VividFace: A Robost and High-Fidelity Video Face Swapping Framework

Hao Shao[1] **Shulun Wang**[2] **Yang Zhou**[2] **Guanglu Song**[2]
Dailan He[1] **Zhuofan Zong**[1] **Shuo Qin**[2] **Yu Liu**[2] ✉ **Hongsheng Li**[1,3] ✉
[1]CUHK MMLab [2]SenseTime Research [3]CPII under InnoHK

Abstract

Video face swapping has seen increasing adoption in diverse applications, yet existing methods primarily trained on static images struggle to address temporal consistency and complex real-world scenarios. To overcome these limitations, we propose the first video face swapping framework, VividFace, a robust and high-fidelity diffusion-based framework. VividFace employs a novel hybrid training strategy that leverages abundant static image data alongside temporal video sequences, enabling it to effectively model temporal coherence and identity consistency in videos. Central to our approach is a carefully designed diffusion model integrated with a specialized VAE, capable of processing image-video hybrid data efficiently. To further enhance identity and pose disentanglement, we introduce and release the Attribute-Identity Disentanglement Triplet (AIDT) dataset, comprising a large-scale collection of triplets where each set contains three face images—two sharing the same pose and two sharing the same identity. Augmented comprehensively with occlusion scenarios, AIDT significantly boosts the robustness of VividFace against occlusions. Moreover, we incorporate advanced 3D reconstruction techniques as conditioning inputs to address significant pose variations effectively. Extensive experiments demonstrate that VividFace achieves state-of-the-art performance in identity preservation, temporal consistency, and visual realism, surpassing existing methods while requiring fewer inference steps. Our framework notably mitigates common challenges such as temporal flickering, identity loss, and sensitivity to occlusions and pose variations. The AIDT dataset, source code, and pre-trained weights will be released to support future research. The code and pretrained weights are available on the project page.

1 Introduction

In recent years, face swapping has emerged as a crucial technology across various domains, from content creation [36] and privacy protection [52] to safe stunt scene production [39] and digital twin generation [35]. As video is a predominant medium for communication, the demand for high-quality face swapping techniques has grown substantially. Video face swapping involves extracting identity features from a source face and seamlessly integrating them with the attributes (such as expressions, poses, *etc.*) and background of a target face while maintaining temporal consistency. However, despite the recent advancements, current face-swapping methods encounter difficulties in video contexts, as most are optimized for static images rather than dynamic video sequences.

Existing face swapping approaches can be broadly categorized into three main methodologies: 3D-based, GAN-based, and diffusion-based methods. Traditional 3D-based methods [3, 4, 47, 34], primarily utilizing 3D Morphable Models (3DMM) [5], often struggle with low-resolution outputs and face blending issues. GAN-based approaches [33, 9, 27, 2, 29, 56] encounter challenges with training instability, mode collapse, and producing low-resolution output, particularly in complex

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

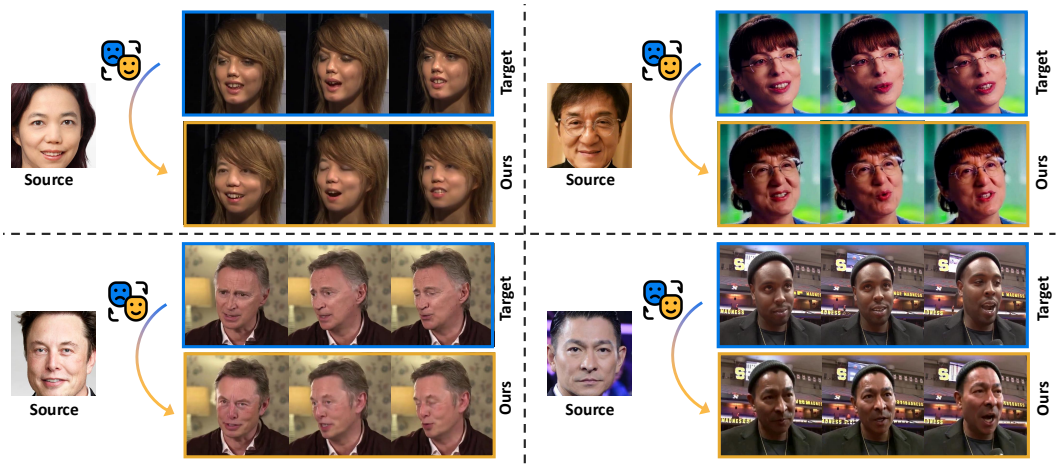


Figure 1: Face swapping results of VividFace at 512×512 resolution. Our method produces high-fidelity and vivid outputs that accurately follow both pose and expression changes. Corresponding videos are provided in the supplementary material.

cases. Recently, diffusion models [20] have gained prominence in image synthesis tasks, offering advantages such as high-fidelity output, enhanced controllability, and high training stability.

Recent advancements like DiffSwap [55] and REFace [1] have highlighted the effectiveness of diffusion models for image-level face swapping. However, significant challenges persist when extending these methods to video face swapping, including maintaining temporal consistency, handling large pose variations, and addressing occlusions. To overcome these challenges, we propose VividFace, the first robust and high-fidelity diffusion-based video face swapping framework. VividFace introduces an innovative hybrid training strategy that integrates diverse static image data with temporal video sequences, effectively overcoming the inherent limitations of using video-only data, such as insufficient diversity resulting from highly similar frames within individual videos. This hybrid approach significantly expands training data diversity, enhancing the robustness and generalization capability of our model. Our framework employs a specifically designed diffusion model optimized for processing both static images and temporal video data. We further introduce the VidFaceVAE adapted to jointly handle face images and video sequences, effectively mitigating temporal flickering typically encountered in existing video swapping methods.

To further enhance identity and attribute preservation, we create and release the Attribute-Identity Disentanglement Triplet (AIDT) dataset, consisting of 1 million image triplets and 0.6 million video triplets. Each triplet includes a source face, a target face with shared identity but different poses and expressions, and a GAN-generated face matching the target's pose and expression but featuring a distinct identity. This structured dataset significantly boosts the model's capability for disentangling identity from pose and expression. While ReliableSwap [53] also utilizes a triplet-based concept, it mainly addresses artifact reduction through augmentation rather than robust disentanglement. Additionally, we develop a comprehensive occlusion augmentation strategy that dynamically introduces various occluding objects over target faces, significantly improving the framework's robustness against real-world occlusions. To effectively manage significant pose variations, we incorporate a 3D Morphable Model (3DMM)-based reconstruction as additional conditioning input. This 3D guidance ensures accurate pose and expression representation, facilitating better generalization across diverse video contexts. To further reduce information leakage and enhance robustness, we retain only the pose and expression features from the reconstruction, discarding texture and identity information.

Experimental results demonstrate our framework's superiority in terms of Fréchet Video Distance (FVD), temporal consistency, and attribute/identity preservation, with fewer inference steps compared to existing methods. Besides, we also demonstrate the stability and generation of our method in multiple complex cases.

To summarize, this paper makes the following contributions:

- The first diffusion-based video face swapping framework, VividFace, featuring a novel image-video hybrid training strategy.

- We provide and plan to release a large-scale AIDT dataset to significantly improve face feature identity-expression disentanglement.
- Robustness enhancements through comprehensive occlusion augmentation and advanced 3D face reconstruction conditioning to handle large pose variations effectively.
- Extensive experimental analyses demonstrating superior temporal consistency, identity preservation, and visual quality, along with comprehensive ablation studies.

2 Related Work

2.1 Face Swapping

The frameworks of face swapping are generally categorized into three types: 3D-based [3, 4, 47, 34], GAN-based [33, 9, 27, 2, 29, 56, 26, 38], and diffusion-based methods [23, 55, 1, 19]. In addition to these three main approaches, FaceShifter [27] introduces a two-stage framework that generates high-fidelity swapped faces by thoroughly and adaptively exploiting and integrating the target attributes. 3D-based frameworks typically employ the parameterized 3DMM [5] model to reconstruct the swapped face. Face2Face [47] transferred expressions from source to target face by fitting a 3DMM face model to both faces. The authors in [34] show that face swapping with robust segmentation preserves identity in intra-subject swaps and reduces recognizability in inter-subject cases. HifiFace [51] introduced a semantic facial fusion module to improve photorealism. However, these 3D-based methods yield low similarity and unrealistic textures due to limited resolution.

GAN [16] has been a powerful tool for generating realistic synthetic images. The popular algorithm DeepFakes [11] utilizes an encoder-decoder architecture for identity-specific face swapping but lacks generalization. To improve adaptability, FSGAN [33] proposes a subject-agnostic approach with a recurrent reenactment module, inpainting and a blending module. E4S [29] reframes face swapping as fine-grained editing by disentangling shape and texture, using regional GAN inversion for precise feature manipulation and occlusion handling. SimSwap [9] introduces an ID Injection Module and Weak Feature Matching Loss for flexible, high-fidelity identity swapping. However, GAN-based methods often struggle with balancing losses and handling shape variations or occlusions, leading to inconsistencies in illumination and identity in complex cases.

Recently, diffusion models have become a leading framework for image & video generation. DiffFace [23] first leverages conditional diffusion models for stable identity-preserving swapping. DiffSwap [55] and FaceAdapter [19] build upon conditional inpainting paradigms to achieve high-fidelity, controllable swapping. REFace [1] improves this by reframing swapping as a self-supervised inpainting task. However, existing diffusion-based methods mainly target static images, overlooking key video challenges like temporal consistency, occlusions, and large pose variations.

2.2 Diffusion Models

Diffusion models [20, 31, 57, 58, 43] have recently emerged as a powerful generative framework, achieving state-of-the-art performance in various domains, including image synthesis [14, 20, 45], editing [22, 44, 28], super-resolution [50, 15], and video generation [18, 6, 32]. Unlike GANs, which often suffer training instability, diffusion models offer a more stable training process by gradually denoising data from random noise, resulting in high-fidelity outputs. Notable advancements include Stable Diffusion [40], which enhances efficiency by operating in the latent space, and SVD [6], which incorporates temporal modules to scale diffusion models for video tasks. Conditioning mechanisms like cross-attention and concatenation enhance controllability, enabling targeted generation across applications [48]. Thus, diffusion models are increasingly popular for versatile, high-quality content creation.

3 Method

3.1 Preliminaries

Our method employs Stable Diffusion (SD) [40] as the backbone network. Stable Diffusion is a text-to-image model built on the Latent Diffusion Model (LDM), which enables efficient image generation by operating within a compressed latent space. SD uses a variational autoencoder (VAE) [24] to map

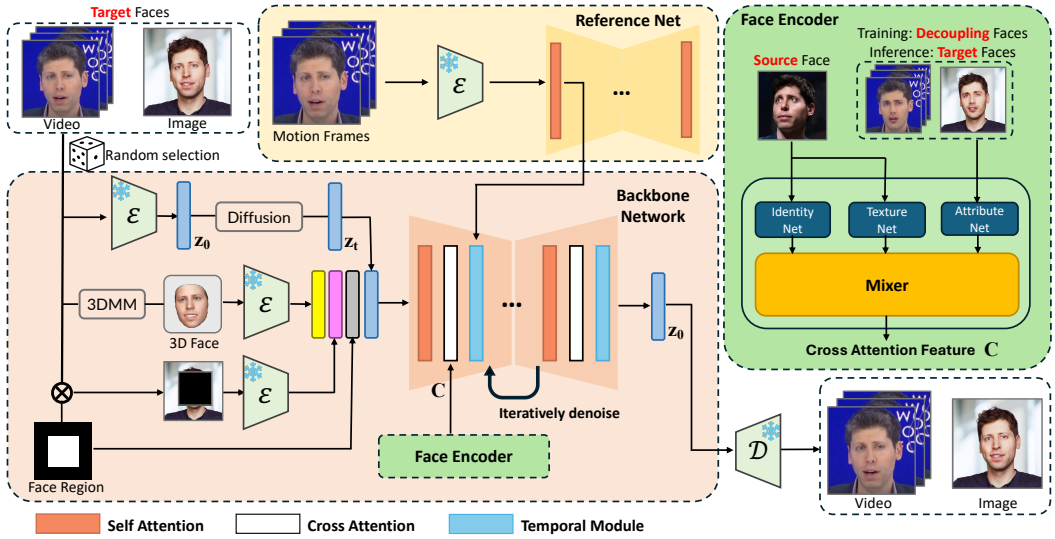


Figure 2: Overview of the proposed framework. During training, our framework randomly chooses static images or video sequences as the training data. In addition to the noise z_t, three other types of inputs are integrated to guide the generation process: (1) a face region mask, which controls the generation of facial imagery; (2) a 3D reconstructed face, which helps guide the pose and expression, especially in cases of large pose variations; and (3) masked source images, which supply background information. These inputs are processed through the Backbone Network, which performs the denoising operation. Within the Backbone Network, we employ cross-attention and temporal attention mechanisms. The temporal attention module ensures temporal continuity and consistency across frames. Our face encoder extracts identity and texture features from the target face, as well as pose and expression details from the source face, and uses these features to produce realistic and high-fidelity results.

the original image x_0 unto a latent representation z_0. reducing computational cost while preserving visual quality. The image is encoded as $z_0 = \mathcal{E}(x_0)$ and decoded back as $x_0 = \mathcal{D}(z_0)$. SD follows the Denoising Diffusion Probabilistic Model (DDPM) [20] framework, introducing Gaussian noise ϵ to the latent z_0 across timesteps t, generating a noisy latent z_t over a series of steps. During inference, the model denoises z_t back to z_0, guided by condition features. The denoising backbone ϵ_θ, based on a U-Net [41], is trained to predict the noise and remove it progressively, using the objective:

$$L = \mathbb{E}_{t,c,z_t,\epsilon}\left[\|\epsilon - \epsilon_\theta(z_t, t, c)\|^2\right],$$

where c represents text features derived from a CLIP encoder [37]. SD uses a U-Net with cross-attention mechanisms, to fuse text embeddings with latent features, enabling fine control over generated images based on text prompts. This allows SD to generate detailed, high-fidelity images while responding effectively to user input.

3.2 Hybrid Face Swapping Framework

Video Face Swapping Task. Video face swapping aims to seamlessly transfer a source face identity onto a target video while preserving the target's pose, expression, lighting, and background. Although recent works like DiffSwap [55, 23, 1] have demonstrated promising results for static image face swapping, extending these approaches directly to videos presents substantial challenges. These include temporal distortions, flickering, occlusion sensitivity, and difficulties in managing significant pose variations.

To effectively tackle these challenges, we propose VividFace, a diffusion-based video face swapping framework specifically designed for robust, temporally coherent, and high-fidelity results. VividFace introduces a novel hybrid training strategy that leverages abundant image-level data alongside temporal video data. This hybrid approach enhances diversity and robustness in training, significantly mitigating issues prevalent in video-only methods. Our framework initially encodes both source

images $x^i_{src} \in \mathbb{R}^{1\times3\times H\times W}$ and video sequences $x^v_{src} \in \mathbb{R}^{T\times3\times H\times W}$ into a unified latent space $z_0 \in \mathbb{R}^{T\times C\times H\times W}$ using a specially designed VAE. Static images are treated equivalently as single-frame videos, ensuring consistent embedding. Subsequently, we train a conditional diffusion model $\epsilon_\theta(z_t, t; \mathbf{C})$ that performs latent space denoising, emphasizing temporal consistency and identity fidelity. Here, $\mathbf{C}$ represents the conditioning vectors, and t indicates the denoising timestep. Due to the absence of ground truth data when source and target images originate from different individuals, our model uses pairs of face images from the same identity during training. As depicted in Figure 2, training batches alternate between static image data and video sequences, ensuring efficient gradient synchronization and optimal learning dynamics.

VidFaceVAE. As shown in Figure 3, our proposed VidFaceVAE is a VAE framework designed to enhance the reconstruction quality of facial data, effectively handling both video sequences and static images. The VidFaceVAE primarily consists of (2+1)D blocks, combining 2D spatial and 1D temporal convolutions to form pseudo-3D operators. For image inputs, the STFM (Spatial Temporal Fusion Module) outputs the result of the 2D ResBlock directly, bypassing the temporal ResBlock. For video inputs, the STFM combines the outputs from both the 2D and temporal blocks using a learnable coefficient β, described as $o = \beta \times o_{\text{spatial}} + (1-\beta) \times o_{\text{temporal}}$, where $o_{spatial}$ and o_{temporal} denote the output from the spatial branch and the temporal branch. We do not involve the temporal downsampling modules in our VAE framework as it needs to process image data. The (2+1)D structure of VidFaceVAE provides two primary advantages: (1) it significantly reduces computational complexity compared to full 3D convolutions, and (2) it facilitates leveraging pretrained 2D VAE parameters and stable diffusion (SD) weights, accelerating training convergence and enhancing overall performance. In contrast to OD-VAE [8], our approach does not utilize 3D causal convolutions due to their limited capacity, particularly when processing static images, and our backbone architecture does not rely on transformer-based designs.

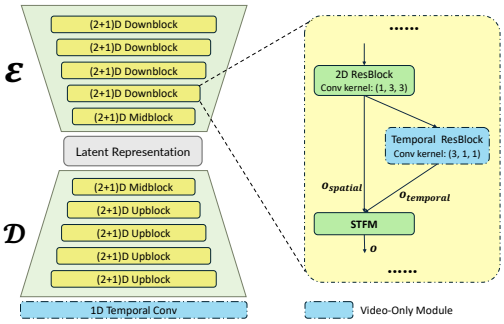


Figure 3: Overview of the proposed VidFaceVAE, capable of simultaneous encoding and decoding of both image and video data. Specific modules are designed for video inputs, which image inputs bypass when necessary.

Temporal Modules. Inspired by recent approaches like EMO [48] and AnimatedDiff [18], we introduce self-attention temporal layers specifically designed for video sequences. During training with video data, we prepare additional motion frames x^{motion}_{src} to enrich temporal context. The temporal attention mechanism combines these motion features with target frames at matching resolutions along the temporal dimension, improving temporal coherence. To smooth the first video clip's generation, motion frames are initialized as zero vectors during training.

3.3 Designs of Condition Vectors

In our framework, several carefully designed condition vectors are used to guide the generation process, ensuring accurate and consistent visual outputs for both static images and video sequences. We formulate video face swapping as a conditional inpainting task, where masked videos with cropped face regions provide the background and lighting conditions. The corresponding face regions guide the diffusion model on where are generated the faces.

In many in-the-wild videos, faces often exhibit significant pose variations, which can lead diffusion models to produce suboptimal results, such as facial distortions and inaccurate pose estimations. To address this issue, we propose using a 3D reconstruction technique to reconstruct the face and use its output as local guidance for pose and expression details. Specifically, we employ 3DMM [5] to extract BFM (Basel Face Model) coefficients, setting the texture and identity component to zero to reduce information leakage. Replacing the reconstructed face with the original target would introduce further leakage, as the ground truth face is identical to the input, which could impair the model's generalization ability. To ensure that the generated face maintains the same identity as the source

Figure 4: Visualization of our AIDT dataset, which includes 1M image triplets and 0.6M video pairs. For video facial data, we only present the target and decoupling faces, as source faces can be derived from any other frame within the same clip.

Figure 5: Visualization of our occlusion data augmentation, which improves the stability and consistency of the generated videos.

face while preserving attributes (such as pose, expression, *etc.*,), we inject cross-attention features **C** extracted by our face encoder as global context to the diffusion model.

Face Encoder. The face encoder module in our framework plays a critical role in extracting and integrating features from the target and source faces to guide the face-swapping process effectively. As illustrated in the right part of Figure 2, the face encoder is composed of three primary networks, each responsible for capturing distinct aspects of facial information: (1) identity net: this network focuses on extracting the core identity features from the target face; (2) texture net: this network is designed to capture detailed texture information from the target face, such as skin tone, fine facial features; (3) attribute net: the net extracts additional facial attributes from the source, such as pose, expression, and other dynamic features that contribute to a realistic and expressive representation.

The straightforward approach is to send the source image to both the identity and texture networks, while the target image is sent to the attribute network. However, a challenge arises when the source and target faces do not belong to the same person, as the ground truth is unavailable in the real world. In most previous methods [1, 55, 23], the source and target images are assumed to be the same, meaning all three networks receive identical input. This results in difficulties for the face encoder in extracting distinct features and leading to information leakage. Specifically, this leakage causes the model to merely “copy and paste" the face region, effectively completing the task by superficially transferring facial features without meaningful feature disentangling or transformation. In contrast, our framework, built on the AIDT dataset (shown in Figure 4), employs source images (same identity, but different attributes) and decoupling images (same attribute, but different identity). These images guide the face encoder to disentangle and fuse facial features, improving generalization across different identities during inference.

Within the Mixer module, the extracted features first undergo cross-attention operations to capture mutual dependencies. The outputs are then scaled by learned weights and fused via weighted sum. This process combines the identity, texture, and attribute features to create a comprehensive cross-attention feature representation **C**. This fused representation offers rich context to guide the diffusion model during face generation, ensuring high fidelity and identity consistency across video frames.

3.4 Training Strategy

Our training process involves three stages to progressively enhance model performance for video face swapping. The first stage focuses on training the VidFaceVAE , where we apply reconstruction, perceptual, and KL divergence losses to ensure high-quality reconstruction and a well-structured latent space. The training data primarily consists of facial images and videos. Given the specifically designed architecture, the spatial modules are initialized using the original 2D VAE. In subsequent stages, the VAE is frozen and no longer updated. In the second stage, we pretrain the model using image data, while the ReferenceNet and temporal modules of the backbone network remain inactive.

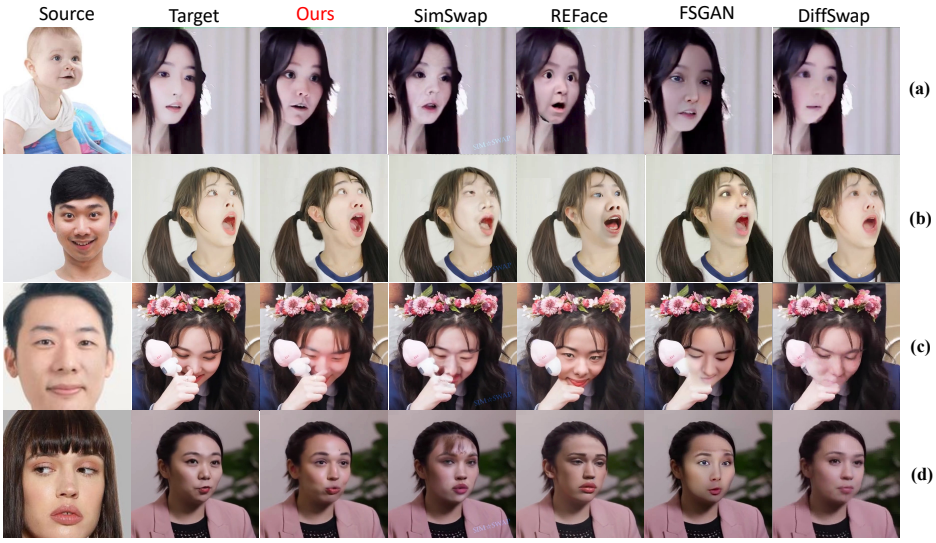


Figure 6: Qualitative comparison at 512×512 resolution. Our method generates high-fidelity results and handles challenging cases effectively, such as large poses (b) and occlusions (c).

The backbone is initialized from the original SD weights. Finally, we perform image-video hybrid training by activating temporal modules and introducing video data, initializing temporal modules from AnimateDiff [18] for effective temporal consistency and smooth frame transitions.

4 AIDT Dataset

In this section, we describe the construction of triplet pairs for our AIDT (Attribute-Identity Disentanglement Triplet) dataset, which includes 1M image triplets and 0.6M video triplets, as shown in Figure 4.The data was generated via a pipeline of collection, detection, tracking, and post-processing. See Appendix B for details. The dataset helps the face encoder to disentangle and fuse distinct facial components—ID features, texture features from the source face, and attribute features from the decoupling face. This enhances generalization, especially when the source and target faces belong to different individuals during inference.

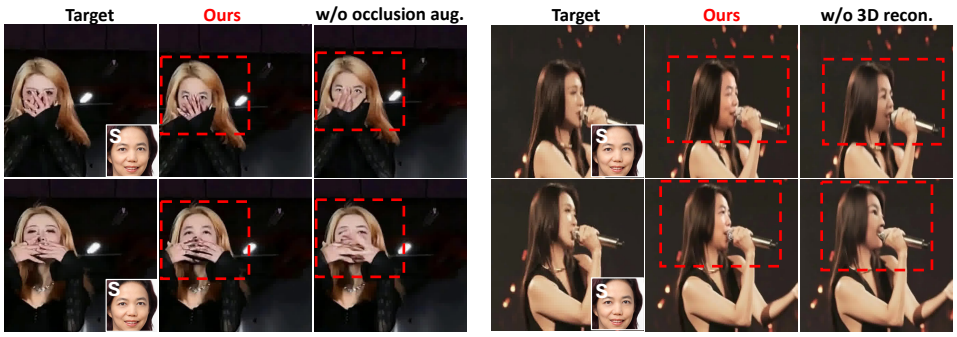
In addition, Figure 5 showcases our occlusion augmentation pipeline. We collect hand and everyday object images, along with their corresponding masks, from web data sources. To simulate temporal dynamics, we design a motion trajectory animation scheme that incorporates scale, rotation, and translation over time. This augmentation strategy improves robustness under occlusion and enhances the temporal consistency of synthesized videos.

5 Experiment

Detailed information regarding dataset preparation, network architectures, hyperparameters, and training procedures can be found in Appendix A.

5.1 Evaluation Protocol

Considering that most previous baselines, such as CelebA [25] and FFHQ [21], are primarily focused on image face swapping, we propose a new benchmark for video face swapping, VidSwapBench. Our benchmark includes 200 source images and 200 high-resolution target videos, with each video containing 128 frames and a single trackable face. These videos and images feature unseen identities and backgrounds, ensuring a diverse and challenging dataset. To evaluate performance, we generate 200 swapped videos using our framework. For comparison, since other methods are based on image-level face swapping, we perform face swapping frame by frame for those methods. For facial data reconstruction, we use SSIM, PSNR and LPIPS [54] to evaluate the quality of reconstructed images and videos. For video face swapping, we use FVD [49] to assess the overall quality of the generated videos. The attribute transfer error is measured by pose and expression errors. We use HopeNet [42]



(a) (b)

Figure 7: Ablation on the occlusion data augmentation and 3D face reconstruction.

and Deep3DFaceRecon [13] to detect these attributes, and the L2 distance to the ground truth is used as the evaluation metric. For ID retrieval, we extract identity features from the source images using ArcFace, and for each swapped video, we perform face retrieval by searching for the most similar faces among all source images. The retrieval is measured by the average cosine similarity of all frames, and we report the Top-1 and Top-5 accuracy.

Table 1: Comparison on VidSwapBench dataset.

Method	$FVD_{32}\downarrow$	$FVD_{128}\downarrow$	ID@1↑	ID@5↑	Pose↓	Expr.↓
MegaFS [56]	1280.3	200.7	72.8	82.1	6.21	0.74
HifiFace [51]	1377.3	386.4	74.1	82.9	6.10	0.74
SimSwap [9]	1242.8	186.6	76.5	88.5	**5.12**	0.76
FSGAN [33]	1507.9	423.8	24.5	40.0	5.19	0.73
DiffFace [23]	2404.7	1404.9	1.5	4.1	18.3	1.58
DiffSwap [55]	1530.2	809.3	14.5	26.3	12.9	1.02
REFace [1]	1336.9	311.9	71.9	86.5	6.67	0.91
Ours	**1201.1**	**122.6**	**78.3**	**90.2**	5.43	**0.72**

Table 2: Comparison on FFHQ dataset.

Method	FID↓	ID@1↑	ID@5↑	Pose↓	Expr.↓
MegaFS [56]	12.0	59.6%	74.1%	3.33	1.11
HifiFace [51]	11.58	75.3%	87.1%	3.28	1.41
SimSwap [9]	13.8	90.6%	96.4%	2.98	1.07
E4S [29]	12.38	70.2%	82.73%	4.50	1.31
DiffFace [23]	8.59	87.2%	94.4%	3.80	2.28
DiffSwap [55]	8.58	78.2%	93.6%	**2.92**	1.10
REFace [1]	5.53	95.4%	98.7%	3.74	**1.04**
Ours	**4.05**	**96.9%**	**99.2%**	3.70	1.07

5.2 Comparisons with Existing Methods

Qualitative Results. Since videos cannot be displayed in the PDF and due to submission policy restrictions on showing generated videos, we provide several comparison videos in the supplementary materials and strongly encourage the reader to view them. We perform quantitative comparison at 512×512 resolution. As shown in Figure 6 (a) and (d), our method generates high-fidelity swapped faces, with attributes that closely match the target faces. In Figure 6 (b), our method successfully transfers both face shape and expression under large pose variations, benefiting from the 3D reconstruction mask, while other methods exhibit generation artifacts. In Figure 6 (c), where a toy and hand occlude the girl's face, most other methods fail to handle the occlusion properly, with the toy and hand either displaced or fused together. Additionally, many methods result in noticeable facial deformations. In contrast, our method successfully recovers the occluded areas and maintains accurate face swapping, thanks to our augmentation strategy.

Quantitative Results. In Table 1, we compare seven open-source methods (four GAN-based and three diffusion-based). The results show that our method outperforms others in ID retrieval and FVD, generating high-fidelity swapped face videos while preserving the source identity. It also achieves comparable performance in pose and expression, maintaining target attributes effectively. Furthermore, since our model supports both image and video face swapping, we also evaluate it on the standard FFHQ dataset. As shown in Table 2, VividFace achieves state-of-the-art results in FID and ID retrieval, while delivering comparable performance in pose and expression preservation.

5.3 Ablation Studies

We conducted comprehensive ablation experiments to analyze the contributions of different components and design choices within our framework. The quantitative results are summarized in Table 3, and qualitative visualizations are presented in Figure 7. **Hybrid vs. Static Training (Exp. 1):** training exclusively with static images results in decreased identity preservation (Top-1 accuracy

Table 3: Ablation on training strategies and module designs

Exp Id	Method	$FVD_{32}\downarrow$	$FVD_{128}\downarrow$	ID retrieval↑ Top-1	ID retrieval↑ Top-5	Pose↓	Expr.↓
0	Baseline	1201.1	122.6	78.3	90.2	5.43	0.72
1	Static Training	1231.9	128.1	75.7	88.1	5.60	0.74
2	Without 3DMM Guidance	1197.4	121.3	78.1	90.4	5.55	0.74
3	Merged ID and Texture Net	1203.8	121.9	76.4	87.1	5.49	0.73
4	Init. Texture/Attr Net with CLIP	1237.9	138.2	76.0	86.1	5.30	0.72

Table 4: Comparison of different VAE architectures

Architecture Encoder	Architecture Decoder	Facial videos SSIM↑	Facial videos PSNR↑	Facial videos LPIPS↓
2D	2D	0.967	37.61	0.048
2D	(2+1)D	0.976	38.77	0.039
(2+1)D	(2+1)D	**0.983**	**41.11**	**0.027**

reduced from 78.3 to 75.7) and higher FVD scores, underscoring the critical advantage of our hybrid training approach in achieving superior temporal consistency and identity fidelity. **Impact of 3D Reconstruction Conditioning (Exp. 2):** removing the 3DMM conditioning slightly reduces pose and expression accuracy, highlighting the effectiveness of 3DMM guidance in handling complex pose variations. **Separate Identity and Texture Networks (Exp. 3):** combining identity and texture features into a single network leads to reduced identity retrieval accuracy, emphasizing the necessity of separate networks for effectively disentangling and extracting distinct facial features. **Initialization Strategy (Exp. 4):** initializing texture and attribute networks with CLIP weights negatively affects identity preservation, demonstrating the effectiveness of our proposed initialization strategy. **Impact of Occlusion Augmentation:** the qualitative results in Figure 7 (a) show severe distortions when occlusion augmentation is omitted. Introducing occlusion augmentation substantially improves stability, consistency, and visual quality under occlusion scenarios. **Effect of 3DMM Conditioning on Large Pose Variations:** Figure 7 (b) illustrates that excluding 3D face reconstruction guidance results in significant instability and distortion for large pose variations.

VAE Architecture Analysis. Table 4 compares reconstruction performances of different VAE architectures. The baseline (SD-VAE) employs a pure 2D structure, the second model utilizes a (2+1)D decoder combined with a 2D encoder, and our proposed VidFaceVAE employs a full (2+1)D encoder-decoder. VidFaceVAE outperforms alternative architectures across all metrics, achieving the highest SSIM (0.983), PSNR (41.11), and lowest LPIPS (0.027). This clearly demonstrates the advantage of effectively integrating spatial and temporal processing in facial video reconstruction.

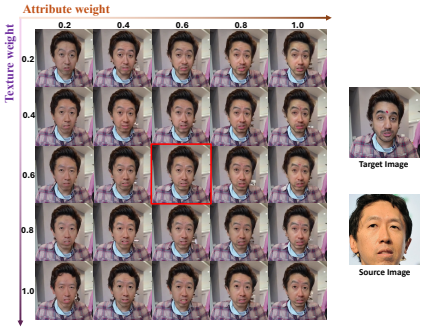


Figure 8: Ablation on the different combinations of texture weights and attribute weights.

5.4 Face Feature Mixing Analysis

We further examine the impact of varying texture and attribute weights within the face encoder. Figure 8 shows that increasing texture weight improves identity similarity but can degrade pose and expression preservation if too high. Higher attribute weights maintain target features but reduce identity fidelity. Optimal performance is achieved by balancing these weights, set to 1.0 for identity and 0.6 for texture and attribute in our experiments.

6 Conclusion

In this paper, we introduced a novel diffusion-based framework for video face swapping, addressing key challenges such as temporal consistency, identity preservation, and large pose variations. Our image-video hybrid training strategy leverages both static images and video data, improving model diversity and robustness. The VidFaceVAE , coupled with a custom Attribute-Identity Disentanglement Triplet (AIDT) dataset and 3D Morphable Model integration, enables accurate face swapping while mitigating issues like flickering and occlusions. Our framework outperforms existing methods in FVD, temporal consistency, and identity preservation, with fewer inference steps. Together with the released dataset, it provides a more efficient solution for high-quality video face swapping and lays the foundation for future advancements.

References

[1] Sanoojan Baliah, Qinliang Lin, Shengcai Liao, Xiaodan Liang, and Muhammad Haris Khan. Realistic and efficient face swapping: A unified approach with diffusion models. *arXiv preprint arXiv:2409.07269*, 2024.

[2] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Towards open-set identity preserving face synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6713–6722, 2018.

[3] Dmitri Bitouk, Neeraj Kumar, Samreen Dhillon, Peter Belhumeur, and Shree K Nayar. Face swapping: automatically replacing faces in photographs. In *ACM SIGGRAPH 2008 papers*, pages 1–8. 2008.

[4] Volker Blanz, Kristina Scherbaum, Thomas Vetter, and Hans-Peter Seidel. Exchanging faces in images. In *Computer Graphics Forum*, volume 23, pages 669–676. Wiley Online Library, 2004.

[5] 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.

[6] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[8] Liuhan Chen, Zongjian Li, Bin Lin, Bin Zhu, Qian Wang, Shenghai Yuan, Xing Zhou, Xinghua Cheng, and Li Yuan. Od-vae: An omni-dimensional video compressor for improving latent video diffusion model. *arXiv preprint arXiv:2409.01199*, 2024.

[9] 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.

[10] 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*, 2023.

[11] DeepFakes. faceswap. `https://github.com/deepfakes/faceswap`, 2020. Accessed: 2024-11-01.

[12] 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.

[13] 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 *IEEE Computer Vision and Pattern Recognition Workshops*, 2019.

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

[15] Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baochang Zhang. Implicit diffusion models for continuous super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10021–10030, 2023.

[16] Ian 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.

[17] Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. Sample and computation redistribution for efficient face detection. *arXiv preprint arXiv:2105.04714*, 2021.

[18] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*.

[19] 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.

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

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

[22] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models, 2023.

[23] 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. *arXiv preprint arXiv:2212.13344*, 2022.

[24] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[25] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5549–5558, 2020.

[26] Jaeseong Lee, Junha Hyung, Sohyun Jung, and Jaegul Choo. Selfswapper: Self-supervised face swapping via shape agnostic masked autoencoder. In *European Conference on Computer Vision*, pages 383–400. Springer, 2024.

[27] 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.

[28] Haonan Lin, Mengmeng Wang, Jiahao Wang, Wenbin An, Yan Chen, Yong Liu, Feng Tian, Guang Dai, Jingdong Wang, and Qianying Wang. Schedule your edit: A simple yet effective diffusion noise schedule for image editing. *arXiv preprint arXiv:2410.18756*, 2024.

[29] Zhian Liu, Maomao Li, Yong Zhang, Cairong Wang, Qi Zhang, Jue 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*, pages 8578–8587, 2023.

[30] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[31] Bingqi Ma, Zhuofan Zong, Guanglu Song, Hongsheng Li, and Yu Liu. Exploring the role of large language models in prompt encoding for diffusion models. *arXiv preprint arXiv:2406.11831*, 2024.

[32] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024.

[33] 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.

[34] Yuval Nirkin, Iacopo Masi, Anh Tran Tuan, Tal Hassner, and Gerard Medioni. On face segmentation, face swapping, and face perception. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 98–105. IEEE, 2018.

[35] Rogue One. Rogue one: a star wars story. *Genre*, 14, 2016.

[36] Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Carl Shift Facenheim, Luis RP, Jian Jiang, Sheng Zhang, et al. Deepfacelab: Integrated, flexible and extensible face-swapping framework. *arXiv preprint arXiv:2005.05535*, 2020.

[37] 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.

[38] Xiaohang Ren, Xingyu Chen, Pengfei Yao, Heung-Yeung Shum, and Baoyuan Wang. Reinforced disentanglement for face swapping without skip connection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20665–20675, 2023.

[39] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13759–13768, 2021.

[40] 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.

[41] Olaf Ronneberger, Philipp Fischer, and Thomas 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.

[42] Nataniel Ruiz, Eunji Chong, and James M Rehg. Fine-grained head pose estimation without keypoints. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 2074–2083, 2018.

[43] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models. *arXiv preprint arXiv:2403.16999*, 2024.

[44] Yujun Shi, Chuhui Xue, Jun Hao Liew, Jiachun Pan, Hanshu Yan, Wenqing Zhang, Vincent YF Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8839–8849, 2024.

[45] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[46] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3667–3676, 2020.

[47] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016.

[48] Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive-generating expressive portrait videos with audio2video diffusion model under weak conditions. *arXiv preprint arXiv:2402.17485*, 2024.

[49] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019.

[50] Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C Kot, and Bihan Wen. Sinsr: diffusion-based image super-resolution in a single step. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25796–25805, 2024.

[51] 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.

[52] Yifan Wu, Fan Yang, Yong Xu, and Haibin Ling. Privacy-protective-gan for privacy preserving face de-identification. *Journal of Computer Science and Technology*, 34:47–60, 2019.

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

[54] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[55] 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.

[56] 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.

[57] Zhuofan Zong, Dongzhi Jiang, Bingqi Ma, Guanglu Song, Hao Shao, Dazhong Shen, Yu Liu, and Hongsheng Li. Easyref: Omni-generalized group image reference for diffusion models via multimodal llm. In *Forty-second International Conference on Machine Learning*, 2024.

[58] Zhuofan Zong, Bingqi Ma, Dazhong Shen, Guanglu Song, Hao Shao, Dongzhi Jiang, Hongsheng Li, and Yu Liu. Mova: Adapting mixture of vision experts to multimodal context. *arXiv preprint arXiv:2404.13046*, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The introduction clearly states the paper's main contributions, including the proposal of the VividFace framework, the hybrid training strategy, the AIDT dataset, occlusion augmentation, and 3DMM-based conditioning. These claims are aligned with the paper's technical content and are appropriately scoped based on the presented methods and experiments.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of the proposed method are discussed in the appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides detailed implementation specifics, including architectural design, training procedures, and evaluation metrics. Additionally, the authors offer access to the codebase and pre-trained models, and commit to releasing the AIDT dataset, enabling thorough reproducibility of all major experimental results and supporting the paper's main claims.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The paper provides access to the code and pre-trained models, along with detailed instructions for reproducing the experiments. Additionally, the authors commit to releasing the AIDT dataset, ensuring that all essential components needed for faithful reproduction of the main results are available or will be made available.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper includes comprehensive descriptions of training and evaluation settings, including data splits, model architecture, optimizer type, learning rate, and other hyperparameters.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

Justification: The paper does not report error bars or statistical significance metrics. Due to the high computational cost associated with training diffusion-based video models, repeated trials under varying random seeds or data splits were not conducted.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper provides relevant information in the experiment section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work complies with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses both positive impacts—such as applications in privacy protection, digital humans, and safe content creation—and potential risks, including misuse for deepfakes or disinformation, in the appendix.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: The paper acknowledges potential misuse risks and outlines safeguards, including planned release under usage restrictions, documentation of intended use, and guidelines for responsible deployment. Additionally, safety considerations are incorporated into dataset construction to avoid harmful or inappropriate content.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We use publicly available datasets and codebases, all of which are properly cited in the paper. License terms (e.g., CC-BY, MIT) are respected, and URLs and version information are provided where applicable.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The paper introduces new assets with clear documentation on their structure, usage, and licensing in the supplementary materials.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs were used solely for grammar checking and language polishing. They did not contribute to the core methodology, experiments, or scientific content of the research.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Implementation Details

We collected approximately 550 hours of facial videos from the internet to train our models, and the facial images are partially sourced from VGGFace2-HQ [10]. In our experiments, we use a latent space of size $13 \times 64 \times 64$ and a U-Net architecture for the ϵ_0 denoising network. Images and video clips sampled from the dataset are resized and cropped to 512×512. The number of motion frames, M, is set to 4, and the generated video length, T, is set to 8 frames. For the face encoder, the identity network is based on ArcFace [12], while the texture and attribute networks are based on DINO [7]. We use the SCRFD [17] for facial bounding box detection. The mixing coefficients of the ace encoder are set to 1.0 for identity features, and 0.6 for both texture and attribute features. The experiments are conducted using 16 NVIDIA A100 GPUs and optimized with AdamW [30]. In the first stage of the VAE training, the learning rate is set to 5e-6 with a batch size of 32. The weights of reconstruction, perceptual, and KL divergence loss are 1.0, 0.1, 1e-6 respectively. For the second and third stages, the learning rate is increased to 1e-5, with the batch size remaining at 32. During inference, we generate video clips using the DDIM sampling algorithm for 32 steps.

B AIDT Dataset details

This section describes our approach to constructing the AIDT (Attribute-Identity Disentanglement Triplet) dataset, which consists of two parts: video clip collection and triplet pairs construction.

B.1 Video Clip Collection Pipeline

Video Collection. We collect facial videos from public platforms to capture a diverse range of visual and auditory content. The video corpus includes two primary categories: static content, such as news broadcasts, interviews, and public speaking events; and dynamic content, encompassing genres like travel guides, vlogs, and musical performances. This variety ensures a comprehensive dataset, representing both controlled and spontaneous human expressions and activities across different scenarios.

Face Detection and Tracking. For each video, we first apply a face detection model to identify face bounding boxes in each frame. Subsequently, all face detections are fed into a tracking procedure to generate face tracklets across frames. Each resulting tracklet is then split into video clips, each containing between 30 and 200 frames. Finally, we crop each video clip to center the head within the frame and to maintain a consistent aspect ratio.

Data Post-Processing. Based on the preliminary clips generated in the previous stage, we further refine the clips by applying three key constraints to collect final training data. We begin by utilizing HyperIQA [46] to filter out low-quality clips, setting an average quality assessment threshold score of 50 to ensure visual fidelity. Second, we enforce identity consistency within each clip by using a face recognition model to extract facial features and compute cosine similarity scores between every pair of frames, thus verifying that each clip represents a single individual. Finally, we utilize an OCR model to exclude clips containing text near the facial region, minimizing visual distractions that could interfere with model training.

B.2 Construction of triplet training dataset

For image data, we first cluster the facial images based on identity similarity. From each cluster, we randomly select two images to form a target-source pair that shares the same identity but has different attributes. To generate the decoupling image, which has a different identity but the same attribute, we use the open-sourced InsightFace to create synthetic images with a distinct identity, while preserving the gender of the original face. This approach helps to avoid the degradation in quality observed when the original and swapped faces belong to different genders. Additionally, we exclude triplets with significant facial expression discrepancies by comparing the face landmarks. For video data, the process is similar, except that both the source and target images come from the same video clip, but not from the same frames as the target or motion images, which reduces the pose variation. Since video data is less abundant than image data, clustering does not yield enough pairs to form a sufficient number of triplets.

Figure 9: Qualitative comparison at 512×512 resolution. Our method generates high-fidelity results and handles challenging cases effectively, such as large poses (b) and occlusions (c). Corresponding videos are provided in the supplementary material.

C More Visualization

In Figure 9, we show additional frames from the same video presented in the main text (Figure 6). We further provide supplementary visual examples in Figure 10, Figure 11, Figure 12, Figure 13, Figure 14, and Figure 15. The corresponding video results for all examples are included in the supplementary materials.

D Limitations

While VividFace demonstrates strong temporal consistency, high-fidelity visual quality, and robustness in challenging scenarios, it also inherits certain limitations associated with diffusion-based architectures. Specifically, compared to GAN-based or lightweight encoder-decoder approaches, diffusion-based models typically require more inference steps to generate each frame, resulting in slower runtime performance. This can limit the practical applicability of our method in real-time or latency-sensitive scenarios, such as live video processing or interactive applications. Although we adopt efficient design choices and optimize the number of inference steps, the trade-off between generation speed and output quality remains an open challenge. Future work could explore accelerated diffusion sampling strategies or hybrid approaches that combine the strengths of both diffusion and feed-forward architectures to further improve efficiency without compromising quality.

Figure 10: Qualitative comparison at 512×512 resolution. Corresponding videos are provided in the supplementary material.

E Broader Impacts

VividFace advances video face swapping with high visual fidelity and temporal consistency, offering potential positive societal impacts in several domains. These include privacy protection through identity anonymization, safer film and content production by reducing the need for risky physical stunts, the creation of digital avatars in virtual environments, and accessible tools for individuals with communication or appearance-related challenges. However, the same capabilities may lead to negative societal consequences if misused. High-quality face-swapping techniques can be exploited for generating deepfakes, which may contribute to misinformation, identity theft, and other forms of digital deception. These risks are particularly relevant in political, social, and journalistic contexts where visual integrity is crucial. To address these concerns, we plan to release our models and datasets with strict usage terms and clear documentation. We encourage responsible use and support the development of detection tools and watermarking techniques to distinguish generated content from real footage. Furthermore, we advocate for continued dialogue in the research community around ethical deployment, regulatory considerations, and public education to minimize harm while enabling beneficial use cases. As with any powerful generative technology, the societal impact of VividFace will depend not only on the tool itself, but on how it is governed and integrated into real-world systems.

Figure 11: Qualitative comparison at 512×512 resolution. Corresponding videos are provided in the supplementary material.

Figure 12: Qualitative comparison at 512×512 resolution. Corresponding videos are provided in the supplementary material.

Figure 13: Qualitative comparison at 512×512 resolution. Corresponding videos are provided in the supplementary material.

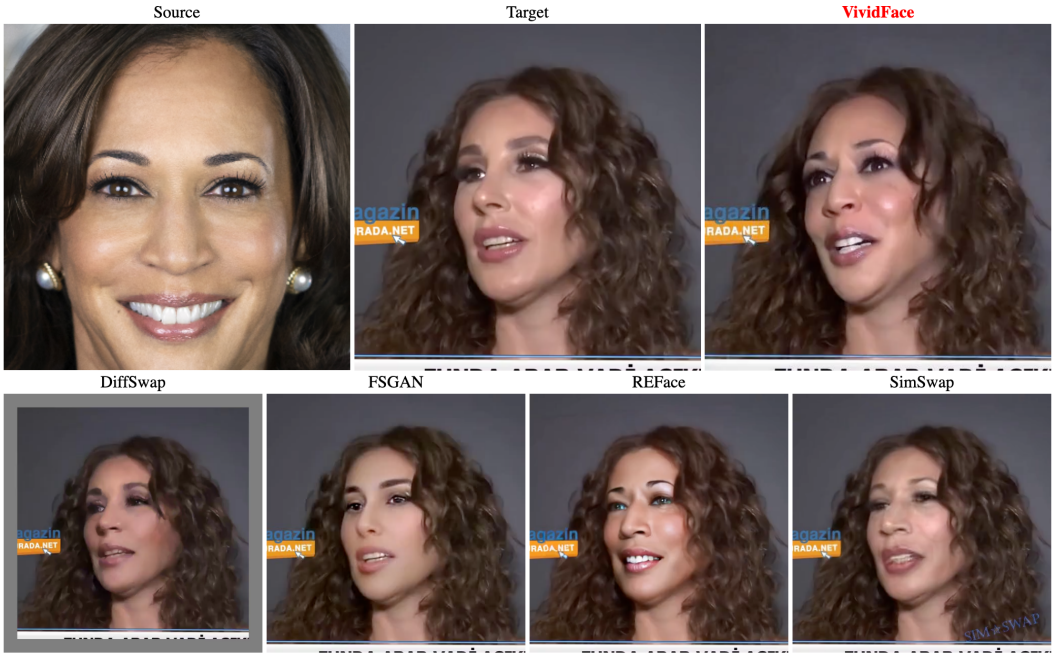


Figure 14: Qualitative comparison at 512×512 resolution. Corresponding videos are provided in the supplementary material.

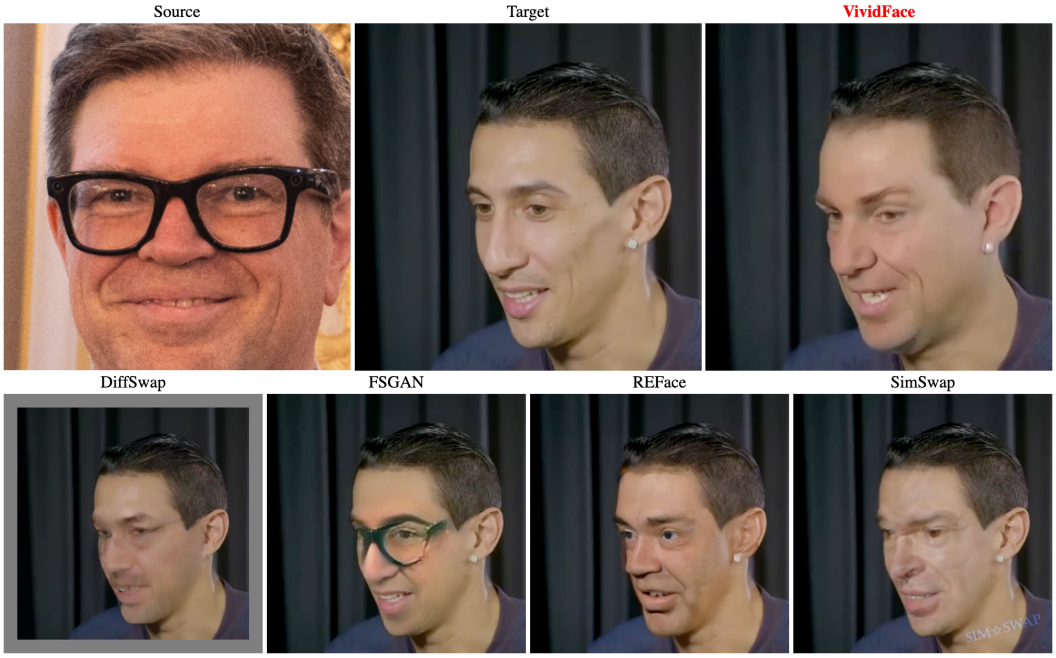


Figure 15: Qualitative comparison at 512×512 resolution. Corresponding videos are provided in the supplementary material.