

OmniFace: Bridging the Image-to-Video Gap for High-Fidelity Face Swapping via Diffusion Transformer

Xu Guo^{1,*}, Fulong Ye^{2,*}, Xinghui Li^{2,*}, Pengqi Tu², Pengze Zhang², Qichao Sun², Songtao Zhao^{2,†}, Xiangwang Hou^{1,†}, and Qian He²

¹ Tsinghua University, Beijing, China

² Intelligent Creation Lab, ByteDance, Beijing, China

guo-x24@mails.tsinghua.edu.cn, zhaosongtao.0815@bytedance.com,
xwhou@mail.tsinghua.edu.cn

Abstract. Video Face Swapping (VFS) requires seamlessly injecting a source identity into a target video while meticulously preserving the original pose, expression, lighting, background, and dynamic information. Existing methods struggle to maintain identity similarity and attribute preservation while preserving temporal consistency. To address the challenge, we propose a comprehensive framework to seamlessly transfer the superiority of Image Face Swapping (IFS) to the video domain. We first introduce a novel data pipeline *SyncID-Pipe* that pre-trains an Identity-Anchored Video Synthesizer and combines it with IFS models to construct bidirectional ID quadruplets for explicit supervision. Building upon paired data, we propose a powerful Diffusion Transformer-based framework *OmniFace*, employing a core Modality-Aware Conditioning module to discriminatively inject multi-model conditions. Meanwhile, we propose a Synthetic-to-Real Training mechanism and an Identity-Coherence Reward Weighting strategy to enhance visual realism and identity consistency under challenging scenarios. To address the issue of limited benchmarks, we introduce *IDBench-V*, a comprehensive benchmark encompassing diverse scenes. Extensive experiments demonstrate *OmniFace* outperforms state-of-the-art methods and further exhibits exceptional versatility, which can be seamlessly adapted to various swap-related tasks.

Keywords: Video Face Swapping · Diffusion Transformer · Paired Data

1 Introduction

Face swapping aims to generate an image or video that combines the identity of a source face with the attributes (such as background, pose, expression, lighting) from a target image or video. This technique has sparked considerable research

* Equal contribution.

† Corresponding author.

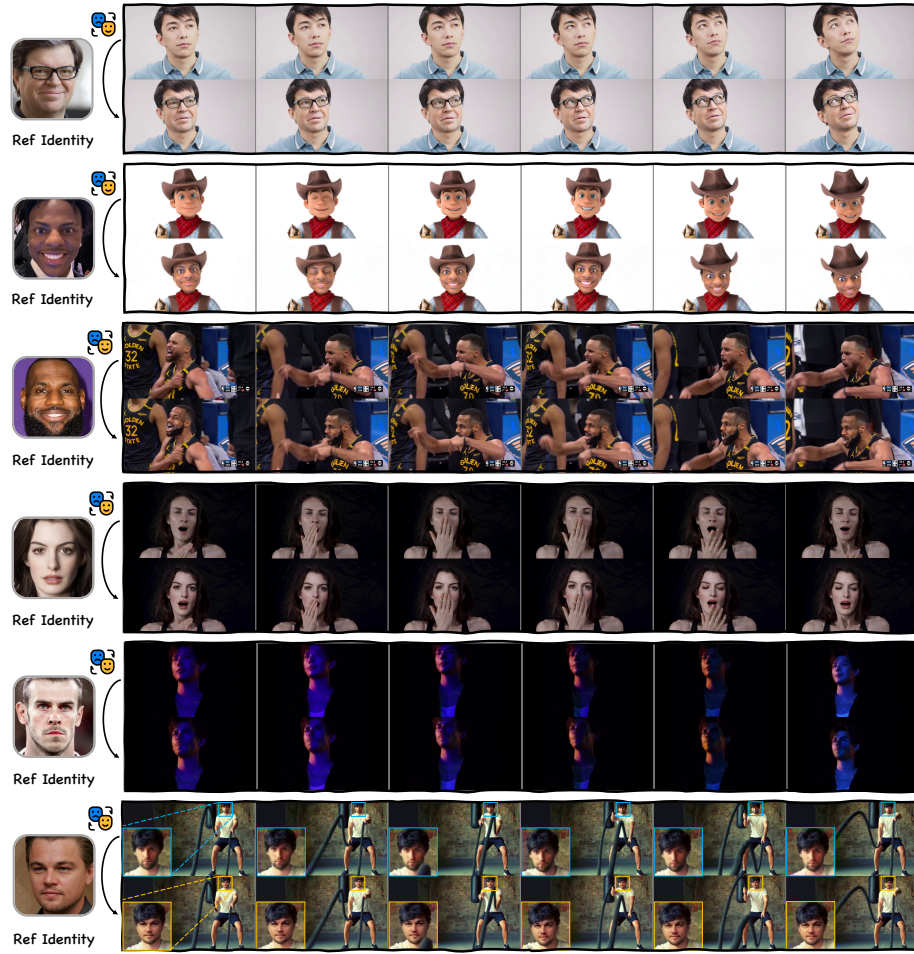


Fig. 1: Showcase of *OmniFace*. *OmniFace* robustly handles challenging scenarios, e.g., complex expressions, animation, large angles, occlusions, complex lighting and small faces.

interest, due to its significant potential for practical applications in film production, creative design and privacy protection. Unlike Image Face Swapping (IFS), Video Face Swapping (VFS) presents more challenges, as it introduces additional critical constraints on temporal identity continuity, pose consistency, and environment preservation.

Existing studies on IFS, such as [10, 45], have achieved remarkable success in maintaining identity similarity and preserving attributes. However, directly applying these IFS methods frame-by-frame to video sequences often leads to significant challenges in temporal consistency, resulting in noticeable flickering and jittering artifacts. Recently, the rapid advancement of diffusion-based video

generation models has greatly propelled the development of VFS. While methods like VividFace [34], DynamicFace [39], HiFiVFS [4], and CanonSwap [27] have improved the coherence and generation quality of VFS, their capabilities in terms of identity similarity and attribute preservation still lag behind those of state-of-the-art IFS models. The fundamental difference between IFS and VFS lies in the dynamic nature of video, which requires consistent preservation of motion and expression across frames. This observation inspires us to explore whether we can bridge the gap between image and video domains by supplementing these dynamic signals, thereby harnessing the strengths of IFS to significantly boost VFS performance.

Building on these insights, we propose a comprehensive framework comprising a novel data pipeline and a customized architecture to enhance VFS performance significantly. Our proposed data curation pipeline, ***SyncID-Pipe***, seamlessly transfers the superiority of IFS to the video domain. Specifically, the pipeline pretrains a pose-driven First-Last-Frame video generation model, which we term the Identity-Anchored Video Synthesizer (IVS). The IVS employs an adaptive pose attention mechanism to inject pose information into First-Last-Frame video foundation models. This enables the model to generate a video consistent with the content of the given start and end frames and the actions specified by the pose sequence. Subsequently, we combine it with IFS models to construct bidirectional ID quadruplets for explicit supervision. Furthermore, to enhance data reliability, we propose an expression adaptation strategy to achieve effective expression transfer, incorporating an enhanced background recomposition mechanism to ensure strict environment alignment in the paired videos.

Building upon paired data, we develop ***OmniFace***, a powerful video face swapping framework based on Diffusion Transformer (DiT) models [31], achieving high-similarity and superior-coherence results. We first introduce a Modality-Aware Conditioning (MC) mechanism that discriminatively injects conditions from multiple modalities, enabling condition decoupling and feature fusion. Furthermore, we design a novel Synthetic-to-Real Training strategy to strengthen visual realism while maintaining identity similarity. To further enhance the preservation of facial dynamics under challenging scenarios, we develop an Identity-Coherence Reward Weighting (IDRW) mechanism, which significantly improves model robustness in complex motions.

Through the aforementioned improvements, our approach enables effective video face swapping across diverse scenarios, as illustrated in Fig. 1. Due to the limited video-face-swapping benchmarks, we introduce ***IDBench-V***, a comprehensive benchmark encompassing a wide spectrum of videos with varying head poses, facial expressions, and lighting conditions. We conduct extensive evaluations on *IDBench-V*, demonstrating that *OmniFace* achieves clear advantages over state-of-the-art methods, both quantitatively and qualitatively. Notably, our proposed framework exhibits exceptional versatility and can be seamlessly adapted to various swap-related tasks. Overall, our contributions are summarized as follows.

- We establish a new, bidirectionally-supervised paradigm for VFS that breaks from the constraints of previous works. This is enabled by our novel *SyncID-Pipe* data pipeline, which effectively transfers the superiority of IFS models to the video domain.
- We propose *OmniFace*, a powerful DiT-based framework featuring a core MC mechanism for effective multi-modal information injection. Furthermore, a Synthetic-to-Real Training and an IDRW strategy ensure high-fidelity results even in challenging scenarios.
- We introduce *IDBench-V*, a comprehensive benchmark for VFS, on which *OmniFace* demonstrates state-of-the-art performance. Moreover, our framework exhibits remarkable versatility, readily adapting to various swap-related tasks. Code and models will be made publicly available.

2 Related Work

Video Foundation Model. The development of diffusion models [11] has significantly advanced video foundation model research. Early latent diffusion methods [2, 9, 13] extended Text-to-Image models with U-Net architectures to the video domain by incorporating temporal modules such as 3D convolutions and temporal attention. Recently, the emergence of Diffusion Transformer (DiT) [31]-based methods for video generation has exhibited superior performance in quality and consistency. These methods [20, 25, 28, 38, 44] employ powerful scaling transformers to generate longer and higher-quality videos. In addition to common Text-to-Video and Image-to-Video models, a growing number of keyframe interpolation models [7] have emerged (e.g., First-Last-Frame models), paving the way for various downstream video generation tasks.

Face Swapping. Early image face swapping primarily focused on GAN-based models, such as FSGAN [29,30], FaceShifter [22], HifiFace [40], and SimSwap [3]. More recently, with the rapid development of diffusion models, several diffusion-based image face swapping models have emerged, including DiffFace [18], DiffSwap [46], FaceAdapter [10], ReFace [1], and DreamID [45], all of which have achieved promising results. Related tasks such as talking-head generation, face animation, and face reenactment have explored motion transfer and temporal consistency [14, 19, 26, 35]. Compared to image face swapping, video face swapping is still in its nascent stages. VividFace [34] models video face swapping as a conditional inpainting task and proposes the first diffusion-based framework. DynamicFace [39] incorporates precise and disentangled facial conditions for flexible and accurate control. HiFiVFS [4] introduces an additional attribute extraction module to capture fine-grained attribute features. CanonSwap [27] proposes a canonical space, performing face swapping within this space before projecting back to the real domain. Stand-In [43] employs its identity-preserving framework on a general Text-to-Video model, guiding it with externally pre-processed landmark-based masks for zero-shot face swapping. While these works have improved the temporal consistency and quality of generated results, their performance in terms of identity similarity and attribute preservation remains

largely unsatisfactory due to a lack of explicit supervision. In contrast, we significantly advance video face swapping by leveraging the superiority of image face swapping to construct explicit supervision.

3 Preliminary

The DiT model employs a transformer as the denoising network to refine the diffusion latent. Our method inherits the video diffusion transformers trained using Flow Matching [23], which conducts the forward process by linearly interpolating between noise and data in a straight line. At the time step t , latent $\mathbf{z}_t$ is defined as: $\mathbf{z}_t = (1 - t)\mathbf{z}_0 + t\epsilon$, where $\mathbf{z}_0$ is the clean video, and $\epsilon \sim \mathcal{N}(0, 1)$ is the Gaussian noise. The model is trained to directly regress the target velocity:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t, \mathbf{z}_0, \epsilon} \left[\left\| (\mathbf{z}_0 - \epsilon) - \mathbf{v}_\theta(\mathbf{z}_t, t, y) \right\|^2 \right], \quad (1)$$

where $\mathbf{v}_\theta$ refers to the diffusion model output and y denotes condition.

4 Methodology

We propose a comprehensive framework to boost Video Face Swapping (VFS) by harnessing the prowess of Image Face Swapping (IFS). We first introduce a novel data curation pipeline *SyncID-Pipe*, which constructs bidirectional ID quadruplets to bridge the gap between VFS and IFS (Sec. 4.1). Building upon paired data, we develop *OmniFace*, a DiT-based video face swapping framework employing a core Modality-Aware Conditioning module to discriminatively inject conditions from multiple modalities (Sec. 4.2). To further enhance visual realism and identity consistency under challenging scenarios, we design a Synthetic-to-Real Training mechanism and an Identity-Coherence Reward Weighting strategy during training (Sec. 4.3). Moreover, our proposed framework exhibits exceptional versatility (Sec. 4.5).

4.1 SyncID-Pipe

IFS has demonstrated better performance in identity and attribute preservation compared to VFS. The fundamental difference between IFS and VFS lies in the dynamic nature of video, which requires consistent preservation of motion and expression across frames. This observation inspires us to explore bridging the gap between image and video domains by supplementing dynamic signals, thereby leveraging the strengths of IFS to significantly boost VFS performance.

Identity-Anchored Video Synthesizer. Building on this insight, we introduce a simple yet effective Identity-Anchored Video Synthesizer (IVS) to generate pair data for explicitly supervised training. As shown in Fig. 2, the IVS is trained to reconstruct a portrait video V_r by leveraging its extracted pose

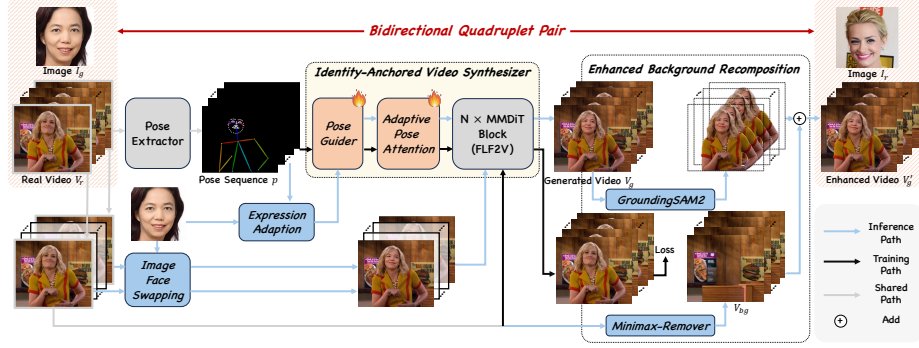


Fig. 2: Overview of SyncID-Pipe. We pre-train the Identity-Anchored Video Synthesizer and combine it with the Image Face Swapping model to construct Bidirectional Quadruplet Pair data.

sequence p . This is achieved by conditioning a First-Last-Frame video foundation model (FLF2V) [7] on the initial and final frames of V_r , along with its pose sequence p . Such reconstruction-based training enables large-scale video pre-training. To minimize modifications to the foundation model and facilitate seamless integration of its pre-trained motion priors with the face swapping framework (detailed in Sec. 4.2), we introduce an Adaptive Pose-Attention mechanism to inject motion information.

Adaptive Pose-Attention. We employ a lightweight Pose Guider composed of several simple convolutional layers to extract pose features and align them with the dimension of the latent feature. To ensure precise spatiotemporal alignment between the pose sequence and the noisy latent video, we reuse the Rotary Position Embedding (RoPE) [36] indices from the noisy latents for the pose condition, which can help maintain the alignment. In each DiT block, we introduce two trainable linear layers $\mathbf{W}'_k$ and $\mathbf{W}'_v$ to align pose features $\mathbf{P}$ with latent feature $\mathbf{Z}$. Formally, the output of the Pose-Attention $\mathbf{Z}_{new}$ is:

$$\mathbf{Z}_{new} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V} + \lambda \cdot \text{Softmax}\left(\frac{\mathbf{Q}(\mathbf{K}')^\top}{\sqrt{d}}\right)\mathbf{V}', \quad (2)$$

where $\mathbf{Q} = \mathbf{Z}\mathbf{W}_q$, $\mathbf{K} = \mathbf{Z}\mathbf{W}_k$, and $\mathbf{V} = \mathbf{Z}\mathbf{W}_v$ are derived from the frozen DiT layers, while $\mathbf{K}' = \mathbf{P}\mathbf{W}'_k$ and $\mathbf{V}' = \mathbf{P}\mathbf{W}'_v$ are from our trainable pose-adaptor layers. The hyperparameter λ controls the strength and flexibility of incorporating pose features. We train IVS by collecting a large-scale portrait dataset and extracting pose sequences. The model is optimized using Flow Matching [23]. Through the aforementioned training, IVS can generate a video consistent with the content of the given keyframes and the actions specified by the pose sequence. This video is subsequently utilized for constructing bidirectional ID quadruplets.

Bidirectional ID Quadruplets Construction. Leveraging the identity preservation and dynamic attribute controllability of our designed IVS model, we ef-

fectively bridge the gap between IFS and VFS. Building upon this capability, we construct bidirectional ID quadruplet training data, anchored by IFS, to enable explicit supervision. As illustrated in Fig. 2, for a source image-video pair (I_r, V_r) with identity ID A and a target image (I_g) with identity ID B, we first utilize a state-of-the-art IFS model [45] to transfer ID B onto the first and last frames of V_r . This process yields high-quality reference frames (I_{ref1}, I_{ref2}) . Subsequently, these reference frames, along with the retargeted pose sequence, are fed into the pre-trained IVS module to synthesize V_g , a video of ID B. The resulting bidirectional ID quadruplet is formatted as $\{I_r, V_r, I_g, V_g\}$, where $\{I_r, V_g, V_r\}$ constitutes the forward-generated paired data, and $\{I_r, V_r, V_g\}$ represents the backward-real paired data. To further enhance the diversity and practicality of the training data, we implement the following strategies:

Source Data Curation. To ensure the robustness of *OmniFace* across diverse and challenging scenarios, we carefully curate target videos encompassing varied makeup styles, extreme lighting conditions, and other adversarial settings. Furthermore, we incorporate talking-head datasets to enhance the ability of the model to preserve subtle facial expressions and accurate lip-synchronization.

Expression Adaptation. Considering that simply using the pose sequence of the target video to drive the generation of the target video leads to identity-expression entanglement and thus causes identity leakage, we use an expression adaptation module to decouple identity and expression, allowing for high-quality expression transfer. During inference, a 3D face reconstruction model [41] extracts identity coefficient from I_g , and the expression and pose coefficients from each V_r frame. We then recombine identity of I_g with expression and pose of each V_r frame to reconstruct a new 3D face model. Projecting this model yields retargeted facial landmarks, which replace the original ones in the pose sequence for final inference.

Enhanced Background Recomposition. Static keyframe-driven IVS often produces videos with background inconsistencies relative to the target videos, particularly with significant background motion. To enhance the applicability of the model in real-world dynamic scenarios, we design an Enhanced Background Recomposition module. As shown in Fig. 2, for forward-generated paired data $\{I_r, V_g, V_r\}$, we first extract foreground masks from V_r and V_g using SAM2 [32]. We then utilize MinimaxRemover [47] to remove the foreground from V_r , yielding a clean background video V_{bg} . Subsequently, the foreground from V_g is pasted back into V_{bg} to form enhanced video V'_g . A feathering operation is then applied at the foreground edges to achieve a smooth and natural blending. Crucially, in the final paired data obtained this way, the supervision video remains the real video V_r . This prevents the model from learning artifacts introduced during the background pasting process. By specifically augmenting these data, we aim to improve the capabilities of the model in background preservation.

4.2 OmniFace Framework

VFS aims to seamlessly transfer identity information from a source face to a target video, while preserving attributes of the target video, such as pose, ex-

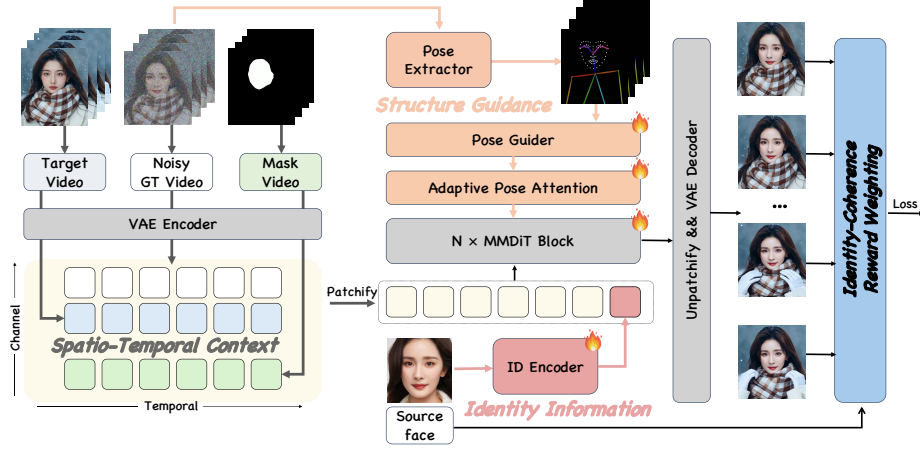


Fig. 3: Overview of *OmniFace* framework. We design customized injection mechanisms for Spatio-Temporal Context, Structural Guidance, and Identity Information, respectively. In Synthetic Training, I_g and V_r serve as the source face and target video, with V_g as the GT video; in Real Augmentation Training, I_r and V'_g serve as the source face and target video, with V_r as the GT video. Target videos with high identity variance are further used for identity-coherence reward weighting.

pression, and background. A core challenge in designing such a model involves effectively injecting and simultaneously disentangling and fusing different types of information, including identity and attributes. To address this, we develop *OmniFace*, a video face swapping framework based on DiT models, as illustrated in Fig. 3. Central to this framework is the **Modality-Aware Conditioning (MC)** mechanism, designed to efficiently inject multiple conditions, tailored to their respective needs. To make conditions disentangled from each other, the MC mechanism decomposes the conditions for video face swapping into three distinct types. The details are as follows:

Spatio-Temporal Context Module. To provide the contextual reference information to be retained (e.g., background, lighting), we inject both the target video and the dilated face mask into the model. Since these conditions must align precisely with the latent noise in both spatial and temporal dimensions, we concatenate them with latent of target video along the channel dimension.

Structural Guidance Module. To achieve fine-grained preservation of dynamic attributes, we incorporate the pose condition as structural guidance. To this end, we employ the Pose-Attention mechanism and initialize corresponding parameters with a pre-trained model described in Sec. 4.1, which fully leverages the prior from the IVS model. The strategy enables efficient structural control while avoiding disruption to the high-level features being processed by the DiT.

Identity Information Module. In contrast to the context and structural information, identity condition represents high-level semantic features and requires comprehensive spatiotemporal interaction. We first employ a dedicated

ID encoder to encode the reference identity into ID embeddings. And then we concatenate them with the latent noise after patchification along the token dimension, which enables full interaction with latent features through the DiT’s inherent attention mechanism.

In summary, the spatio-temporal context module provides required attribute information and leverages the mask to indicate facial regions to the model. The structural guidance module further enhances motion attributes, improving the preservation of exaggerated movements and fine-grained expressions. Finally, the identity information module efficiently injects identity-related features.

4.3 OmniFace Training Pipeline

A significant challenge in training video face swapping models involves balancing identity similarity, attribute preservation, visual realism, and ensuring temporal consistency of identity. To mitigate the issue, we introduce a Synthetic-to-Real Training mechanism along with an Identity-Coherence Reward Weighting strategy. The training process is divided into the following stages:

Synthetic Training. In our constructed bidirectional ID quadruplet data $\{I_r, V_r, I_g, V_g\}$, $\{I_r, V_g, V_r\}$ represents the forward-generated paired data, while $\{I_g, V_r, V_g\}$ represents the backward-real paired data. In this initial stage, we train our model using the forward-generated paired data. Since V_g is synthesized by the IVS module, it remains distributionally consistent with our underlying video foundation model. (as further elaborated in the supplementary material). This alignment significantly accelerates model convergence and enables the attainment of higher identity similarity, yielding superior similarity compared to direct training with backward-real paired data.

Real Augmentation Training. The previous stage yields a VFS model with high identity similarity, however, its realism and background preservation remain limited due to the fact that the forward-generated paired data is model-generated and inherently lacks fidelity in background and realism. Therefore, we introduce a real augmentation training stage to fine-tune the model, which uses backward-real paired data $\{I_g, V_r, V'_g\}$ augmented by the Enhanced Background Recomposition strategy introduced in Sec. 4.1. Experiments show that this stage enables the model to maintain strong identity similarity while achieving excellent background.

Identity-Coherence Reward Weighting. While the aforementioned mechanisms establish a robust baseline, temporal identity consistency remains a challenge in complex scenarios, particularly in videos with significant motions. We observe that identity similarity fluctuates, remaining high in frontal views and mild movements but degrading considerably during profile views or intense actions. To specifically address this final-mile problem, as shown in Fig. 3, we introduce the Identity-Coherence Reward Weighting (IDRW) mechanism, inspired by [6, 24]. The core insight behind IDRW is to incentivize the model to focus its learning capacity on difficult frames. We formalize this as a policy optimization problem, where the model’s generative process is treated as a policy, π_θ , that aims to produce video frames with maximal identity fidelity. Consequently,

frames with low identity fidelity are inherently more valuable as learning signals, as they represent the greatest potential for policy improvement. Instead of learning a complex Q-function via temporal difference updates in some methods [33], we leverage the specific nature of the video face-swapping task to define an efficient and explicit Q-value. Given a state y representing the conditional inputs and an action $\hat{x}_0$ corresponding to the generated video frame, we define the Q-value as:

$$Q(y, \hat{x}_0) = \frac{1}{\cos(E(\hat{x}_0), E(I_t)) + \delta}, \quad (3)$$

where $E(\cdot)$ denotes the feature embedding extracted by [5], $\cos(\cdot, \cdot)$ is the cosine similarity, I_t is the source identity image, and δ is a small constant for numerical stability.

As implemented in our pipeline, we first perform a full sampling pass without backpropagation to generate a video and compute the Q-value for each frame. These frame-wise Q-values are then averaged to obtain Q_c for each VAE-encoded chunk. These Q_c are then used as weights for the flow matching loss [23] during the IDRW training step:

$$\mathcal{L}_{\text{IDRW}}(\theta) = \sum_{c=1}^C \mathbb{E}_{t, \epsilon} \left[Q_c \cdot \|(\mathbf{z}_c - \epsilon) - \mathbf{v}_\theta((1-t)\mathbf{z}_c + t\epsilon, t, y)\|^2 \right], \quad (4)$$

where the loss is summed over all C VAE-encoded chunks in the video. This loss can also be interpreted as a reward-weighted likelihood maximization objective as formally proven in **supplementary material**.

4.4 OmniFace Inference Pipeline

During inference, we utilize Classifier-Free Guidance (CFG) [12] to enhance identity similarity. This guidance vector, $\mathbf{d}$, is defined as the difference between the velocity predictions with and without the source identity condition, $\mathcal{C}_{\text{id}}$:

$$\mathbf{d} = \mathbf{v}_\theta(z_t, \mathcal{C}_{\text{pose}}, \mathcal{C}_{\text{ref}}, \mathcal{C}_{\text{id}}) - \mathbf{v}_\theta(z_t, \mathcal{C}_{\text{pose}}, \mathcal{C}_{\text{ref}}, \emptyset), \quad (5)$$

where $\mathcal{C}_{\text{pose}}$ is the pose sequence, $\mathcal{C}_{\text{ref}}$ is the concatenation of the target video and its mask, and $\emptyset$ denotes the null identity condition. The final velocity is then guided by:

$$\mathbf{v}_{\text{output}} = \mathbf{v}_{\text{cond}} + s \cdot \mathbf{d}, \quad (6)$$

where $\mathbf{v}_{\text{cond}}$ is the fully conditional prediction and s is the guidance scale. This allows for controlling the trade-off between fidelity and quality.

4.5 OmniFace Versatility

Notably, our proposed *OmniFace* extends beyond face swapping task. By constructing explicit paired data for various human-centric swapping tasks—such as outfit, accessory, and hairstyle swapping—simply through replacing the Image Face Swapping (IFS) model with a general-purpose image editing model (e.g., Nano banana [8]) within the SyncID-Pipe, *OmniFace* can be extended to a wider range of swap category tasks. We will detail this extensibility in Sec. 5.5.

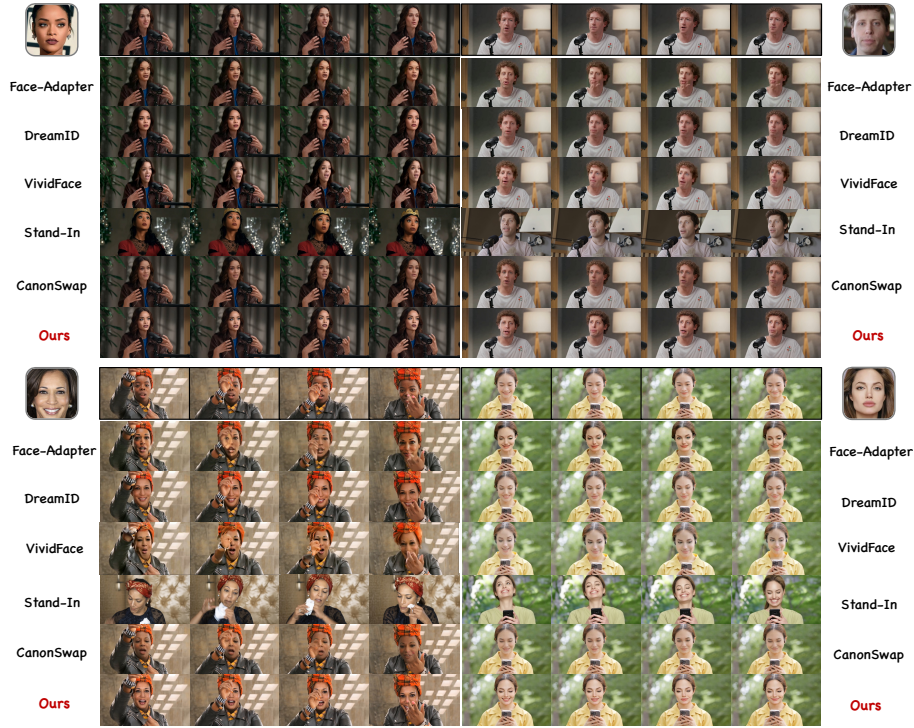


Fig. 4: Qualitative comparisons with state-of-the-art methods. Please zoom in for more details.

5 Experiments

5.1 Setup

IDBench-V. Considering that most previous benchmarks, such as CelebA-HQ [21] and FFHQ [17], are primarily focused on image-level tasks. We introduce *IDBench-V*, a new comprehensive benchmark for video face swapping. Following the scale of recent works [27, 34], our benchmark comprises 200 real-world target video-source image pairs, covering a diverse range of challenging scenarios. These include small faces, extreme head poses, severe occlusions, complex lighting, dynamic expressions, and cluttered multi-person scenes. IDBench-V provides a rigorous and holistic platform for evaluation in real-world usage scenarios.

Implementation Details. We use the AdamW optimizer with a constant learning rate of 1.0×10^{-5} for all training stages. The training process for OmniFace commences with 50k iterations with a global batch size of 16 on exclusively synthetic ground truth (GT) data V_g . Subsequently, the model is fine-tuned for an additional 80k iterations with a global batch size of 32 on a hybrid dataset of real GT V_r and synthetic GT V_g . The training regimen culminates in the application of our IDRW, where the model is further refined on a curated subset

Table 1: Quantitative comparisons with baseline methods.

Method	Identity Consistency				Attribute Preservation				Video Quality	
	ID-Arc $\uparrow$	ID-Ins $\uparrow$	ID-Cur $\uparrow$	Variance $\downarrow$	Pose $\downarrow$	Expression $\downarrow$	Background $\uparrow$	Subject $\uparrow$	VFID $\downarrow$	Smoothness $\uparrow$
REFace [WACV'25] [1]	0.472	0.471	0.474	0.0191	5.102	2.785	0.909	0.913	7.084	0.988
Face-Adapter [ECCV'24] [10]	0.440	0.496	0.450	0.0081	5.156	3.037	0.942	0.945	3.460	0.988
DreamID [SIGGRAPH Asia'25] [45]	0.616	0.702	0.664	0.0058	3.013	2.930	0.940	0.951	3.108	0.989
VividFace [NeurIPS'25] [34]	0.412	0.438	0.422	0.0046	6.744	3.136	0.935	0.947	3.125	0.984
Stand-In [CVPR'26] [43]	0.403	0.403	0.367	0.0057	19.819	2.995	0.931	0.951	3.368	0.982
CanonSwap [ICCV'25] [27]	0.397	0.431	0.407	<u>0.0030</u>	2.430	2.477	0.950	0.954	2.176	0.991
Ours	0.659	0.713	0.688	0.0029	<u>2.446</u>	2.430	0.951	0.966	<u>2.243</u>	0.992

of data from the previous stages—specifically, samples exhibiting high variance in ID similarity. All training was performed on NVIDIA A100 GPUs. Refer to supplementary material for more details.

Baselines. We conduct comparisons against existing face swapping SOTA models on *IDBench-V*. For image face swapping, we compare with REFace [1], Face-Adapter [10], and DreamID [45], noting that these image-based methods achieve video face swapping by processing frames individually. For video face swapping, we evaluate against VividFace [34], Stand-In [43] and CanonSwap [27].

Evaluation Metrics We evaluate the performance of various video face swapping methods across three key dimensions: Identity Consistency, Attribute Preservation, and Video Quality. For Identity Consistency, we employ ArcFace [5], InsightFace [37], and CurricularFace [15] to compute ID similarity. To quantify temporal stability, we additionally calculate the variance of these frame-wise similarities. Attribute Preservation is assessed by evaluating the fidelity of pose and expression transferred from the target video. (details are provided in the supplementary material) Furthermore, we incorporate three metrics from VBench [16]: background consistency, subject consistency, and motion smoothness, which respectively evaluate the consistency of the background, the primary subject, and the overall motion. Finally, for Video Quality, we evaluate perceptual video quality in unpaired scenarios using the VFID with a ResNext [42] feature extractor.

5.2 Quantitative Comparisons

Metric Evaluation. As shown in Tab. 1, *OmniFace* comprehensively outperforms state-of-the-art models in terms of identity similarity metrics. Regarding attribute preservation, *OmniFace* is optimal across almost all metrics, with only a slight inferiority to CanonSwap in terms of pose. It is worth noting that CanonSwap, due to its very low identity similarity, results in minimal alteration to the original video, thereby exhibiting good attribute preservation and video quality. Our proximity to CanonSwap in attribute preservation demonstrates our model’s excellent capability in this regard, while its identity similarity is significantly higher than CanonSwap. This superiority is attributed to our ID quadruplet effectively transferring the high identity similarity from DreamID to VFS. Interestingly, owing to the effectiveness of our training strategy, our identity similarity even slightly surpasses that of DreamID. In terms of video quality, our model also achieves outstanding results, showing substantial improvements compared to IFS models such as REFace, Face-Adapter, and DreamID. This col-

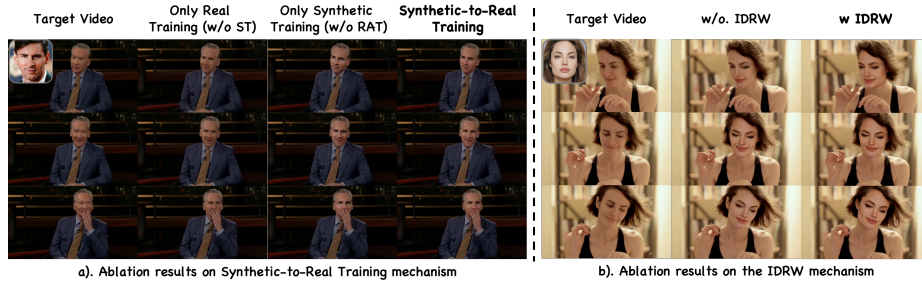


Fig. 5: Ablation studies of *OmniFace*.

lectively demonstrates that our model not only achieves high identity similarity and robust attribute preservation but also generates high-quality videos.

User Study. We invited 19 volunteers to conduct a human evaluation of the models on IDBench. Each sample was rated across three dimensions: Identity Similarity, Attribute Preservation, and Video Quality, with scores ranging from 1-5. As presented in Tab. 2, show that our model achieved the best performance across all metrics, thereby demonstrating the superior capabilities of our model.

Table 2: User study of *OmniFace*.

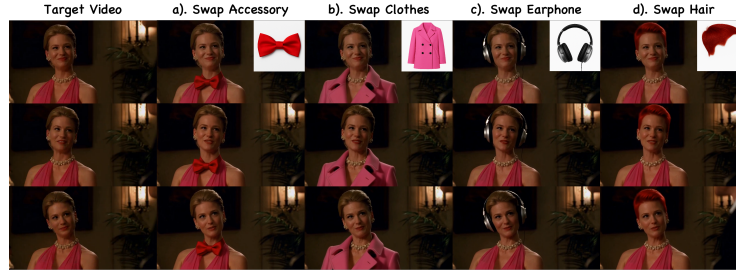
Method	ID Sim $\uparrow$	Attr $\uparrow$	Quality $\uparrow$
REFace [1]	1.45	2.15	1.11
Face-Adapter [10]	2.17	2.93	1.14
DreamID [45]	<u>3.78</u>	3.89	3.06
VividFace [34]	2.32	3.17	2.85
Stand-In [43]	2.45	1.60	2.91
Canonswap [27]	1.99	<u>3.91</u>	<u>3.42</u>
Ours	3.85	4.22	4.15

5.3 Qualitative Analysis

We conduct a qualitative comparison with VFS methods VividFace, Stand-In and Canonswap as well as IFS methods Face-Adapter and DreamID. As illustrated in Fig. 4, *OmniFace* demonstrates excellent performance across identity similarity, expression preservation, background preservation, and occlusion. In the first two cases presented in the first row, our method demonstrates superiority in identity similarity compared to other approaches, consistently across both male and female subjects. Specifically, our identity similarity is significantly better than that of Face-Adapter, VividFace, Stand-In, and CanonSwap. Perceptually, the identity similarity is close to DreamID, however, we observe that our IVS module, by incorporating dynamic expression information, leads to superior expression performance compared to DreamID. Stand-In, which utilizes an inpainting approach for face swapping, introduces substantial alterations to the original video. In the second row, the left case highlights the robust performance of our model under occlusion, outperforming all other models. The right case showcases the superiority of our model in handling complex expressions.

Table 3: Ablation study of *OmniFace*.

Method	ID-Arc $\uparrow$	Variance $\downarrow$	Pose $\downarrow$	Expression $\downarrow$	VFID $\downarrow$
a) w/o Quadruplet	0.510	0.0036	2.468	2.432	2.242
b) w/o ST	0.604	0.0035	2.742	2.445	2.145
c) w/o RAT	0.657	0.0042	2.557	2.443	3.845
d) w/o IDRW	0.631	0.0041	2.687	2.488	2.206
e) Ours	0.659	0.0029	2.446	2.430	2.243

**Fig. 6: Versatility** of our methods. Please zoom in for more details.

5.4 Ablation Studies

To demonstrate the effectiveness of our proposed method, we conduct the following ablation studies: a). Directly training using the self-reconstruction inpainting-based approach (w/o Quadruplet); b). Training exclusively with backward-real paired data (w/o ST); c). Training solely with forward-generated paired data (w/o RAT); d). Training without the Identity-Coherence Reward Weighting stage (w/o IDRW). As shown in Tab. 3 a), Traditional inpainting-based method yields significantly lower identity similarity, demonstrating the effectiveness of *SyncID-Pipe* to construct explicitly supervised data to bridge the gap between VFS and IFS. b). As demonstrated in Fig. 5 a), the w/o ST setting yields higher realism but lower identity similarity (i.e., a good VFID score but poor ID-Arc). Conversely, w/o RAT achieves superior identity similarity but at the cost of realism (i.e., a good ID-Arc score but poor VFID). In contrast, our Synthetic-to-Real Training strategy (line w/o IDRW) ultimately strikes a good balance, maintaining high identity similarity while preserving realism. Furthermore, the IDRW mechanism significantly enhances identity similarity under complex motions, leading to a noticeable improvement. A comparison between lines (d) and (e) reveals that IDRW not only improves ID-Arc to some extent but also substantially reduces variance, which represents the consistency of inter-frame similarity. As depicted in Fig. 5 b), the top and bottom frames present profile views, while the middle frame shows a frontal view. Without IDRW, the model performs well on frontal views but poorly on profile views. However, IDRW substantially boosts identity similarity in profile views, as exemplified by the topmost frame.

5.5 Versatility

As shown in Fig. 6, by expanding training data, our model can be extended to a wider range of human-centric swapping tasks, including accessory, outfit, headphone, and hairstyle swapping.

6 Ethical Considerations

High-fidelity video face swapping can support benign applications such as film production, creative design, privacy-preserving anonymization, and consented identity editing. However, it can also be misused to create non-consensual deep-fakes, impersonate individuals, spread misinformation, or harass targeted persons. We therefore emphasize that *OmniFace* should be used only with consent from the depicted identities and with clear disclosure of generated content. We will release code and models only for research use under a responsible-use license, and encourage safeguards such as provenance tracking, watermarking, and deepfake detection.

7 Conclusion

This work presents a comprehensive framework for Video Face Swapping (VFS). Our *SyncID-Pipe* data pipeline effectively transfers the superiority of image face swapping to video, enabling *OmniFace*—a DiT-based model for VFS—to achieve superior performance on our proposed comprehensive benchmark *IDBench-V*. The method demonstrates strong versatility and provides a systematic solution for high-fidelity VFS.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Project U23A20281.

References

1. Baliah, S., Lin, Q., Liao, S., Liang, X., Khan, M.H.: Realistic and efficient face swapping: A unified approach with diffusion models. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1062–1071. IEEE (2025)
2. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
3. Chen, R., Chen, X., Ni, B., Ge, Y.: Simswap: An efficient framework for high fidelity face swapping. In: MM ’20: The 28th ACM International Conference on Multimedia (2020)

4. Chen, X., He, K., Zhu, J., Ge, Y., Li, W., Wang, C.: Hifivfs: High fidelity video face swapping. arXiv preprint arXiv:2411.18293 (2024)
5. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019)
6. Ding, S., Hu, K., Zhang, Z., Ren, K., Zhang, W., Yu, J., Wang, J., Shi, Y.: Diffusion-based reinforcement learning via q-weighted variational policy optimization. Advances in Neural Information Processing Systems **37**, 53945–53968 (2024)
7. Gao, Y., Guo, H., Hoang, T., Huang, W., Jiang, L., Kong, F., Li, H., Li, J., Li, L., Li, X., et al.: Seedance 1.0: Exploring the boundaries of video generation models. arXiv preprint arXiv:2506.09113 (2025)
8. Google: Gemini 2.5 flash image. <https://aistudio.google.com/models/gemini-2-5-flash-image> (2025), accessed: 2026-06-29
9. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725 (2023)
10. Han, Y., Zhu, J., He, K., Chen, X., Ge, Y., Li, W., Li, X., Zhang, J., Wang, C., Liu, Y.: Face-adapter for pre-trained diffusion models with fine-grained id and attribute control. In: European conference on computer vision. pp. 20–36. Springer (2024)
11. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
12. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
13. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. Advances in neural information processing systems **35**, 8633–8646 (2022)
14. Hong, F.T., Zhang, L., Shen, L., Xu, D.: Depth-aware generative adversarial network for talking head video generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3397–3406 (2022)
15. Huang, Y., Wang, Y., Tai, Y., Liu, X., Shen, P., Li, S., Li, J., Huang, F.: Curricular-face: adaptive curriculum learning loss for deep face recognition. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5901–5910 (2020)
16. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21807–21818 (2024)
17. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
18. Kim, K., Kim, Y., Cho, S., Seo, J., Nam, J., Lee, K., Kim, S., Lee, K.: Diff-face: Diffusion-based face swapping with facial guidance. Pattern Recognition **163**, 111451 (2025)
19. Kligvasser, I., Cohen, R., Leifman, G., Rivlin, E., Elad, M.: Anchored diffusion for video face reenactment. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 4087–4097. IEEE (2025)
20. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)

21. Lee, C.H., Liu, Z., Wu, L., Luo, P.: Maskgan: Towards diverse and interactive facial image manipulation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5549–5558 (2020)
22. Li, L., Bao, J., Yang, H., Chen, D., Wen, F.: Advancing high fidelity identity swapping for forgery detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5074–5083 (2020)
23. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
24. Liu, J., Liu, G., Liang, J., Li, Y., Liu, J., Wang, X., Wan, P., Zhang, D., Ouyang, W.: Flow-grpo: Training flow matching models via online rl. arXiv preprint arXiv:2505.05470 (2025)
25. Liu, Y., Zhang, K., Li, Y., Yan, Z., Gao, C., Chen, R., Yuan, Z., Huang, Y., Sun, H., Gao, J., et al.: Sora: A review on background, technology, limitations, and opportunities of large vision models. arXiv preprint arXiv:2402.17177 (2024)
26. Liu, Z., Lu, J., Lu, R., Liang, C., Wang, S.: Consisttalk: Intensity controllable temporally consistent talking head generation with diffusion noise search. Proceedings of the AAAI Conference on Artificial Intelligence **40**(9), 7449–7457 (2026)
27. Luo, X., Zhu, Y., Liu, Y., Lin, L., Wan, C., Cai, Z., Li, Y., Huang, S.L.: Canonswap: High-fidelity and consistent video face swapping via canonical space modulation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10064–10074 (2025)
28. Ma, G., Huang, H., Yan, K., Chen, L., Duan, N., Yin, S., Wan, C., Ming, R., Song, X., Chen, X., et al.: Step-video-t2v technical report: The practice, challenges, and future of video foundation model. arXiv preprint arXiv:2502.10248 (2025)
29. Nirkin, Y., Keller, Y., Hassner, T.: FSGAN: Subject agnostic face swapping and reenactment. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 7184–7193 (2019)
30. Nirkin, Y., Keller, Y., Hassner, T.: Fsganv2: Improved subject agnostic face swapping and reenactment. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(1), 560–575 (2023)
31. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
32. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollár, P., Feichtenhofer, C.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
33. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
34. Shao, H., Wang, S., Zhou, Y., Song, G., He, D., Qin, S., Zong, Z., Ma, B., Liu, Y., Li, H.: Vividface: A diffusion-based hybrid framework for high-fidelity video face swapping. arXiv preprint arXiv:2412.11279 (2024)
35. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation. Advances in neural information processing systems **32** (2019)
36. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)
37. The InsightFace Team: Insightface products. <https://github.com/deepinsight/insightface>, accessed: 2026-06-29
38. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)

39. Wang, R., Chen, Y., Xu, S., He, T., Zhu, W., Song, D., Chen, N., Tang, X., Hu, Y.: Dynamicface: High-quality and consistent face swapping for image and video using composable 3d facial priors. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13438–13447 (2025)
40. Wang, Y., Chen, X., Zhu, J., Chu, W., Tai, Y., Wang, C., Li, J., Wu, Y., Huang, F., Ji, R.: Hiface: 3d shape and semantic prior guided high fidelity face swapping. arXiv preprint arXiv:2106.09965 (2021)
41. Wang, Z., Zhu, X., Zhang, T., Wang, B., Lei, Z.: 3d face reconstruction with the geometric guidance of facial part segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1672–1682 (2024)
42. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. arXiv preprint arXiv:1611.05431 (2016)
43. Xue, B., Duan, Z.P., Yan, Q., Wang, W., Liu, H., Guo, C.L., Li, C., Li, C., Lyu, J.: Stand-in: A lightweight and plug-and-play identity control for video generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 23314–23324 (2026)
44. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
45. Ye, F., Hua, M., Zhang, P., Li, X., Sun, Q., Zhao, S., He, Q., Wu, X.: Dreamid: High-fidelity and fast diffusion-based face swapping via triplet id group learning. arXiv preprint arXiv:2504.14509 (2025)
46. Zhao, W., Rao, Y., Shi, W., Liu, Z., Zhou, J., Lu, J.: 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. pp. 8568–8577 (2023)
47. Zi, B., Peng, W., Qi, X., Wang, J., Zhao, S., Xiao, R., Wong, K.F.: Minimax-remover: Taming bad noise helps video object removal. arXiv preprint arXiv:2505.24873 (2025)