the fundamental geometric "canvas" of the source identity.
- **Middle-grained (Layers 5-9):** These layers govern localized facial features and semantic attributes, such as eye shape, mouth structure, and specific expression nuances.
- **Fine-grained (Layers 10-14):** These layers represent high-frequency details, including skin texture, lighting conditions, and color schemes.

The final 3D latent f_{3D} captures the holistic structural identity by aggregating these complementary frequencies. Thus,

$$f^t_{\text{3D}} = P^t_{\text{3D}}(\frac{1}{\text{L}}\sum_{i=1}^{\text{L}} F_{\text{3D}}(x^t)), \tag{3}$$

$$f^s_{\text{3D}} = P^s_{\text{3D}}(\frac{1}{\text{L}}\sum_{i=1}^{\text{L}} F_{\text{3D}}(x^s)), \tag{4}$$

$$f_{\text{3D}} = f^t_{\text{3D}} + f^s_{\text{3D}} \tag{5}$$

Here, P^t_{3D} and P^s_{3D} are MLPs to project the 3D latents of the source and target images and L is the number of layers of the latent features. Finally, all latent features are then combined to create the final conditioning feature f as follows

$$f = w_{\text{3D}} f_{\text{3D}} + w_{\text{ID}} f_{\text{ID}} + w_{LM} f_{LM} \tag{6}$$

where w_{3D}, w_{ID} and w_{LM} are weights for aggregation as illustrated in Figure 4.

3.4. Training Objective of DiffSwap++

Our training objective fuses three complementary losses as illustrated in Fig 5, viz., denoising diffusion, identity preservation, and perceptual similarity, optimized in two sequential steps. In the *face-reconstruction step*, a U-Net denoiser ϵ_θ learns to infer the Gaussian noise ϵ injected at a random diffusion timestep t. Given the noisy sample z_t, its timestep embedding x^t, a spatial mask M^{tar}, and our fully conditioned 3D aware feature token f_{recon} distilled from the target image, the network is supervised with the mean–squared error:

$$\mathcal{L}_{\text{Recon}} = E_{t,x_0,\epsilon}\left\|\epsilon_\theta(z_t, M^{tar}, f_{recon}, t) - \epsilon\right\|_2^2 \tag{7}$$

The *face–swapping step* fine-tunes a DDIM sampler, so that its intermediate latents $\hat{z}_i (i = 1 \dots N)$ transmit the source identity while blending seamlessly with the target context. To calculate our losses the intermediate latents are processed through a decoder $\mathcal{D}$ and, for $\mathcal{L}_{\text{ID}}$, an ArcFace [7] identity encoder $\mathcal{E}$:

$$\mathcal{L}_{\text{ID}} = \sum_{i=1}^{\text{N}} \left(1 - \left\langle \mathcal{E}(\mathcal{D}(\hat{z}_{0(z_i,t_i)}) \otimes M^t), \mathcal{E}(x^s \otimes M^s) \right\rangle\right) \tag{8}$$

where M^s and M^t are source and target face masks respectively and $\langle \cdot, \cdot \rangle$ denotes cosine similarity. Visual coherence with the target background is promoted via a learned perceptual image patch similarity (LPIPS) metric[42],

$$\mathcal{L}_{\text{PS}} = \sum_{i=1}^{\text{N}} L_{\text{LPIPS}}(\mathcal{D}(\hat{z}_{0(z_i,t_i)}), x^t) \tag{9}$$

The total loss $\mathcal{L}_{Total}$ is thus a‘ combination of reconstruction, identity and perceptual similarity losses denoted as

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{Recon}} + w_{\text{LID}}\mathcal{L}_{\text{ID}} + w_{\text{PS}}\mathcal{L}_{\text{PS}} \tag{10}$$

where w_{ID} and w_{PS} trade off identity preservation against perceptual realism. This two-step optimization first instills a strong generative prior through accurate face reconstruction, then aligns the model with the face-swapping objectives without sacrificing realism or identity fidelity.

4. Experiments

Datasets. DiffSwap++ is trained on the CelebA-HQ [14] dataset where the images are of high quality version of the CelebA [19] dataset. The dataset comprises of 30k images, of resolution 1024×1024, of which 28k are used for training and the remaining for evaluation. In addition, for out-of-domain (OOD) data, we use the FFHQ dataset [15] for evaluation. FFHQ contains 70k high quality face images and we utilize the last 2k images for our evaluation. For biometric evaluations for source identity consistency in swapped faces, we use the CelebV-Text [41] dataset. We sample 50 frames from each of 20 videos from the dataset to perform face-swapping.

Implementation Details. In all our experiments, the images are resized to 512×512. The denoising network (ϵ_θ) is a U-Net architecture [30]. For the 2D conditioning features, ID features are extracted using ArcFace [7]. We obtain 68 face landmark positions from DLib [17]. The 3D conditioning feature latents are generated by leveraging an encoder with a feature pyramid architecture from pSp [28] that is pre-trained to generate 3D latent representations of 2D faces in 3DSwap [21]. These conditioning features are integrated into the diffusion process using cross attention [22] with each U-Net layer.

Our model is trained on 4 NVIDIA A6000 GPUs (48GB). We train our model for 20 epochs with a global batch size of 4 and initialize our training for stable diffusion [29]. During training, we perform $T = 1000$ noising time-steps and $N = 4$ DDIM [35] steps. During inference, 50 DDIM [35] steps are used. For the weighted sum in equation 6 to calculate f, we utilize the hyperparameters $w_{\text{3D}} = 1.0$, $w_{\text{ID}} = 10.0$, and $w_{\text{LM}} = 0.05$. For the weighted sum in equation 10 to calculate $\mathcal{L}_{Total}$, we utilize the hyperparameters $w_{\text{ID}} = 0.3$ and $w_{\text{PS}} = 0.1$.

Evaluation Protocol. We benchmark the face-swapping performance of DiffSwap++ on the CelebA-HQ [14] and FFHQ [15] test sets. Specifically, each method is applied to an identically constructed set of 1000 source faces and 1000 target faces, yielding 1000 source–target pairs and ensuring strictly comparable conditions across all baselines. To assess the potential security risks, we report Fréchet Information Distance (FID) [12] to indicate the difficulty of detecting generated samples compared to real data. We measure impersonation potential using top-1 and top-5 retrieval accuracy, which proxies the success rate of bypassing identity verification systems. For the identity metric, ArcFace [7] is used to embed every source and swapped face, after which cosine similarity determines the nearest neighbors among the source set. Pose and expression measured as the ℓ_2 distance between head-pose estimates from HopeNet [10] and expression coefficients from DeepFace3DRecon [8] are evaluated to ensure the contextual consistency necessary to deceive human observers in dynamic environments. All baseline methods are evaluated using their officially released code and publicly available pretrained checkpoints. Note that the results for ReFace [1] are obtained from a model we trained for 20 epochs using the exact hyperparameter settings reported in the original paper, to ensure a fair comparison.

Table 1. Inference time and memory usage per image.

Method	Type	Time (s)	VRAM (GB)
DiffFace	Diffusion	37.77	8.15
DiffSwap	Diffusion	5.20	8.07
FaceDancer	GAN	0.80	1.65
HiFiFace	GAN	0.04	2.80
REFace	Diffusion	3.44	18.68
DiffSwap++	**Diffusion**	**3.35**	**18.69**

Inference Efficiency. Table 1 compares the computational efficiency of our method against state-of-the-art baselines. In terms of inference speed, DiffSwap++ requires 3.35s per image at inference, substantially faster than diffusion-based baselines: $\sim 11\times$ faster than DiffFace and $\sim 1.5\times$ faster than DiffSwap, while remaining on par with REFace. In terms of memory, DiffSwap++ uses 18.69 GB VRAM. Although higher than lightweight methods such as FaceDancer [31], it is comparable to high-fidelity frameworks and fits within the capacity of modern GPUs.

Table 2. Comparison on CelebA [19] dataset.

Method	FID↓	ID retrieval ↑ Top-1	ID retrieval ↑ Top-5	Pose↓	Expr.↓
HiFiFace [39]	13.97	86.4%	93.9%	3.01	1.19
FaceDancer [31]	12.25	79.0%	92.9%	2.32	**0.65**
DiffFace [16]	9.49	93.0%	97.2%	3.19	1.01
DiffSwap [43]	8.29	16.9%	34.9%	**2.13**	0.76
SimSwap [6]	10.61	90.3%	93.3%	2.21	0.78
REFace [1]	7.01	92.7%	96.9%	3.30	1.01
DiffSwap++ (w/o 3D)	8.80	86.0%	94.1%	3.48	1.06
DiffSwap++ (NeRF Feat.)	9.64	92.0%	96.9%	4.16	1.12
DiffSwap++ (e4e Enc.)	7.24	91.4%	97.0%	3.94	1.05
DiffSwap++	**6.42**	**94.7%**	**98.2%**	3.29	1.00

Table 3. Comparison on FFHQ [15] dataset.

Method	FID↓	ID retrieval ↑ Top-1	ID retrieval ↑ Top-5	Pose↓	Expr.↓
HiFiFace [39]	11.85	74.0%	86.3%	3.20	1.39
FaceDancer [31]	9.45	75.9%	87.6%	2.43	**0.70**
DiffFace [16]	8.72	86.9%	94.0%	3.67	1.30
DiffSwap [43]	**6.21**	34.9%	55.7%	2.43	1.03
SimSwap [6]	9.48	77.8%	81.7%	**2.39**	0.92
REFace [1]	6.45	91.0%	95.5%	3.69	1.07
DiffSwap++ (w/o 3D)	9.95	84.2%	92.4%	3.56	1.09
DiffSwap++ (NeRF Feat.)	10.75	93.4%	98.1%	5.32	1.16
DiffSwap++ (e4e Enc.)	5.74	94.1%	97.6%	4.07	1.09
DiffSwap++	6.57	**95.1%**	**98.4%**	3.70	1.07

4.1. Comparison with the State-of-the-Art

Face-Swapping on CelebA-HQ and FFHQ. In Table 2, we compare DiffSwap++ against state-of-the-art face-swapping methods, as well as a variant of DiffSwap++ trained without 3D features, using 1,000 source–target pairs from CelebA-HQ [14]. In Table 3, we further evaluate DiffSwap++ under out-of-domain (OOD) face swapping on the challenging FFHQ [15] dataset. Across both datasets, DiffSwap++ achieves the highest top-1/top-5 identity retrieval accuracy and the lowest Fréchet Information Distance (FID), demonstrating that it produces swaps that are both photorealistic and highly effective at inducing false acceptances in recognition systems. Importantly, pose and expression similarity remain competitive with the strongest baselines, indicating that improved identity fidelity is not obtained at the expense of geometric or affective consistency.

We note that identity retrieval is inherently at odds with pose and expression preservation: methods that minimally modify the target face typically achieve strong pose/expression scores but fail to transfer the source identity. Consequently, approaches reporting low identity retrieval often exhibit favorable pose and expression metrics, as they avoid the structural changes required to preserve the source identity in the swapped result. This trade-off is reflected in our results: DiffSwap [43], SimSwap [6], and FaceDancer [31] largely preserve the target pose and expression, yielding strong scores on those metrics, but their limited identity transfer leads to higher FID and substantially lower identity-matching performance. In contrast, DiffSwap++ strikes a more balanced compromise, achieving state-of-the-art identity preservation while maintaining the target's pose and expression with minimal degradation.

In Table 2 and Table 3, we further observe that our 3D-aware latent features are more effective for training than alternative 3D feature extraction strategies. Specifically, we evaluate (i) replacing our 3D features with NeRF features extracted using a pi-GAN [4] model pretrained on CelebA [19], and (ii) substituting the pSp [28] encoder in our 3D pipeline with an e4e [37] encoder. Although both variants converge faster (8 epochs for e4e and 10 epochs for NeRF) and improve ID retrieval, further supporting the robustness of our 3D guidance, their overall performance remains inferior to DiffSwap++.

Biometric Evaluation for Deepfake Detection. To gauge how faithfully a face swapping method propagates source identity under a biometric lens, it is natural to analyze the Impostor Attack Presentation Acceptance Rate (IAPAR) metric across different acceptance thresholds [25, 24]. Biometric security systems use IAPAR to estimate their effectiveness in identifying fraudulent faces instead of incorrectly accepting them as genuine users. We utilize IAPAR to strictly evaluate the impersonation success of the swap acting as a presentation attack, distinguishing this metric from standard detection metrics like AUC.

For each model, we construct a face-swap corpus on CelebV-Text [41] by randomly selecting 20 videos, sampling 50 frames per video as targets. We replace each target face with a single source frame drawn from a different video, yielding 1000 swapped frames per method. FaceNet-512 [32] embeddings are extracted from every swapped frame and its corresponding source image, and cosine similarity scores are thresholded over the full $[0, 1]$ range; IAPAR is then the fraction of swapped frames whose similarity exceeds the threshold, i.e., the rate at which an imposter presentation (the swap) is incorrectly accepted as the genuine source. We compute the False Reject Rate (FRR) in an analogous way, but using 50 original frames from the same video as the source image. In Fig. 6, we plot the percentage of ground truth frames that fail to match the identity of the corresponding source image at thresholds ranging from $0-1$. DiffSwap++ maintains a high threshold demonstrating high-fidelity face-swaps that preserve the source face identity. SimSwap [6] demonstrates a higher threshold than DiffSwap++. However, qualitative inspection in Fig. 7 shows that SimSwap generates visually implausible swaps failing to properly integrate the source identity onto the target features, leading to swaps which are similar in identity to the source. On the other hand, DiffSwap++ generates swaps that maintain realism.

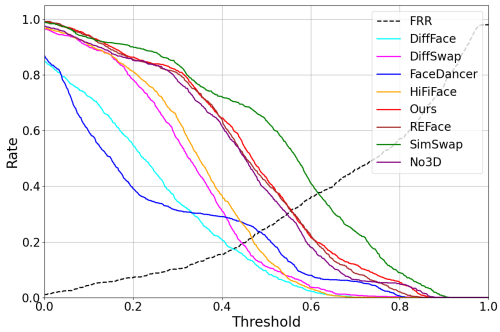


Figure 6. A comparison of IAPAR for all models at different thresholds against FRR.

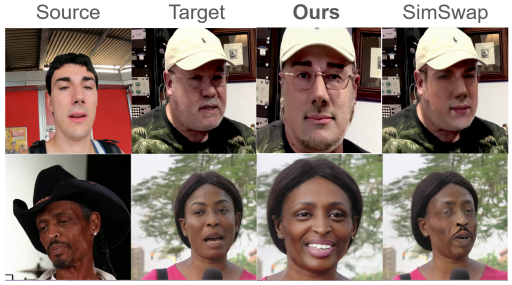


Figure 7. Comparison between **DiffSwap++** and SimSwap on CelebV-Text video frames.

4.2. Ablation Analysis

To understand how the 3D conditioning latent should be incorporated into the face-swap network, we experiment with five integration schemes for the 14×512 feature tensor produced by the 3D Feature Extractor [3]. The first strategy performs a global average over the layer axis, yielding a single 512-D descriptor (‘*Avg. Pool*’). The next three strategies pool only over specific layer groups that roughly correspond to ‘*Coarse*’ (layers 1–4), ‘*Middle*’ (5–9), and ‘*Fine-Grained*’ (10–14) geometry and appearance, respectively. Finally, we replace pooling with a learnable linear projection that selects a weighted combination of all 14 layers. Each variant is trained for five epochs under identical hyper-parameters, and its quantitative performance on FFHQ[15] is reported in Table 4. This table reveals the specific influence mechanisms of different granularity levels.

Impact on Identity: The “Coarse” variant yields the lowest identity retrieval (44.30%), indicating that while head shape is necessary, it is insufficient for robust recognition without textural details. The “Fine-Grained” variant performs better (64.40%) by capturing skin texture, yet fails to achieve state-of-the-art results without the structural support of the coarse layers. The superior performance of “Avg. Pool” (79.30%) confirms that identity is a composite of both structural geometry and textural appearance.

Impact on Pose Consistency: Notably, the “Coarse” variant results in higher pose error (l_2 distance of 3.88) compared to “Fine-Grained” (3.56). We attribute this to the entanglement of source pose within the coarse layers; relying exclusively on coarse source latents creates a geometric conflict with the target’s pose mask. By utilizing the full hierarchy, “Avg. Pool” balances structural identity transfer while mitigating these geometric conflicts, as evidenced by its robust performance across all metrics.

We additionally provide results for variants of our model, each trained for five epochs, in Table 4 where it is evident that 4 DDIM steps provide the best balance during training and that both identity and landmark features are required for training a face swapping model of this nature.

Table 4. Ablation studies on 3D Condition Generation and DDIM Steps/Loss Components on FFHQ dataset.

Method	**FID↓**	**ID retrieval ↑**		**Pose↓**	**Expr.↓**
		Top-1	**Top-5**		
3D Condition Generation					
Avg. Pool	5.72	**78.80%**	**90.20%**	4.40	**1.04**
Coarse	6.59	46.90%	65.90%	4.11	**1.04**
Middle	7.21	54.20%	73.80%	4.12	1.09
Fine-Grained	**5.68**	65.60%	79.80%	**3.85**	**1.04**
Linear Proj.	6.27	50.90%	69.80%	3.88	1.05
DDIM Steps and Loss Components					
DiffSwap++	**5.72**	**78.80%**	**90.20%**	**4.40**	**1.04**
DiffSwap++ (w/o ID Feat.)	15.84	0.10%	0.70%	5.88	1.15
DiffSwap++ (w/o LM Feat.)	42.19	0.30%	0.60%	20.84	2.54
DiffSwap++ (2 DDIM Steps)	9.28	32.30%	50.30%	5.58	1.27
DiffSwap++ (6 DDIM Steps)	27.7	11.40%	27.30%	4.91	1.40

4.3. Qualitative Evaluation

Qualitative comparisons between all models are provided in Fig. 8. Three recurrent failure modes observed in prior works are highlighted in Fig. 9 and compared to the face-swaps produced by DiffSwap++. First, **visual artifacts** hallucinated eyeglass rims, hollow pupils, or irregular chin textures are prominent in the REFace [1] outputs. These irregularities serve as clear indicators for forgery detection algorithms. In contrast, DiffSwap++minimizes these cues, producing anatomically consistent geometry that significantly complicates the task of distinguishing synthetic content from authentic. Second, several methods synthesize **anatomically implausible faces** that break facial geometry; this is readily apparent in the face-swaps generated by HiFiFace [39], but is absent in the ones by DiffSwap++, whose structure adheres to plausible facial morphology. Third, approaches such as DiffSwap often **leave the target face intact**, effecting only superficial alterations; although this strategy yields high pose and expression scores, it fails to convey the source identity and therefore fares poorly on retrieval metrics. By contrast, DiffSwap++ achieves a more balanced transformation, preserving identity while respecting target pose and expression, and doing so without introducing artifacts or geometric distortions.

User Study for Face-Swap Evaluation. To evaluate the risk of deceiving human moderators, we conducted a



Figure 8. Qualitative comparison of all models on FFHQ.

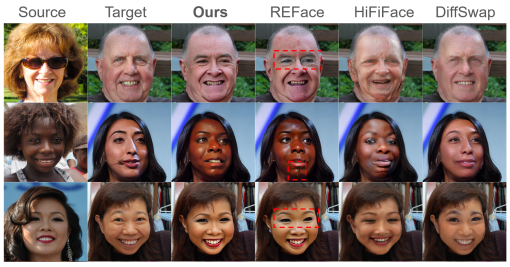


Figure 9. Comparison highlighting issues with other models (artifacts highlighted in red).

Table 5. User Study Results

Metric	Method	Mean	Var	95% CI
Identity	**Ours**	**3.46**	0.32	[3.06, 3.86]
	REFace	3.20	0.23	[2.86, 3.54]
	Hifi	2.58	0.26	[2.22, 2.94]
	DS	2.46	0.33	[2.05, 2.87]
Expression	Ours	3.14	0.32	[2.74, 3.54]
	REFace	3.22	**0.19**	[2.91, 3.53]
	Hifi	2.72	0.39	[2.27, 3.17]
	DS	**3.57**	0.42	[3.11, 4.03]
Pose	Ours	3.65	0.29	[3.26, 4.04]
	REFace	3.48	**0.04**	[3.33, 3.63]
	Hifi	3.08	0.29	[2.70, 3.46]
	DS	**4.11**	0.25	[3.75, 4.47]

user study involving 10 volunteer participants. As human inspection remains a primary layer of defense in media verification, assessing the perceptual convincingness of the output is crucial. Each participant was presented with 10 sets of face-swapped images along with their corresponding source and target images. Each set included one face swap from DiffSwap++, REFace [1], HifiFace [39], and DiffSwap [43]. Participants were asked to rate each face swap on three criteria: identity preservation, expression preservation, and pose preservation using a 5-point Likert scale, where 1 indicates the lowest quality and 5 indicates the highest. To eliminate bias, the order of face swaps within each set was randomized, and the names of the models were not disclosed. The quantitative results, summarized in Table 5, demonstrate that DiffSwap++ achieves the highest mean perceptual quality in identity preservation (3.46). While the performance gap with the strong baseline REFace (3.20) is not statistically significant ($p = 0.26$), indicating comparable reliability between the two methods, the 95% confidence interval for DiffSwap++ ($[3.06, 3.86]$) extends higher than that of REFace ($[2.86, 3.54]$), suggesting a higher upper bound for identity preservation. In contrast, our method significantly outperforms HifiFace ($p = 0.006$) and DiffSwap ($p = 0.014$), the latter of which sacrifices identity retention (2.46) for expression. Regarding stability, DiffSwap++ maintains a balanced variance profile ($\sigma^2 \approx 0.32$) across all metrics. This stands in contrast to DiffSwap, which exhibits higher instability in expression ($\sigma^2 = 0.42$), and REFace, which, while highly stable ($\sigma^2 = 0.04$ in pose), fails to reach the high-fidelity peaks achieved by our method in both Identity and Pose. Thus, DiffSwap++ offers the most effective balance of high mean fidelity and consistent generation quality.

5. Conclusion

In this work, we propose DiffSwap++, a diffusion-based face-swapping pipeline that incorporates 3D facial features during training to effectively disentangle identity from pose and expression, enabling more accurate identity preservation in face-swapped images. Our approach is efficient at inference and achieves state-of-the-art performance in identity preservation, while maintaining comparable levels of pose and expression fidelity. We validate the effectiveness of our method through extensive experiments on three public datasets, including a biometric-style evaluation designed to rigorously assess identity preservation capabilities.

Acknowledgments

This material is based upon work supported by the Center for Identification Technology Research and the National Science Foundation under Grant No. 1650503 and 2413228. We would also like to thank Kumara Kahatapitiya at Stony Brook University for our valuable discussions.

References

[1] S. Baliah, Q. Lin, S. Liao, X. Liang, and M. H. Khan. Realistic and efficient face swapping: A unified approach with diffusion models, 2024.

[2] F. Bao, C. Li, J. Sun, and J. Zhu. Why are conditional generative models better than unconditional ones?, 2022.

[3] E. R. Chan, C. Z. Lin, M. A. Chan, K. Nagano, B. Pan, S. D. Mello, O. Gallo, L. Guibas, J. Tremblay, S. Khamis, T. Karras, and G. Wetzstein. Efficient geometry-aware 3d generative adversarial networks, 2022.

[4] E. R. Chan, M. Monteiro, P. Kellnhofer, J. Wu, and G. Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis, 2021.

[5] E. R. Chan, K. Nagano, M. A. Chan, A. W. Bergman, J. J. Park, A. Levy, M. Aittala, S. D. Mello, T. Karras, and G. Wetzstein. Generative novel view synthesis with 3d-aware diffusion models, 2023.

[6] R. Chen, X. Chen, B. Ni, and Y. Ge. Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 2003–2011. ACM, Oct. 2020.

[7] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):5962–5979, Oct. 2022.

[8] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set, 2020.

[9] P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis, 2021.

[10] B. Doosti, S. Naha, M. Mirbagheri, and D. Crandall. Hopenet: A graph-based model for hand-object pose estimation, 2020.

[11] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks, 2014.

[12] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018.

[13] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[14] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

[15] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks, 2019.

[16] K. Kim, Y. Kim, S. Cho, J. Seo, J. Nam, K. Lee, S. Kim, and K. Lee. Diffface: Diffusion-based face swapping with facial guidance. *Pattern Recognition*, 163:111451, 2025.

[17] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(60):1755–1758, 2009.

[18] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 2144–2151. IEEE, 2011.

[19] C.-H. Lee, Z. Liu, L. Wu, and P. Luo. Maskgan: Towards diverse and interactive facial image manipulation, 2020.

[20] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping, 2020.

[21] Y. Li, C. Ma, Y. Yan, W. Zhu, and X. Yang. 3d-aware face swapping. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12705–12714, 2023.

[22] H. Lin, X. Cheng, X. Wu, and D. Shen. Cat: Cross attention in vision transformer. In *2022 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE, 2022.

[23] D. S. Ma, J. Correll, and B. Wittenbrink. The chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, 47(4):1122–1135, 2015.

[24] M. Micheletto, R. Casula, G. Orrù, S. Carta, S. Concas, S. M. L. Cava, J. Fierrez, and G. L. Marcialis. Livdet2023 – fingerprint liveness detection competition: Advancing generalization, 2023.

[25] M. Micheletto, G. Orrù, L. Ghiani, and G. L. Marcialis. Improving fingerprint presentation attack detection by an approach integrated into the personal verification stage, 2025.

[26] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks, 2018.

[27] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[28] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation, 2021.

[29] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. 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.

[30] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[31] F. Rosberg, E. E. Aksoy, F. Alonso-Fernandez, and C. 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.

[32] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[33] S. Sengupta, A. Kanazawa, C. D. Castillo, and D. W. Jacobs. Sfsnet: Learning shape, reflectance and illuminance of

faces 'in the wild'. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6296–6305. IEEE, 2018.

[34] J. Seo, W. Jang, M.-S. Kwak, H. Kim, J. Ko, J. Kim, J.-H. Kim, J. Lee, and S. Kim. Let 2d diffusion model know 3d-consistency for robust text-to-3d generation, 2024.

[35] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[36] S. Tang, F. Zhang, J. Chen, P. Wang, and Y. Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion, 2023.

[37] O. Tov, Y. Alaluf, Y. Nitzan, O. Patashnik, and D. Cohen-Or. Designing an encoder for stylegan image manipulation, 2021.

[38] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018.

[39] Y. Wang, X. Chen, J. Zhu, W. Chu, Y. Tai, C. Wang, J. Li, Y. Wu, F. Huang, and R. Ji. Hififace: 3d shape and semantic prior guided high fidelity face swapping, 2021.

[40] Y. Xu, B. Deng, J. Wang, Y. Jing, J. Pan, and S. He. High-resolution face swapping via latent semantics disentanglement, 2022.

[41] J. Yu, H. Zhu, L. Jiang, C. C. Loy, W. Cai, and W. Wu. Celebv-text: A large-scale facial text-video dataset, 2023.

[42] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018.

[43] Y. Zhao, W. Zuo, Y. Lin, W. Wu, and D. Chen. 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 (CVPR)*, pages 22919–22929, 2023.

DiffSwap++: 3D Latent-Controlled Diffusion for Identity-Preserving Face Swapping

Supplementary Material

Weston Bondurant, Arkaprava Sinha, Hieu Le, Srijan Das, Stephanie Schuckers
The University of North Carolina at Charlotte
wbondura@charlotte.edu

A. Overview

The Supplementary material is organized as follows:

- Section B: Additional Quantitative Results
- Section C: Additional Qualitative Results

B. Additional Quantitative Results

B.1. Ablation Studies

Additional results for variants of our model, each trained for five epochs, are provided in Table 6. It is further evident that 4 DDIM steps provide the best balance during training and that both identity and landmark features are required for training a face swapping model of this nature.

Table 6. Ablation studies on 3D Condition Generation and DDIM Steps/Loss Components on CelebA dataset.

Method	FID↓	ID retrieval ↑		Pose↓	Expr.↓
		Top-1	Top-5		
3D Condition Generation					
Avg. Pool	5.78	**79.30%**	**90.40%**	3.73	0.99
Coarse	6.46	44.30%	66.70%	3.88	**0.97**
Middle	7.24	54.40%	74.60%	3.87	1.05
Fine-Grained	**5.53**	64.40%	82.70%	**3.56**	0.99
Linear Proj.	6.20	55.00%	75.60%	3.71	1.01
DDIM Steps and Loss Components					
DiffSwap++	**5.78**	**79.30%**	**90.40%**	**3.73**	**0.99**
DiffSwap++ (w/o ID Feat.)	10.6	0.60%	0.20%	5.05	1.09
DiffSwap++ (w/o LM Feat.)	60.19	0.60%	1.20%	21.03	2.50
DiffSwap++ (2 DDIM Steps)	11.29	30.50%	53.10%	5.06	1.25
DiffSwap++ (6 DDIM Steps)	36.89	12.20%	33.50%	4.46	1.39

B.2. Facial Recognition Evaluation Beyond ArcFace

To further validate the robustness of our identity preservation, we expand our face recognition evaluation by incorporating the CosFace [38] recognition model. As shown in Table 7, DiffSwap++ achieves a 96.80% Top-1 retrieval accuracy, providing additional evidence for the robustness of our identity preservation. Although DiffFace [16] and REFace [1] achieve marginally higher retrieval accuracies under this specific metric, their generated face swaps exhibit noticeably lower visual realism. This is corroborated by the FID results in the main manuscript, where DiffSwap++ achieves the lowest FID compared to REFace and DiffFace.

Table 7. Identity retrieval accuracy utilizing the CosFace recognition model.

Metric	Ours	DiffFace	REFace	DiffSwap	HifiFace	SimSwap	FaceDancer
Top-1 (%)	**96.80**	98.20	98.02	61.80	95.50	93.90	88.20
Top-5 (%)	**99.10**	99.40	99.67	79.81	97.80	94.80	96.30

B.3. Dataset Illumination Variance

Table 8 displays extracted spherical harmonics lighting coefficients using SfSNet [33]. We measured standard deviations across the X, Y, and Z axes, alongside overall directional intensity representing lighting contrast and harshness. We display that the primary datasets utilized in our methodology (CelebA [19] and FFHQ [15]) exhibit significantly greater spatial variances than CFD [23], a controlled face dataset where each image has the same flat lighting. These metrics empirically confirm that our chosen in-the-wild test datasets successfully capture real-world illumination conditions.

Table 8. Illumination variance across datasets measured via spherical harmonics.

Dataset	Lateral (X)	Vertical (Y)	Depth (Z)	Intensity (σ)
CFD	0.0596	0.0342	0.0711	0.0353
CelebA	0.1044	0.1119	0.0856	0.0973
FFHQ	0.1245	0.1254	0.0893	0.1100

C. Additional Qualitative Results

C.1. Extended Baseline Comparisons

Qualitative comparisons for images generated from the AFLW2000-3D dataset [18], which feature more extreme pose and expression, are provided in Figure 1. We also provide extended comparisons from the CelebA [19] dataset in Figure 2. These figures highlight the capability of DiffSwap++ to produce face swaps which are consistently more

realistic and better preserve the identity of the source face compared to other popular models.

Furthermore, Figure 3 provides a visually prominent comparison specifically between DiffSwap++ and SimSwap [6] on the CelebV-Text [41] dataset. As highlighted in the figure, SimSwap results often feature minimal changes to the target faces and introduce visual artifacts. This failure to properly integrate the source identity directly corresponds to their lower identity retrieval scores compared to DiffSwap++.

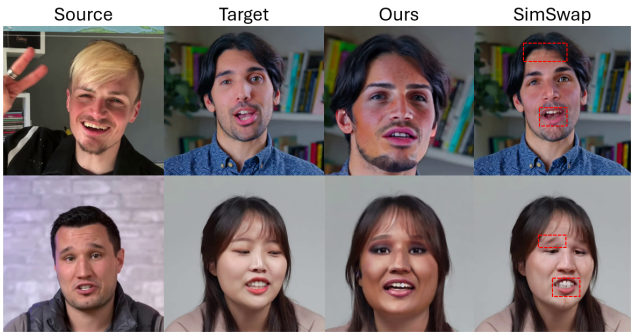


Figure 3. Targeted visual comparison highlighting artifact generation and lack of identity transfer in SimSwap compared to DiffSwap++.

C.2. User Study Examples

Figure 4 provides a representative visual sample of the interface and images used in our user study. To ensure a fair and consistent evaluation, we confirm that all images presented to participants were of identical quality, including the exact same resolution, aspect ratio, and display settings.

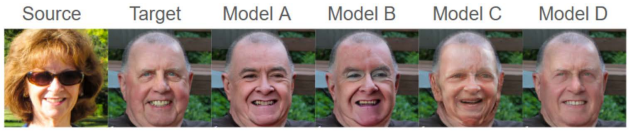


Figure 4. A representative example of the presentation format utilized during the human evaluation user study.

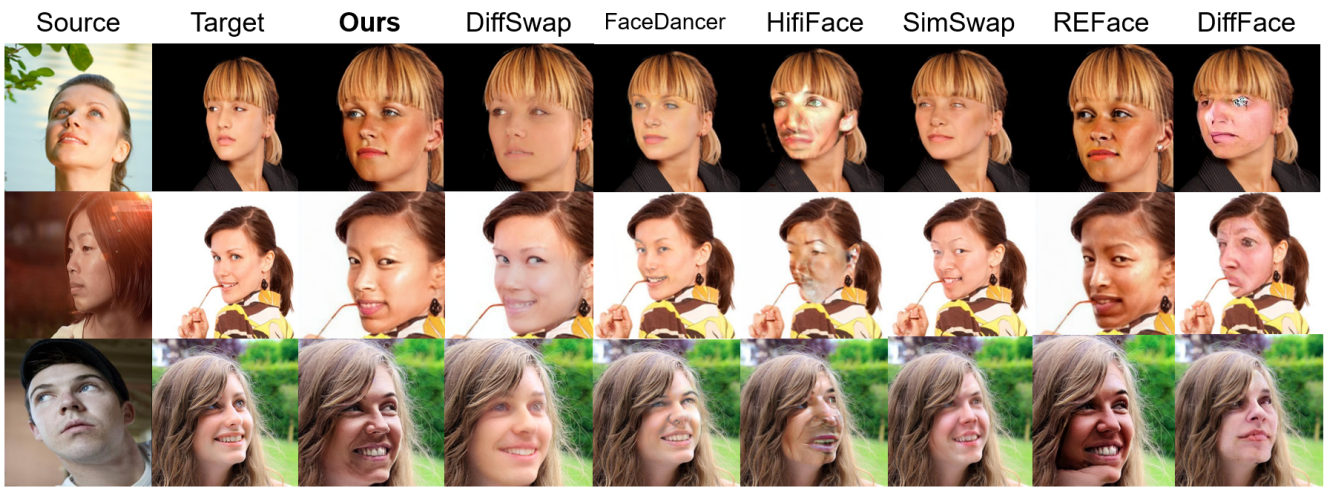


Figure 1. Qualitative comparison for extreme pose and expression.

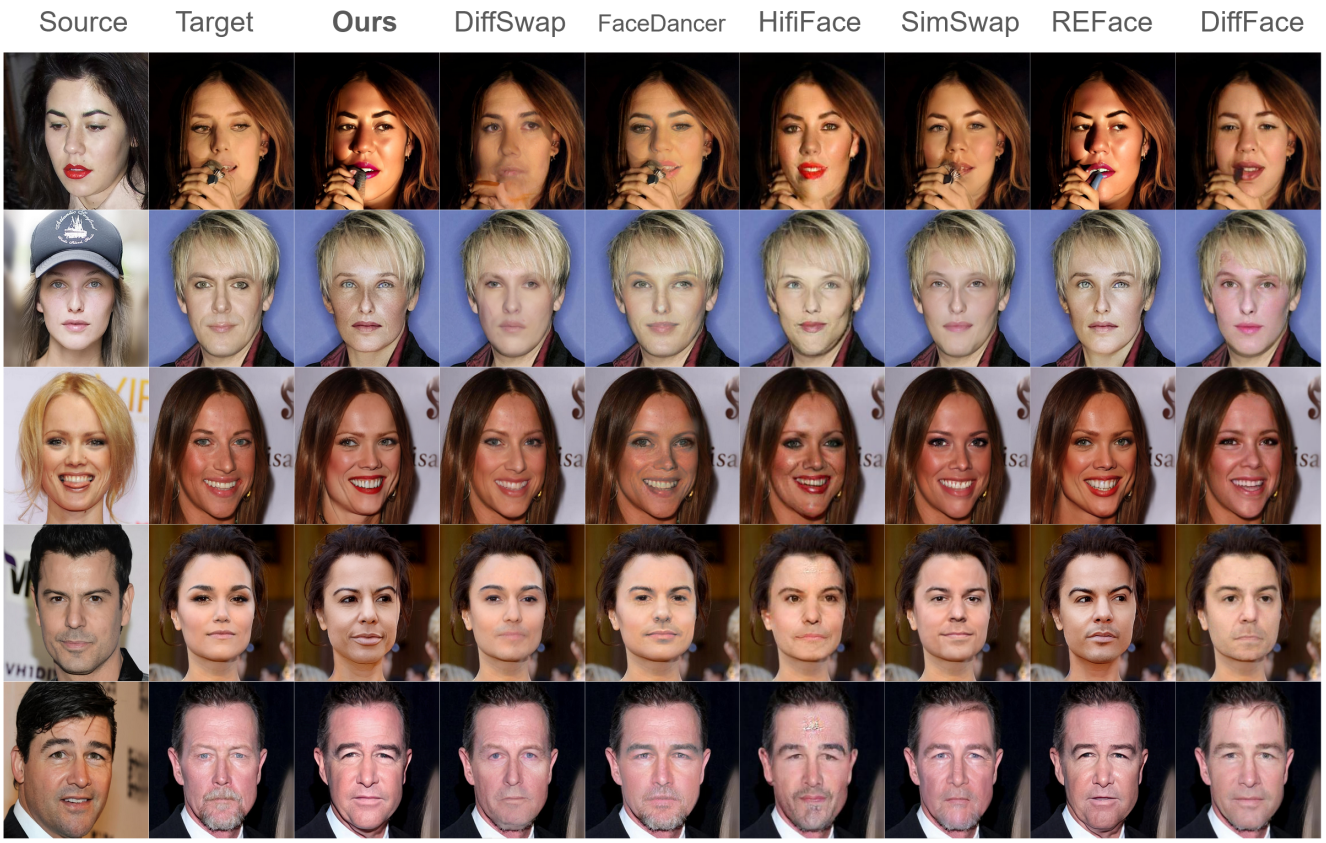


Figure 2. Qualitative comparison of all models on CelebA.