FROM SHORT CLIPS TO REAL-TIME STREAMS: IDENTITY-AWARE SUPERVISION AND CONTEXT-EFFICIENT GENERATION FOR VIDEO FACE SWAPPING

Anonymous authors
Paper under double-blind review

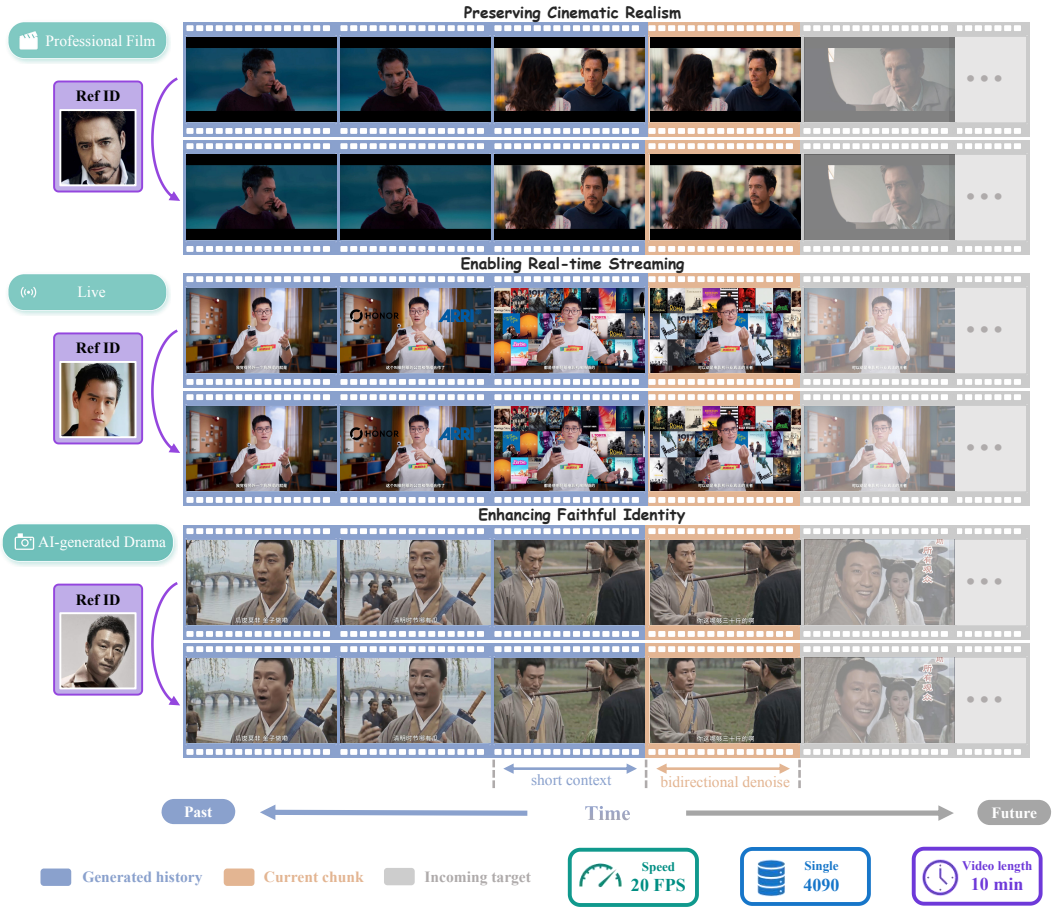

Figure 1: **Real-time streaming video face swapping with STREAMINGSWAP.** Examples span film, live, and AI-generated drama, with reference videos above and swapped outputs below.

ABSTRACT

Video face swapping transfers the reference identity to a target video while preserving its other attributes and motion. Despite recent progress in reference-guided flow-based methods, jointly achieving faithful identity transfer, high fidelity, and real-time generation over long video streams remains challenging. To address this challenge, we rethink the task-specific structure of reference-guided video face swapping. First, face swapping changes only identity while preserving most visual content, yet flow-matching supervision is weakly sensitive to these changes. We therefore develop *Joint Rec-ID Loss*, which balances explicit identity supervision and latent reconstruction according to the noise level. Second, the target video already provides non-edited attributes and temporal dynamics, enabling

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

generation with fewer spatial tokens and shorter temporal contexts through *Face-RoI Generation* and *Short-Context Rollout Training*. By integrating these designs, we propose **StreamingSwap**, a reference-guided flow-matching framework for real-time streaming video face swapping. We also introduce **HumanFaceVid**, comprising 1,000 hours of cinematic training data and evaluation sets of 200 short and 200 long videos. Experiments demonstrate that our method achieves competitive identity transfer and visual quality at 20 FPS on a single NVIDIA RTX 4090 GPU, with stable generation over videos up to 10 minutes long.

1 Introduction

Video face swapping transfers identity from a source image to a target video while preserving other target attributes, such as expression, pose, gaze, illumination, and background. It has broad applications in film production, AI-generated dramas, and live streaming. GAN-based approaches (Chen et al., 2020; Shiohara et al., 2023; Luo et al., 2025; Henry, 2025) offer strong identity transfer and efficient inference, but processing frames independently can introduce visual artifacts and temporal flicker. To improve realism and temporal coherence, diffusion-based approaches leverage the generative priors of pretrained video diffusion models. Among these, inpainting-based approaches (Han et al., 2024; Wang et al., 2025; Chen et al., 2024; Shao et al., 2024) mask the target face and reconstruct it conditioned on source identity and auxiliary attributes, but removing direct access to fine-grained target cues makes non-edited attributes difficult to preserve faithfully.

Recent reference-guided methods, represented by LivingSwap (Luo et al., 2026) and DreamID-V (Guo et al., 2026), construct paired supervision to retain the complete target video as a visual reference, enabling more faithful preservation of fine-grained target attributes. However, the unmasked target face contains not only the attributes to be preserved but also the identity to be replaced, creating a competing identity signal that makes source-identity injection difficult. Meanwhile, these methods are fine-tuned from large-scale video diffusion models (Wan et al., 2025) and remain computationally expensive due to iterative full-frame denoising over long temporal contexts. Consequently, extending their high-fidelity results to real-time, long-duration generation while ensuring faithful source-identity transfer remains difficult.

These challenges motivate us to rethink reference-guided video face swapping through its task-specific structure. First, video face swapping changes only identity while preserving most visual content in the target video. This leaves flow-matching supervision with a weak signal for identity transfer, as identity differences account for only a small fraction of the latent reconstruction error. Second, the target video directly provides the non-edited attributes and temporal dynamics to be preserved. Yet existing reference-guided methods still process background over long temporal contexts, resulting in substantial computational redundancy.

Guided by these observations, we introduce **StreamingSwap**, the first reference-guided flow-matching framework for real-time, long-duration video face swapping. To strengthen source-identity transfer, *Joint Rec-ID Loss* combines clean-latent reconstruction with explicit identity supervision, balancing their relative contributions according to the noise level. For efficient generation, *Face-RoI Generation* restricts video conditioning and generation to face-region tokens, avoiding redundant background computation. To enable streaming, *Short-Context Rollout Training* trains the bidirectional model through autoregressive rollout with a short generated history, reducing the mismatch between training and inference while retaining bidirectional generation within each window. We further introduce **HumanFaceVid**, a cinematic video face-swapping dataset constructed from OpenHumanVid (Li et al., 2025), comprising 1,000 hours of training data and evaluation sets of 200 short and 200 long videos. Across diverse scenarios, **StreamingSwap** improves identity similarity while preserving target attributes and maintaining high video quality. It processes videos at 20 FPS on a single NVIDIA 4090 GPU and maintains stable generation for videos up to 10 minutes.

Our contributions are summarized as follows:

- We identify and examine two task-specific properties of reference-guided video face swapping: latent flow-matching supervision is weakly sensitive to identity changes, while target-provided visual information allows generation with fewer spatial tokens and shorter temporal contexts.

- We introduce **STREAMINGSWAP**, which combines *Joint Rec-ID Loss*, *Face-RoI Generation*, and *Short-Context Rollout Training* to enable real-time streaming video face swapping. It achieves 20 FPS on a single NVIDIA 4090 GPU with stable generation on videos up to 10 minutes long.
- We construct **HumanFaceVid**, comprising 1,000 hours of cinematic training data derived from OpenHumanVid and evaluation sets of 200 short and 200 long videos for assessing identity fidelity, target-attribute preservation, visual quality, and temporal stability.

2 PRELIMINARIES

2.1 FLOW MATCHING

Flow matching (FM) (Lipman et al., 2022) learns a continuous velocity field that transports samples from a Gaussian noise distribution to a target distribution defined in either data or latent space. Let x_1 and x_0 denote a clean video latent and random Gaussian noise in the same latent space, respectively. The intermediate latent along the linear probability path is defined as

$$x_t = (1 - \sigma_t)x_1 + \sigma_t x_0, \tag{1}$$

where $\sigma_t = 1 - t$ denotes the noise level at a continuous timestep $t \in [0, 1]$. Thus, $t = 0$ ($\sigma_t = 1$) corresponds to pure noise and $t = 1$ ($\sigma_t = 0$) corresponds to clean latent. Given a trainable velocity model v_θ, the flow matching objective is defined as

$$\mathcal{L}_{\text{flow}} = \mathbb{E}_{(x_1,c),x_0\sim\mathcal{N}(0,I),t\sim\mathcal{U}(0,1)} \left[\|v_\theta(x_t, c, t) - v_t^\star\|_2^2 \right], \tag{2}$$

where (x_1, c) is a training pair of a clean video latent and its conditioning signal, sampled from the training dataset. $v_t^\star = \frac{\mathrm{d}x_t}{\mathrm{d}t} = x_1 - x_0$ is the target velocity. $\mathcal{N}(0, I)$ and $\mathcal{U}(0, 1)$ denote the standard Gaussian distribution and the uniform distribution over $[0, 1]$, respectively.

2.2 REFERENCE-GUIDED VIDEO FACE SWAPPING

Reference-guided video face swapping conditions generation on the target video to preserve its expression, pose, illumination, motion, and background.

Model conditioning. Reference-guided models incorporate source-image and target-video representations as complementary conditions in the denoising network. A representative design is DreamID-V (Guo et al., 2026), which builds on a Wan Transformer (Wan et al., 2025) and injects video and source-image conditions through channel-wise and token-wise concatenation, respectively. Formally, at each time step t, its velocity predictor $v_\theta(x_t, c, t)$ takes the noisy latent x_t together with conditioning information $c = \mathcal{C}(I, V)$ derived from the source image I and target video V. To construct these conditioning inputs, a frozen VAE encoder $\mathcal{E}_\phi$ first encodes the target video and its face mask M into x_v and x_m, respectively. Then, the noisy latent x_t is concatenated with these clean video conditions and embedded into video tokens:

$$H_{\text{vid}} = \mathcal{P}_{3\text{D}}(\text{Concat}_{\text{ch}}[x_t, x_v, x_m]) \in \mathbb{R}^{(hwF)\times d}, \tag{3}$$

where $\mathcal{P}_{3\text{D}}$ denotes the 3D patch embedding layer, h, w, and F are the height, width, and temporal length of the resulting token grid, and d is the dimension of token feature. $\text{Concat}_{\text{ch}}$ denotes channel-wise concatenation. The grid is flattened into a sequence, with each token retaining its corresponding spatial and temporal position. The source image is encoded by the same VAE and a separate 2D patch embedding $\mathcal{P}_{2\text{D}}$, yielding identity tokens that are concatenated with the video tokens:

$$H_{\text{id}} = \mathcal{P}_{2\text{D}}(\mathcal{E}_\phi(I)), \qquad H = \text{Concat}_{\text{seq}}[H_{\text{id}}, H_{\text{vid}}], \tag{4}$$

where $\text{Concat}_{\text{seq}}$ denotes sequence-wise concatenation. The Transformer then processes H through bidirectional self-attention to predict the final output of the velocity predictor $v_\theta(x_t, c, t)$.

Paired data and reversing supervision. Since naturally paired face-swapping videos are difficult to obtain, LivingSwap (Luo et al., 2026) and DreamID-V (Guo et al., 2026) construct paired data using pretrained face-swapping pipelines. Following the notation of DreamID-V, let (I_r, V_r) denote a real image–video pair with identity r, and let I_g depict another identity g. A face-swapping pipeline $\mathcal{S}$ produces

$$V_g = \mathcal{S}(I_g, V_r), \tag{5}$$

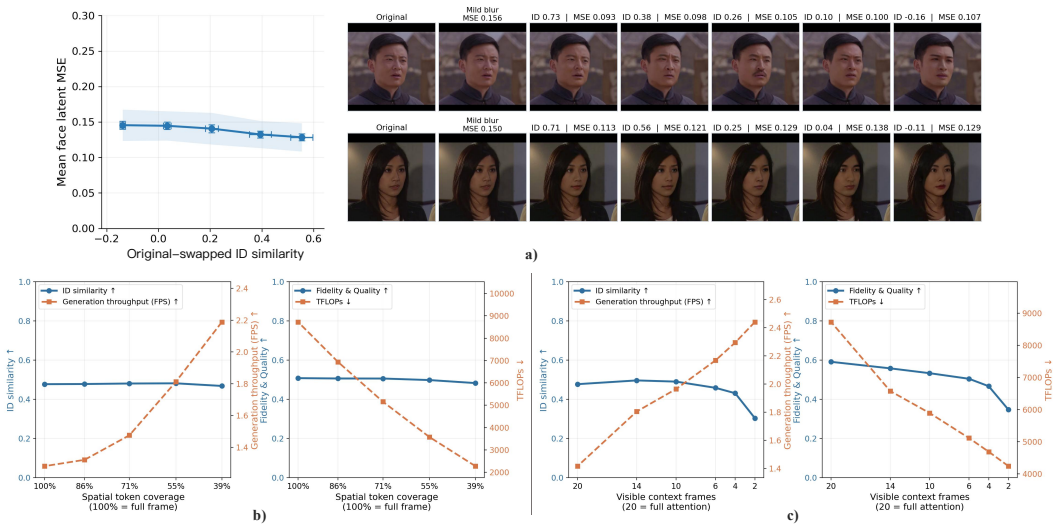

Figure 2: Task-specific analysis of reference-guided video face swapping. (a) Latent flow-matching supervision is weakly sensitive to identity changes. (b) Face-region processing and (c) short temporal context preserve most generation quality while substantially reducing computation.

where V_g carries identity g while approximately preserving the non-identity attributes and temporal dynamics of V_r. The resulting quadruplet (I_r, V_r, I_g, V_g) supports role-reversed training: I_r provides the source identity, V_g serves as the target reference video, and V_r provides ground-truth supervision. This reversal avoids using synthesis artifacts as ground-truth targets. Training follows Eq. 2 with $x_1 = \mathcal{E}_\phi(V_r)$ and $c = c_{\text{swap}} = \mathcal{C}(I_r, V_g)$, with x_t constructed from x_1 using Eq. 1.

3 Task-Specific Analysis

The difficulty of combining faithful identity transfer, high visual quality, and real-time long-video generation motivates us to revisit video face swapping through its task-specific structure. Compared with general-purpose video editing, reference-guided face swapping has two distinctive properties.

First, face swapping has a restricted editing scope: only facial identity should change, while all other facial attributes and content outside the face region should remain consistent with the target video. However, the flow matching objective optimizes a reconstruction loss over the latent representation of the entire image, which may be insufficiently sensitive to identity changes when most visual content remains unchanged. To quantify this potential mismatch, we investigate *how strongly the flow-matching objective penalizes identity-transfer failures* **(Q1)**.

Second, the desired output is densely specified by visual inputs: the source image provides the identity, while the target video supplies not only non-identity facial attributes within the edited region but also the surrounding content and temporal dynamics. Nevertheless, existing reference-guided methods generate full frames, including regions that require no editing, and attend over extended temporal contexts to maintain consistency. To examine this potential computational redundancy, we investigate *how much spatial and temporal context is actually required to preserve fidelity* **(Q2)**.

We investigate these questions through two controlled analysis experiments.

3.1 Identity Sensitivity of Flow-Matching Supervision

Setup. Under the linear probability path in Eq. 1, the predicted velocity determines the clean-latent estimate

$$\hat{x}_1 = x_t + \sigma_t v_\theta(x_t, c_{\text{swap}}, t). \tag{6}$$

Because $x_1 = x_t + \sigma_t v_t^\star$, for $\sigma_t > 0$,

$$\|v_\theta(x_t, c_{\text{swap}}, t) - v_t^\star\|_2^2 = \frac{1}{\sigma_t^2} \|\hat{x}_1 - x_1\|_2^2 . \tag{7}$$

Thus, at fixed noise, flow matching is equivalent to a scaled clean-latent reconstruction objective. Under role-reversed training, a model that ignores I_r and reconstructs the target video V_g is penalized according to the latent distance between V_g and the ground-truth video V_r. This latent distance therefore provides a diagnostic of the flow-matching supervision sensitivity to identity differences. We randomly sample 64 real videos V_r. For each video, we select five identity images $\{I_g^{(k)}\}_{k=1}^{5}$ with varying identity similarities to I_r, including I_r as the self-swap control. Following the paired-data construction used for training, we use InSwapper (Henry, 2025) to generate $V_g^{(k)} = \mathcal{S}(I_g^{(k)}, V_r)$. We encode V_r and each $V_g^{(k)}$ using $\mathcal{E}_\phi$ to compute their face-region latent MSE following Eq. 7, and measure identity similarity between the two videos. To contextualize the MSE scale, the qualitative visualization additionally includes a mildly blurred identity-preserving version of V_r as a reference.

Results. As shown in Fig. 2(a), identity similarity changes by more than 0.75 across the constructed pairs, whereas mean face-region latent MSE varies by less than 0.05. The identity-preserving self-swap (third column) and mild-blur controls (second column) contextualize this error scale: visually minor deviations already produce latent MSEs around 0.15, comparable to or larger than those caused by visible identity changes (last column). This mismatch is consistent with the observation that swapping identity leaves most facial content, including expression, pose, and illumination, largely unchanged. At a fixed noise level, the corresponding flow-matching penalty therefore changes little as the identity discrepancy increases. These results suggest that flow-matching supervision can support identity transfer but provides only a weak signal for identity changes.

3.2 Spatial and Temporal Context Requirements

Setup. We examine the spatial and temporal context required by DreamID-V (Guo et al., 2026) using the same 64 evaluation cases. We independently vary the spatial coverage and temporal attention range of the video tokens $H_{\text{vid}} \in \mathbb{R}^{(hwF)\times d}$ defined in Eq. 3.

Spatially, we restrict token selection along the $h \times w$ spatial grid while retaining all F temporal frames. We progressively narrow the retained spatial region from the full frame to an enlarged face editing region. For the retained video-token indices $\mathcal{J}_\rho$, the selected token sequence and its spatial coverage are given by

$$H_{\text{vid}}^{(\rho)} = H_{\text{vid}}[\mathcal{J}_\rho], \qquad \rho = \frac{|\mathcal{J}_\rho|}{hwF}, \tag{8}$$

where $\rho = 100\%$ corresponds to full-frame processing. Temporally, we use an attention mask to restrict each video token's temporal context to K neighboring latent frames. We vary K over $\{20, 14, 10, 6, 4, 2\}$, with $K = 20 = F - 1$ corresponding to full temporal bidirectional attention.

We report identity similarity, expression and lighting errors, imaging quality, generation throughput in FPS, and computational cost in TFLOPs. For visualization, Fig. 2 aggregates expression error, lighting error, and imaging quality into a normalized fidelity-and-quality score.

Results. Spatially, Fig. 2(b) shows that reducing the retained-token ratio from 100% to 39% largely preserves identity similarity and the fidelity-and-quality score, while increasing generation throughput from approximately 1.3 to 2.2 FPS and reducing computation from $8{,}900$ to $2{,}200$ TFLOPs. This result indicates that background tokens contribute substantial computation but limited information once the target face region is available in reference-guided methods.

Temporally, Fig. 2(c) shows that reducing the visible context from full attention to six neighboring latent frames causes only moderate quality degradation while increasing generation throughput and reducing computational cost. Performance drops sharply with fewer neighboring frames, particularly at $K = 2$. These results suggest that a short temporal context is sufficient for reference-guided generation, as the target video already provides the content and motion cues needed for long-range consistency. The temporal context therefore serves primarily to maintain local continuity.

Summary. Together, the two experiments motivate the following design principles:

- **Explicit identity supervision.** Training should directly supervise source identity to complement the weak sensitivity of latent reconstruction to identity changes.
- **Focused spatial and temporal computation.** Generation should exploit the target-provided content and motion to concentrate computation on face-region tokens and a short temporal context.

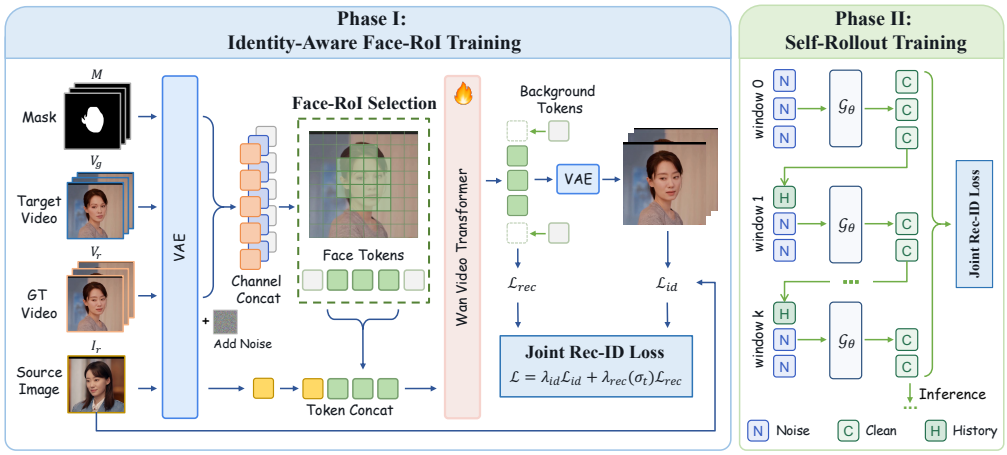

Figure 3: Overview of **StreamingSwap**. Phase I restricts video processing to face-RoI tokens and balances latent reconstruction with identity supervision through noise-dependent weighting. Phase II fine-tunes the model through self-rollout with short generated histories, enabling streaming inference while retaining bidirectional generation within each chunk.

4 Method

4.1 Overview

Our pipeline comprises data preparation, two training phases, and streaming inference. We first construct the reversed pairs as in App. B. Phase I adapts a reference-guided video flow model with *Joint Rec-ID Loss* and *Face-RoI Generation*. Phase II employs *Short-Context Rollout Training*, exposing the model to its own generated short history for autoregressive long-video inference (Fig. 3).

4.2 Phase I: Identity-Aware Face-RoI Training

We adopt the architecture and reference-guided training setup described in Sec. 2.2, initializing the model with pretrained DreamID-V weights (Guo et al., 2026). Guided by our task-specific analysis, we introduce two modifications.

Face-RoI Generation. The analysis in Sec. 3.2 suggests that processing background tokens incurs substantial computational cost while providing limited performance gains. We therefore restrict spatial token coverage to an enlarged face region. Specifically, we downsample the binary face masks M to the latent grid using max pooling and spatially dilate the resulting masks to define the compute region. We select all patches intersecting this region, denoting their token indices by $\mathcal{J}_{\text{roi}}$ and the latent positions they cover by Ω_{roi}. As illustrated in Fig. 3 (left), the Transformer input is

$$H_{\text{roi}} = \text{Concat}_{\text{seq}}\left[H_{\text{id}}, H_{\text{vid}}[\mathcal{J}_{\text{roi}}]\right]. \tag{9}$$

Selected video tokens retain their original spatiotemporal positional encodings. To support batched training, we pad sequences to the maximum length within each batch and mask out padding tokens in attention. We recover the clean RoI latent estimate $\hat{x}_{1,\text{roi}}$ from the predicted velocity following Eq. 6. This estimate is scattered back to the full latent grid and blended with the target-video latent $x_g = \mathcal{E}_\phi(V_g)$:

$$\tilde{x}_1 = M_{\text{blend}} \odot \text{Scatter}_{\Omega_{\text{roi}}}\left(\hat{x}_{1,\text{roi}}\right) + (1 - M_{\text{blend}}) \odot x_g, \tag{10}$$

where $\text{Scatter}_{\Omega_{\text{roi}}}$ restores the predicted latents to their original positions. The feathered blending mask M_{blend} is derived from the same downsampled face masks, with weights transitioning smoothly from 1 in the inner face region to 0 outside the compute region.

Joint Rec-ID Loss. As shown in Sec. 3.1, latent reconstruction is relatively insensitive to facial identity changes. We therefore combine clean-latent reconstruction with explicit identity supervision (Fig. 3, left). To reduce training memory usage, we decode a small subset of frames from $\tilde{x}_1$ and evaluate their identity using a frozen face-recognition encoder f_{id} (Deng et al., 2019a):

$$\mathcal{L}_{\text{id}} = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} \left[1 - \cos\Big(f_{\text{id}}(\hat{I}_k), f_{\text{id}}(I_r)\Big)\right], \tag{11}$$

where $\hat{I}_k$ is a decoded frame and $\mathcal{K}$ indexes the selected frames with valid face detections. Before blending, the predicted clean RoI latent is supervised against the corresponding ground-truth latent:

$$\mathcal{L}_{\text{rec}} = \|\hat{x}_{1,\text{roi}} - x_{1,\text{roi}}\|_2^2, \tag{12}$$

where $x_{1,\text{roi}}$ is the restriction of x_1 to Ω_{roi}. Phase I minimizes

$$\mathcal{L} = \mathbb{E}_{(I_r,V_g,V_r),x_0,t}\left[\lambda_{\text{id}}\mathcal{L}_{\text{id}} + \lambda_{\text{rec}}(\sigma_t)\mathcal{L}_{\text{rec}}\right], \tag{13}$$

where the expectation is over role-reversed training samples, Gaussian noise, and flow time. The identity weight λ_{id} remains fixed, while the reconstruction weight depends on the noise level:

$$\lambda_{\text{rec}}(\sigma_t) = \begin{cases} \lambda_{\text{high}}, & \sigma_t > \tau, \\ \lambda_{\text{low}}, & \sigma_t \le \tau, \end{cases} \qquad \lambda_{\text{high}} > \lambda_{\text{low}}. \tag{14}$$

This schedule emphasizes reconstruction supervision at high noise levels to guide clean-latent recovery, including facial identity, while giving identity supervision greater relative weight at low noise levels to refine identity-specific facial details (App. F).

4.3 Phase II: Short-Context Rollout Training

Recent forcing-based methods adapt bidirectional video models for streaming generation through causal autoregressive modeling (Yin et al., 2025; Huang et al., 2025; Zhu et al., 2026). Our analysis in Sec. 3.2 motivates a simpler adaptation for reference-guided face swapping: the target video supplies content and motion cues, so a short generated history can provide context for local temporal continuity. We therefore retain bidirectional attention within each window and condition the generation of new frames on this short history.

Let N_{ctx} and N_{gen} denote the numbers of history and newly generated latent frames, respectively. For generation window n, we take the latest N_{ctx} latent frames of previous output $\tilde{x}_1^{(n-1)}$ as $x_{\text{ctx}}^{(n)}$ and generate the next chunk as

$$\hat{x}_{1,\text{roi}}^{(n)} = \mathcal{G}_\theta\left(x_0^{(n)}; x_{\text{ctx}}^{(n)}, c_{\text{swap}}^{(n)}\right), \tag{15}$$

where $x_0^{(n)}$ contains Gaussian noise for N_{gen} latent frames and $\mathcal{G}_\theta$ denotes the multi-step denoising process using the velocity model v_θ. The condition $c_{\text{swap}}^{(n)}$ includes the source image and aligned target-video and mask inputs covering both the history and new frames. History latents remain fixed at $\sigma = 0$ throughout denoising. The predicted RoI latents are composited using Eq. 10, and the latest composited frames provide history for the next window. The first window generates $N_{\text{ctx}} + N_{\text{gen}}$ latent frames without history. We use $N_{\text{ctx}} = 3$ and $N_{\text{gen}} = 9$ to balance generation quality, throughput, and latency (App. G).

Training with ground-truth history creates a mismatch with streaming inference, where prediction errors can propagate through the generated history and accumulate across chunks. Inspired by Self Forcing (Huang et al., 2025), we fine-tune the Phase-I model through autoregressive self-rollout, exposing it to its own generated history during training (Fig. 3, right). We use a 4-step denoising schedule, as the Phase-I model achieves comparable identity and target-attribute performance with 4 and 25 steps at substantially higher throughput (App. E). Following the stochastic gradient-truncation strategy of Self Forcing, we randomly select a denoising step as the stopping point and backpropagate only through the clean-latent prediction at that step. Gradients are stopped through both the generated history and preceding denoising steps. Rather than a distribution-matching objective, we use the Phase-I identity and Face-RoI reconstruction losses in Eq. 13 to supervise the newly generated frames, retaining the noise-dependent weights in Eq. 14.

5 Experiments

We initialize **StreamingSwap** from DreamID-V (Guo et al., 2026) and train each phase for 1,000 iterations. We evaluate identity similarity, target-attribute preservation, FVD, and end-to-end inference efficiency. All speed measurements use a single NVIDIA RTX 4090 GPU. Detailed metrics and implementation settings are provided in App. D.

Table 1: Quantitative comparison on short and long videos. **StreamingSwap** combines effective identity transfer with strong target-attribute preservation and video quality at over 20 FPS.

Method	Identity	Attribute Preservation				Quality	Efficiency	
	ID Sim.↑	Pose↓	Expr.↓	Light↓	Gaze↑	FVD↓	FPS↑	Latency (s)↓
Short videos								
BlendFace	0.533	6.218	**2.471**	0.301	0.875	130.77	16.548	0.276
InSwapper	**0.735**	6.084	2.624	0.252	0.873	138.09	**28.621**	**0.050**
FaceAdapter	0.129	8.563	2.681	0.329	0.823	257.99	0.353	2.833
VividFace	0.175	8.870	2.695	0.314	0.786	204.15	0.283	28.267
VFace	0.373	11.258	3.004	0.366	0.761	252.24	0.129	20.662
DreamID-V	0.734	9.338	3.009	0.509	0.764	231.00	1.815	44.891
StreamingSwap (Ours)	0.577	**5.354**	2.484	**0.242**	**0.881**	**129.09**	21.174	1.782
Long videos								
BlendFace	0.533	7.173	**2.464**	0.306	0.867	154.05	16.548	0.276
InSwapper	**0.723**	6.701	2.584	0.248	0.867	147.01	**28.621**	**0.050**
DreamID-V	0.715	10.827	2.967	0.489	0.798	249.41	1.815	44.891
StreamingSwap (Ours)	0.576	**6.336**	2.466	**0.238**	**0.881**	**136.23**	21.174	1.782



Figure 4: Qualitative comparison on short and long videos. Higher identity similarity can coexist with reduced target fidelity and realism, whereas **StreamingSwap** balances them.

5.1 Comparisons with Existing Methods

We compare **StreamingSwap** with two representative GAN-based methods and four diffusion-based methods that can run on a single NVIDIA RTX 4090 GPU, with quantitative and qualitative results in Table 1 and Fig. 4. Long-video comparisons include three strong-performing methods. Across both benchmarks, **StreamingSwap** achieves effective identity transfer while preserving target fidelity and visual realism. Although InSwapper and DreamID-V achieve higher identity similarity, the frame-level scores in Fig. 4 illustrate that higher embedding similarity does not necessarily imply better perceptual quality or target fidelity. DreamID-V exhibits pronounced changes in lighting and head pose despite high ID Sim. scores, while InSwapper's high scores coexist with less realism in the displayed examples. In contrast, **StreamingSwap** maintains recognizable source identity while more closely following the target lighting, expressions, and head poses. These comparisons highlight the importance of balancing identity similarity and target-attribute reconstruction, as further examined in App. F. Overall, **StreamingSwap** offers a balanced quality–efficiency, exceeding 20 FPS and achieving substantially higher throughput than the evaluated diffusion baselines.

Table 2: Ablation of the main components. Phase-I variants remove identity supervision or Face-RoI. The Phase-II variant replaces rollout training with ground-truth-history conditioning.

Variant	ID Sim.↑	Pose↓	Expr.↓	Light↓	Gaze↑	FPS↑
Phase I on short videos						
DreamID-V (Initialization, no CFG)	0.445	8.504	2.695	0.427	0.798	1.815
w/o ID supervision	0.273	5.507	1.922	0.198	0.881	30.280
w/o Face-RoI	0.574	6.168	2.460	0.262	0.865	12.784
StreamingSwap (Phase I)	0.583	6.086	2.476	0.260	0.869	30.280
Phase II on long videos						
w/o self-rollout	0.573	6.318	2.481	0.262	0.714	21.174
StreamingSwap (Phase II)	0.576	6.336	2.466	0.238	0.881	21.174

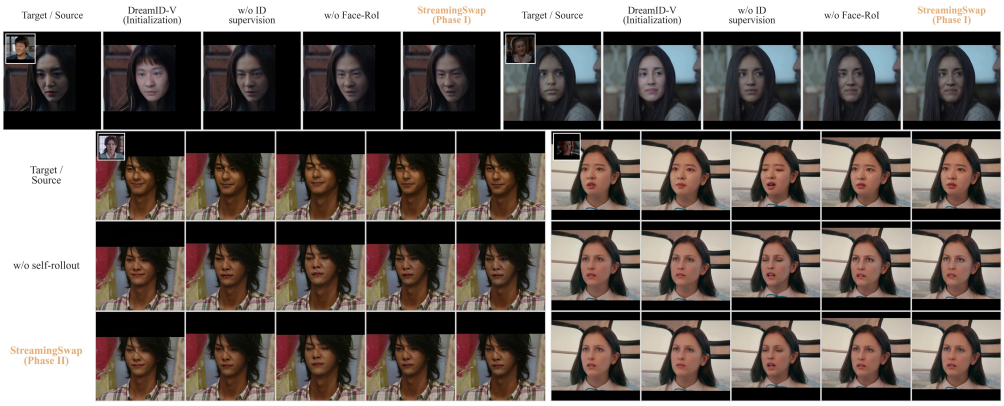

Figure 5: Qualitative ablation of **StreamingSwap**. Top: identity supervision strengthens source-identity transfer, while Face-RoI preserves visual quality. Bottom: self-rollout aligns training with streaming inference, mitigating error accumulation and progressive visual degradation.

5.2 Ablation of Identity-aware Supervision and Efficient Context

Setup. We separately remove identity supervision and Face-RoI in Phase I, and replace self-rollout with ground-truth-history conditioning in Phase II. Table 2 and Fig. 5 summarize the results.

Identity supervision. Removing identity supervision weakens source-identity transfer despite improving target-attribute scores, as also reflected in the visual examples. This confirms that flow matching supervision alone is insufficient to ensure the intended identity edit.

Face-RoI processing. Face-RoI increases throughput while retaining comparable identity, attribute scores, and visual appearance, supporting the removal of redundant background computation.

Self-rollout training. These results suggest that self-rollout training mitigates error accumulation across streaming windows by exposing the model to its own generated history.

Phase I vs Phase II. Phase II enables streaming generation with comparable performance but lower throughput: each window processes three history frames and nine new frames, yielding roughly three quarters of the non-streaming throughput before additional per-window overhead.

6 Conclusion

We presented **StreamingSwap**, a reference-guided framework for streaming video face swapping that aligns supervision and computation with the task requirements. Our analysis motivates *Joint Rec-ID Loss* to strengthen identity transfer, alongside *Face-RoI Generation* and *Short-Context Rollout Training* to reduce spatial and temporal computation. We also introduced **HumanFaceVid** to support training and evaluation on short and long videos. Experiments demonstrate competitive identity transfer and target fidelity at over 20 FPS on a single NVIDIA RTX 4090 GPU, with stable generation on videos up to 10 minutes long. These results show that task-specific supervision and compact context can make reference-guided generation substantially more efficient.

AI USE STATEMENT

We used Codex to assist with Chinese–English translation, language refinement, interpretation of experimental results, and figure and table organization and LaTeX formatting. We also used a pre-trained face-swapping model to construct synthetic training videos, as described in Appendix B. The synthetic data were filtered using identity similarity and other quality metrics, followed by manual inspection of randomly sampled examples. The authors take responsibility for the final content of this work, including all AI-assisted text.

ETHICS STATEMENT

This work develops video face-swapping techniques for research and creative applications, including film production, AI-generated dramas, and live streaming. We recognize that realistic identity manipulation can be misused for impersonation, misinformation, and non-consensual content creation, and that efficient streaming may increase these risks. The additional application demos use video clips obtained from Bilibili and reference images located through Google Images. The face-swapped outputs are synthetic research demonstrations and should not be interpreted as authentic recordings or endorsements by the depicted individuals. We oppose deceptive and non-consensual uses of this technology and advocate obtaining appropriate permissions and consent, respecting privacy and content ownership, and clearly labeling manipulated media.

REPRODUCIBILITY STATEMENT

The method section describes Face-RoI processing, joint reconstruction–identity supervision, short-context rollout training, and streaming inference. Appendix B details the data preparation pipeline. Appendix D provides additional training and evaluation settings, including model initialization, hyperparameters, sampling and window configurations, metric definitions, and hardware and acceleration settings. We also specify the timing scope for end-to-end efficiency measurements. Together, these descriptions support reproduction of the method and its experimental evaluation.

REFERENCES

Ahmed A Abdelrahman, Thorsten Hempel, Aly Khalifa, Ayoub Al-Hamadi, and Laslo Dinges. L2cs-net: Fine-grained gaze estimation in unconstrained environments. In *2023 8th International Conference on Frontiers of Signal Processing (ICFSP)*, pp. 98–102. IEEE, 2023.

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*, 2020.

Xu Chen, Keke He, Junwei Zhu, Yanhao Ge, Wei Li, and Chengjie Wang. Hifivfs: High fidelity video face swapping. *arXiv preprint arXiv:2411.18293*, 2024.

DeepFakes. https://github.com/deepfakes/face- swap. *Accessed: 2020-12-20*, 2020.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019a.

Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019b.

Gege Gao, Huaibo Huang, Chaoyou Fu, Zhaoyang Li, and Ran He. Information bottleneck disentanglement for identity swapping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021.

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

Google. MediaPipe Face Landmarker. https://ai.google.dev/edge/mediapipe/solutions/vision/face_landmarker.

Xu Guo, Fulong Ye, Xinghui Li, Pengqi Tu, Pengze Zhang, Qichao Sun, Songtao Zhao, Xiangwang Hou, and Qian He. Dreamid-v: Bridging the image-to-video gap for high-fidelity face swapping via diffusion transformer. *arXiv preprint arXiv:2601.01425*, 2026.

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. *arXiv preprint arXiv:2405.12970*, 2024.

Ruhs Henry. Facefusion: Industry leading face manipulation platform, 2025. URL https://github.com/facefusion/facefusion. Accessed: 2025-09-23.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025.

InsightFace. Lightweight facial landmark models with coordinate regression. https://github.com/deepinsight/insightface/tree/master/alignment/coordinate_reg.

Hui Li, Mingwang Xu, Yun Zhan, Shan Mu, Jiaye Li, Kaihui Cheng, Yuxuan Chen, Tan Chen, Mao Ye, Jingdong Wang, et al. Openhumanvid: A large-scale high-quality dataset for enhancing human-centric video generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7752–7762. IEEE, 2025.

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.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Xiangyang Luo, Ye Zhu, Yunfei Liu, Lijian Lin, Cong Wan, Zijian Cai, Shao-Lun Huang, and Yu Li. Canonswap: High-fidelity and consistent video face swapping via canonical space modulation. *arXiv preprint arXiv:2507.02691*, 2025.

Zekai Luo, Zongze Du, Zhouhang Zhu, Hao Zhong, Muzhi Zhu, Wen Wang, Yuling Xi, Chenchen Jing, Hao Chen, and Chunhua Shen. Preserving source video realism: High-fidelity face swapping for cinematic quality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40653–40663, 2026.

ModelTC. ComfyUI-LightVAE: High-performance VAE custom nodes. GitHub repository. URL https://github.com/ModelTC/ComfyUI-LightVAE. Accessed: 2026-09-24.

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*, 2018.

Hao Shao, Shulun Wang, Yang Zhou, Guanglu Song, Dailan He, Shuo Qin, Zhuofan Zong, Bingqi Ma, Yu Liu, and Hongsheng Li. Vividface: A diffusion-based hybrid framework for high-fidelity video face swapping. *arXiv preprint arXiv:2412.11279*, 2024.

Kaede Shiohara, Xingchao Yang, and Takafumi Taketomi. Blendface: Re-designing identity encoders for face-swapping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Runqi Wang, Sijie Xu, Tianyao He, Yang Chen, Wei Zhu, Dejia Song, Nemo Chen, Xu Tang, and Yao Hu. Dynamicface: High-quality and consistent video face swapping using composable 3d facial priors. *arXiv preprint arXiv:2501.08553*, 2025.

Bowen Xue, Zheng-Peng Duan, Qixin Yan, Wenjing Wang, Hao Liu, Chun-Le Guo, Chongyi Li, Chen Li, and Jing Lyu. 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.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22963–22974. IEEE, 2025.

Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 2018.

Jintao Zhang, Haofeng Huang, Pengle Zhang, Jia Wei, Jun Zhu, and Jianfei Chen. Sageattention2: Efficient attention with thorough outlier smoothing and per-thread int4 quantization. *arXiv preprint arXiv:2411.10958*, 2024.

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*, 2023.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. *arXiv preprint arXiv:2602.02214*, 2026.

A Related Work

Video face swapping transfers the identity of a source face to a target video while preserving target-specific attributes such as expression, pose, illumination, motion, and background. Existing approaches can be broadly grouped into GAN-based methods, which emphasize efficient feed-forward synthesis, and diffusion-based methods, which leverage generative priors to improve visual quality and temporal coherence.

A.1 GAN-based video face swapping

GAN-based face swapping learns efficient feed-forward generators under adversarial supervision (Goodfellow et al., 2014). Early systems used shared encoders with identity-specific decoders and therefore required separate training for each identity pair (DeepFakes, 2020). Later subject-agnostic methods inject source-identity features into target representations and improve identity transfer or identity-attribute disentanglement through specialized architectures, objectives, and identity encoders (Li et al., 2019; Gao et al., 2021; Chen et al., 2020; Shiohara et al., 2023). CanonSwap extends this direction to videos by modifying identity in a canonical space before restoring target motion (Luo et al., 2025). InSwapper is a widely used one-shot model distributed through the InsightFace ecosystem, although its architecture and training procedure are not publicly documented Henry (2025). While these methods offer efficient inference, most image-based variants process videos frame by frame and thus do not explicitly model temporal dependencies.

A.2 Diffusion-based video face swapping

Diffusion models have enabled a second line of face-swapping methods with stronger generative priors. One branch follows an inpainting-oriented formulation, in which the target face is masked or otherwise reconstructed from the source identity and explicit target-attribute conditions. DiffSwap formulates face swapping as conditional inpainting and combines masked diffusion with identity and 3D facial guidance (Zhao et al., 2023). Face-Adapter introduces lightweight modules for injecting source identity, spatial structure, and detailed target attributes into pretrained image diffusion models (Han et al., 2024). Extending this formulation to videos, DynamicFace decomposes facial conditions using 3D priors and incorporates temporal modeling, while HiFiVFS builds on a pretrained video diffusion model with disentangled target-attribute extraction and detailed identity injection (Wang et al., 2025; Chen et al., 2024). VividFace jointly trains on images and videos and introduces a video-oriented VAE, identity-attribute disentanglement data, and 3D facial guidance to improve temporal consistency and robustness (Shao et al., 2024). Related identity-control frameworks instead introduce lightweight conditional branches into pretrained video generators and can be adapted to face swapping, although face swapping is not their primary formulation Xue et al. (2026). These approaches improve synthesis quality and temporal modeling, but reconstructing a removed target face from auxiliary conditions can discard fine-grained appearance cues that are difficult to recover explicitly.

A more recent reference-guided branch retains the complete target video as a visual condition and learns face swapping from synthetically constructed pairs. LivingSwap reverses GAN-generated training pairs so that real videos provide reconstruction targets, and uses edited keyframes together with temporal stitching to maintain identity across video chunks (Luo et al., 2026). DreamID-V constructs bidirectional identity supervision through SyncID-Pipe and trains a Diffusion Transformer with modality-aware conditioning, curriculum learning, and identity-coherence optimization (Guo et al., 2026). By directly observing the target face, reference-guided models can preserve fine-grained expressions, illumination, makeup, and occlusions more faithfully. However, access to the complete target face also allows the model to reconstruct most of the video while responding weakly to the source identity, and existing systems retain the computational cost of full-frame, chunk-based diffusion generation. Our work follows the reference-guided formulation but studies a different boundary: identity-sensitive supervision and efficient streaming inference using face-region generation and a fixed-size generated history.

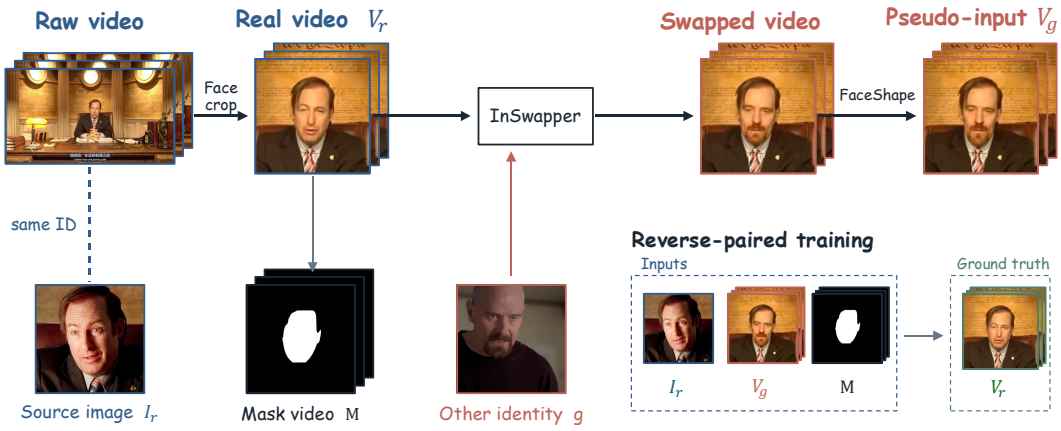

Figure 6: Data preparation and role-reversed training for **HumanFaceVid**. We extract face-centered clips V_r and aligned masks M, and select a source image I_r with the same identity. InSwapper transfers another identity to V_r, followed by facial-contour warping to construct the pseudo-input V_g. Training conditions on (I_r, V_g, M) to reconstruct the original real video V_r, avoiding synthesis artifacts as ground-truth targets.

B Data Preparation

We construct **HumanFaceVid** from OpenHumanVid (Li et al., 2025). We detect and track faces, discard short or unreliable tracks, and obtain frame-aligned binary face masks by combining BiSeNet face parsing (Yu et al., 2018) with InsightFace landmark convex hulls (InsightFace). Source images are selected from cross-video identity clusters, yielding real image–video pairs (I_r, V_r) of the same identity. For each real video, a face-swapping model InSwapper (Henry, 2025) transfers a different identity to produce the generated video V_g in Eq. 5. We additionally apply temporally smoothed random facial-contour warps to V_g, guided by MediaPipe landmarks (Google), to reduce residual geometric identity cues. The resulting data are used in the reversed formulation, with the original real video retained as ground-truth supervision.

C Applications in Different Scenarios

Figures 7–9 illustrate applications of **StreamingSwap** to movies, live and online videos, and AI-generated short dramas.

Movie footage requires identity editing to preserve both cinematic lighting and the nuances of the original performance. The displayed results follow the target video's illumination, subtle facial expressions, and head poses, including shots with partial occlusion, integrating the source identity while retaining the scene's visual character.

Live and online videos require sustained streaming generation that follows a speaker's expressions and head movements. Our method supports continuous processing with stable facial appearance in the displayed sequences, including partially occluded faces. Its streaming throughput makes it suitable for live-video applications whose frame-rate and buffering requirements match the measured processing speed and latency.

AI-generated short dramas present another use case: reinforcing character identity when large motions, complex lighting, or exaggerated expressions cause identity drift. By applying a consistent source identity to the generated footage, **StreamingSwap** provides an efficient post-processing step that strengthens identity consistency while retaining the original performance and scene context, without regenerating the entire scene.

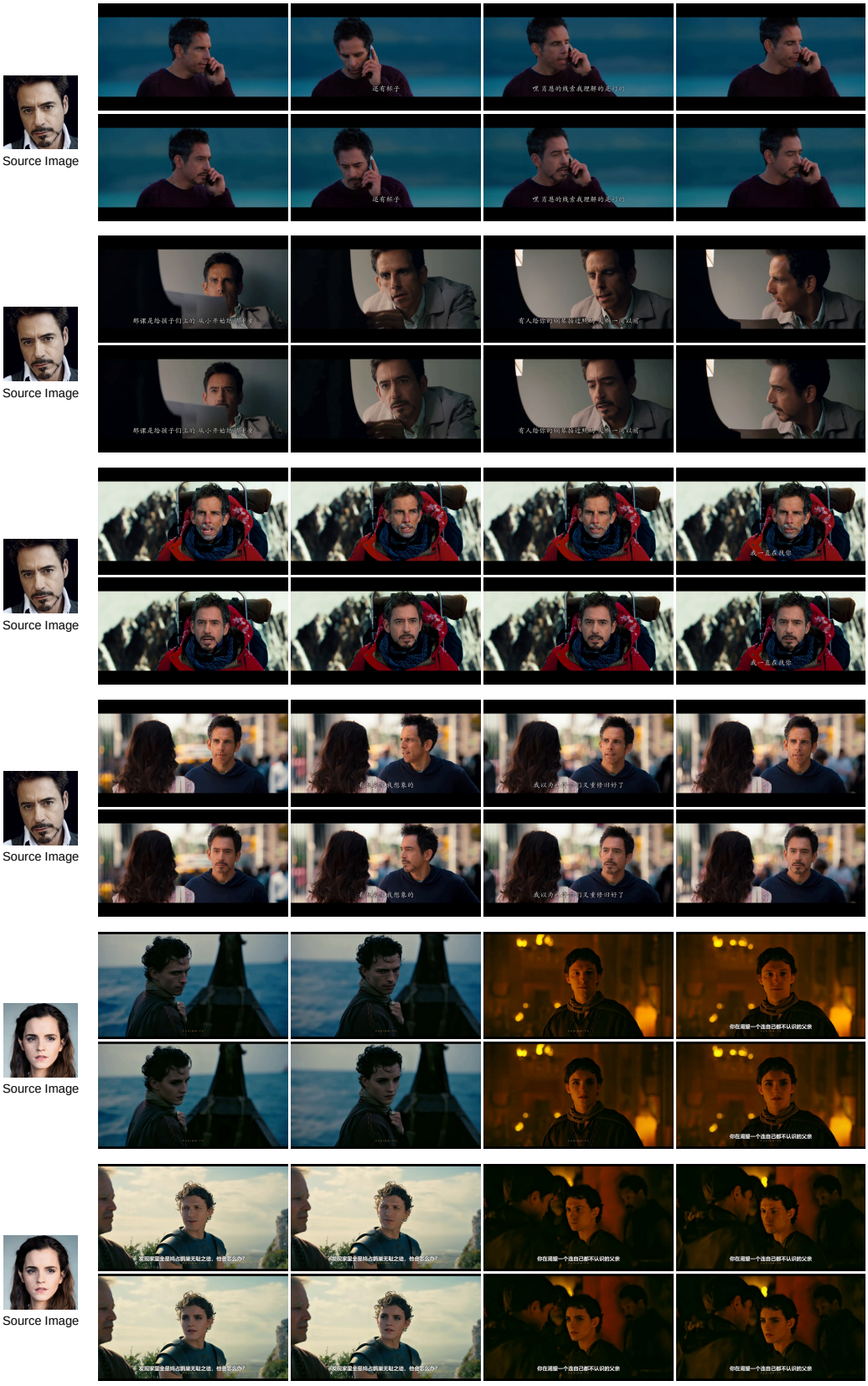

Figure 7: More qualitative results for movies.



Figure 8: More qualitative results for live videos.

D EXPERIMENTAL SETTINGS

Metrics. We evaluate identity transfer, target-attribute preservation, video quality, and inference efficiency. Following DreamID-V (Guo et al., 2026), we measure identity transfer using cosine similarity between ArcFace (Deng et al., 2019a) embeddings of the source image and generated frames, reported as ID Sim. We also annotate the displayed outputs with frame-level ID Sim. scores to facilitate comparison between embedding similarity and visual appearance. Following LivingSwap (Luo et al., 2026), we assess target-attribute preservation through distances between generated and corresponding target frames in estimated head pose (Pose) (Ruiz et al., 2018), expression coefficients (Expr.) (Deng et al., 2019b), and lighting coefficients (Light) (Deng et al., 2019b). Gaze preservation is measured by cosine similarity between their estimated 3D gaze vectors (Gaze) (Abdelrahman et al., 2023). We evaluate video quality using Fréchet Video Distance (FVD). For efficiency, FPS measures the number of newly generated video frames output per second by the complete pipeline, excluding reused history frames. Latency measures the end-to-end time required for a single in-

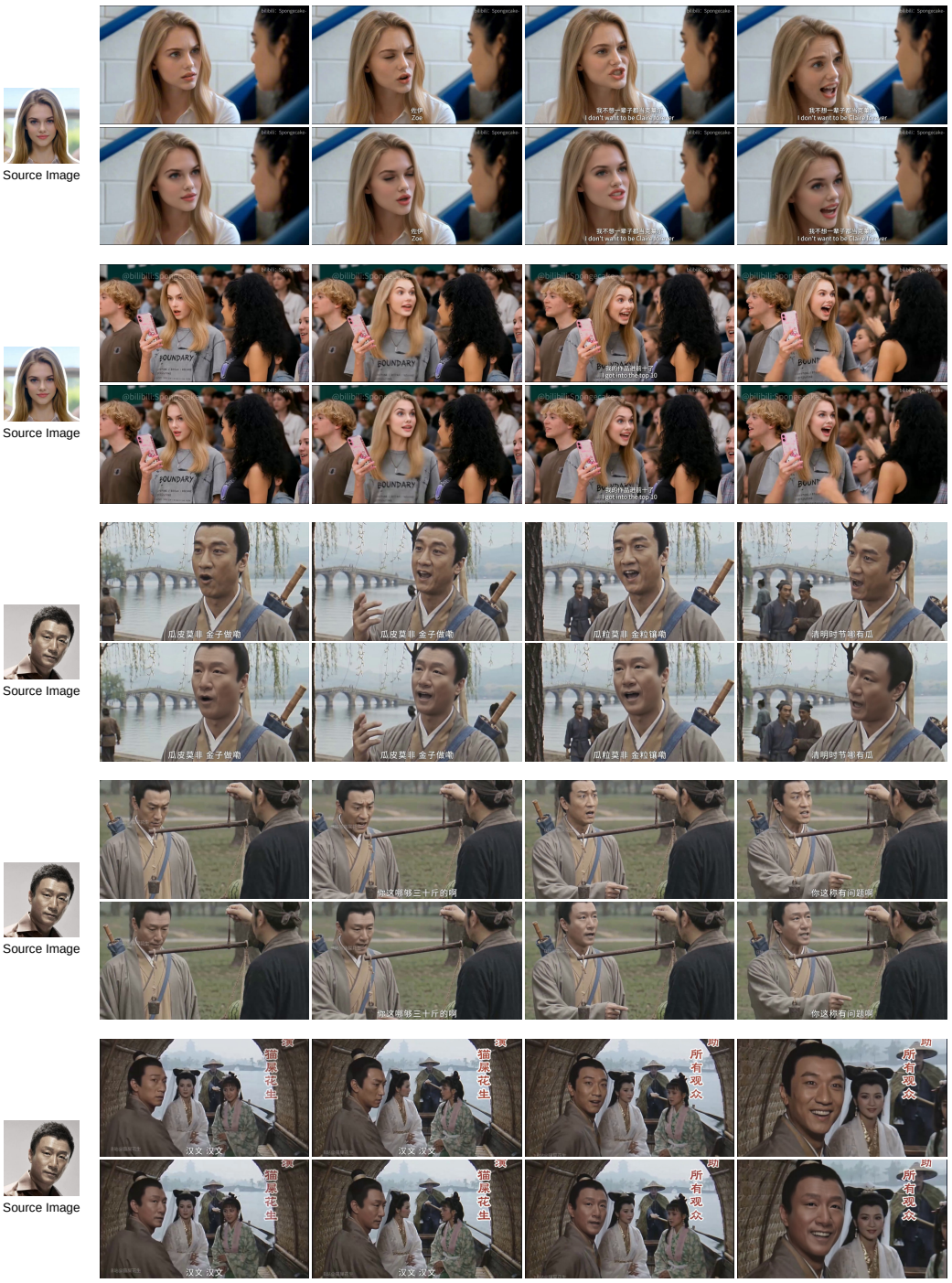

Figure 9: More qualitative results for AI-generated dramas.

ference run. Both measurements include VAE encoding, denoising, VAE decoding, and output compositing.

Implementation Details. We initialize our model from the pretrained DreamID-V checkpoint (Guo et al., 2026) and train Phase I and Phase II for 1,000 iterations each. Both phases use a learning rate of 10^{-5} and a batch size of 8. In Phase II, each training rollout consists of five autoregressive generation windows. Training uses four NVIDIA H100 GPUs, while all inference-speed measure-

Table 3: Ablation of training phase and denoising steps. DreamID-V is the initialization model. Phase I uses bidirectional clip generation, while Phase II adds self-rollout training for streaming generation. All model are evaluated on short videos bench

Model	Steps	ID Sim.↑	Pose↓	Expr.↓	Light↓	Gaze↑	FPS↑
DreamID-V (Initialization)	16	0.445	8.504	2.695	0.427	0.798	1.815
StreamingSwap (Phase I)	25	0.573	6.318	2.481	0.262	0.868	6.702
StreamingSwap (Phase I)	4	**0.583**	6.086	**2.476**	0.260	0.869	**30.280**
StreamingSwap (Phase II, Stream)	4	0.577	**5.354**	2.484	**0.242**	**0.881**	21.174

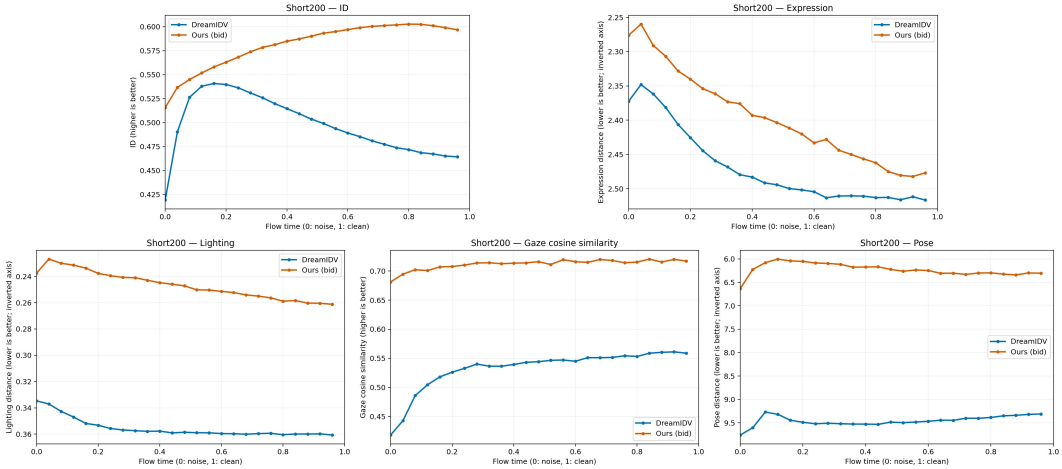

Figure 10: Identity and attribute metrics along the denoising trajectory on short200. Our bidirectional model and DreamID-V exhibit distinct trends across metrics. Flow time progresses from noise to clean data; error axes are inverted so that higher positions indicate better performance.

ments are conducted on a single NVIDIA GeForce RTX 4090 GPU. For inference acceleration, we use SageAttention (Zhang et al., 2024) for DiT self-attention and cross-attention, and compile Transformer blocks individually using `torch.compile`. We also replace the native Wan VAE with LightTAE for Wan 2.1 (ModelTC) for lightweight encoding and decoding. For fair efficiency comparisons, we apply the same acceleration strategies to all compatible methods.

E Ablation of Training Phase and Denoising Steps

Setup. We compare the DreamID-V initialization with 16 denoising steps, the Phase-I bidirectional model with 25 and 4 steps, and the final Phase-II streaming model with 4 steps. We evaluate identity similarity, target-attribute preservation, and end-to-end throughput to assess the effects of training phase and sampling budget. To contextualize the sampling results, we also track identity and attribute metrics along the denoising trajectories of DreamID-V and the Phase-I model.

Results and interpretation. As shown in Table 3, the four-step Phase-I model improves identity similarity from 0.445 to 0.583 over the initialization and improves all four target-attribute metrics. Compared with 25-step sampling, it achieves comparable or slightly better identity, pose, expression, and lighting scores, together with higher gaze similarity, while increasing throughput from 6.702 to 30.280 FPS, a $4.52\times$ speedup. The final streaming model retains most of this identity similarity (0.577) and further improves pose, lighting, and gaze scores, with a slight increase in expression error. Its throughput decreases to 21.174 FPS because streaming repeatedly processes overlapping history and incurs per-window overhead, while reused history frames are excluded from the output frame count. These results support four-step sampling as an efficient choice for rollout training and streaming inference.

The trajectories in Fig. 10 provide further context for the sampling results. For our model, identity similarity improves through most of denoising before declining slightly near the endpoint, whereas expression and lighting errors generally increase after reaching their minima early in the trajectory. Thus, progressing further along the denoising trajectory does not consistently improve the evaluated

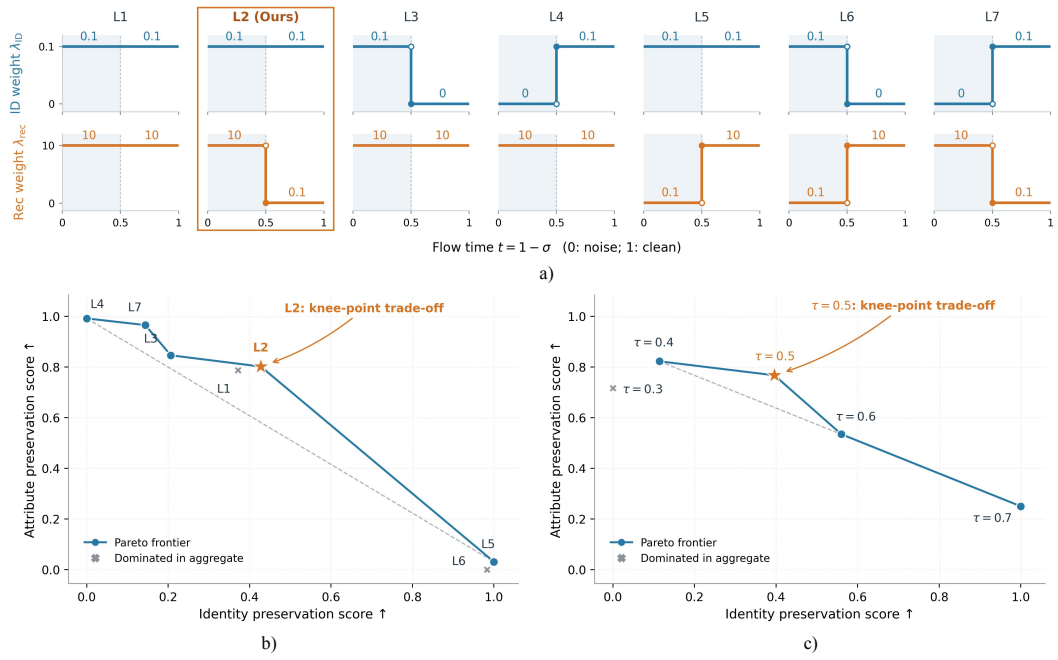

Figure 11: Joint Rec-ID supervision ablations on short200. (a) Weighting schedules. (b) Trade-offs across L1–L7. (c) Threshold sweep for L2. Scores average direction-aligned, min–max-normalized identity and attribute metrics within each sweep. Blue lines connect Pareto-optimal configurations; stars mark the endpoint-distance knee points.



Figure 12: Qualitative comparison of joint reconstruction–identity supervision schedules (L1–L7). L2 balances source-identity transfer and target-attribute preservation.

metrics. One possible explanation is that late-stage refinement introduces appearance details that do not necessarily improve identity or attribute similarity as measured by the evaluation encoders. This interpretation is consistent with the slightly better scores obtained with four-step sampling, although the trajectories alone do not establish why changing the step count produces these gains. Four-step sampling may still sacrifice some fine visual detail relative to 25 steps, so its favorable identity and attribute scores should not be interpreted as uniformly better perceptual quality.

F Ablation of Joint Rec-ID Supervision

Our analysis in Sec. 3.1 motivates explicit identity supervision to complement latent reconstruction. We examine how the weighting schedule, switching threshold, and reconstruction objective affect the balance between identity transfer and target-attribute preservation.

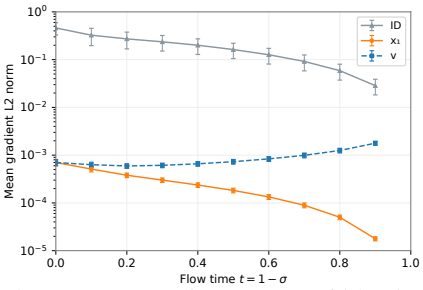

Figure 13: Gradient norms of identity supervision, reconstruction and velocity across noise levels.

Table 4: Reconstruction ablation on Short200.

Objective	ID Sim ↑	Pose ↓	Expr. ↓	Light ↓	Gaze ↑
x_1-pred	0.5830	6.0862	2.4756	0.2601	0.8686
v-pred	0.5870	6.1843	2.5052	0.2664	0.8667

Table 5: Ablation of streaming window configurations. N_{ctx} and N_{gen} denote the numbers of history and newly generated latent frames, respectively. † marks the default configuration.

N_{ctx}	N_{gen}	ID Sim.↑	Pose↓	Expr.↓	Light↓	Gaze↑	FPS↑	Latency (s)↓
1	1	0.583	7.151	2.593	0.272	0.860	11.396	**0.350**
1	3	0.588	6.833	2.547	0.256	0.868	19.332	0.592
3	3	**0.592**	6.780	2.562	0.256	0.867	16.063	0.848
3†	9†	0.577	6.487	2.488	0.239	0.880	**21.174**	1.782
5	9	0.576	6.424	2.488	0.241	0.877	18.774	2.136
5	12	0.568	**6.292**	**2.453**	**0.233**	**0.884**	19.672	2.652

†Default configuration.

Setup. We compare seven piecewise-constant weighting configurations, L1–L7, with the noise range divided at $\sigma = 0.5$ (Fig. 11(a)). The identity weight is either 0.1 or 0, and the reconstruction weight is either 10 or 0.1. For L2, we further vary the switching threshold τ over $\{0.3, 0.4, 0.5, 0.6, 0.7\}$, keeping the identity weight at 0.1 and the high- and low-noise reconstruction weights at 10 and 0.1, respectively. We also compare clean-endpoint and velocity reconstruction under the L2 schedule and examine the gradient norms of the identity and two reconstruction losses during training. All configurations share the same initialization, training budget, and four-step evaluation protocol on the short200 benchmark of **HumanFaceVid**.

Weighting configurations. L2 balances identity transfer and target-attribute preservation by retaining identity supervision throughout training while reducing the reconstruction weight at low noise levels. Compared with L1, it strengthens identity transfer with small changes in attribute preservation and imaging quality. Removing identity supervision from either noise interval weakens identity transfer (L3/L4/L7), whereas weakening high-noise reconstruction compromises target fidelity despite stronger identity similarity (L5/L6). The qualitative results in Fig. 12 illustrate this trade-off, particularly the changes in expression and lighting under L5/L6. Together with its knee-point position in Fig. 11(b), these results support L2 as a balanced choice.

Switching threshold. Increasing τ narrows the noise interval receiving strong reconstruction supervision, favoring identity transfer over target-attribute preservation. Across the sweep, identity similarity increases alongside expression and lighting errors, while pose and gaze vary non-monotonically. Under the normalized aggregation in Fig. 11(c), $\tau = 0.4$ achieves the highest attribute score and $\tau = 0.7$ the highest identity score. We use $\tau = 0.5$, the knee point of this trade-off.

Reconstruction objective. We compare clean-endpoint and velocity reconstruction while keeping the velocity-prediction architecture unchanged. The training gradient norms in Fig. 13 indicate that endpoint reconstruction maintains a more consistent gradient scale relative to identity supervision across noise levels. Under the L2 schedule, endpoint reconstruction improves all four target-attribute metrics, with slightly lower identity similarity (Table 4). We therefore adopt endpoint reconstruction to better preserve target attributes while retaining comparable identity transfer.

G Ablation of Streaming Window Size

Setup. We vary the history length N_{ctx} and generation length N_{gen} to examine the quality–efficiency trade-off in streaming inference. Following Sec. 4.3, each continuation generates N_{gen} new latent

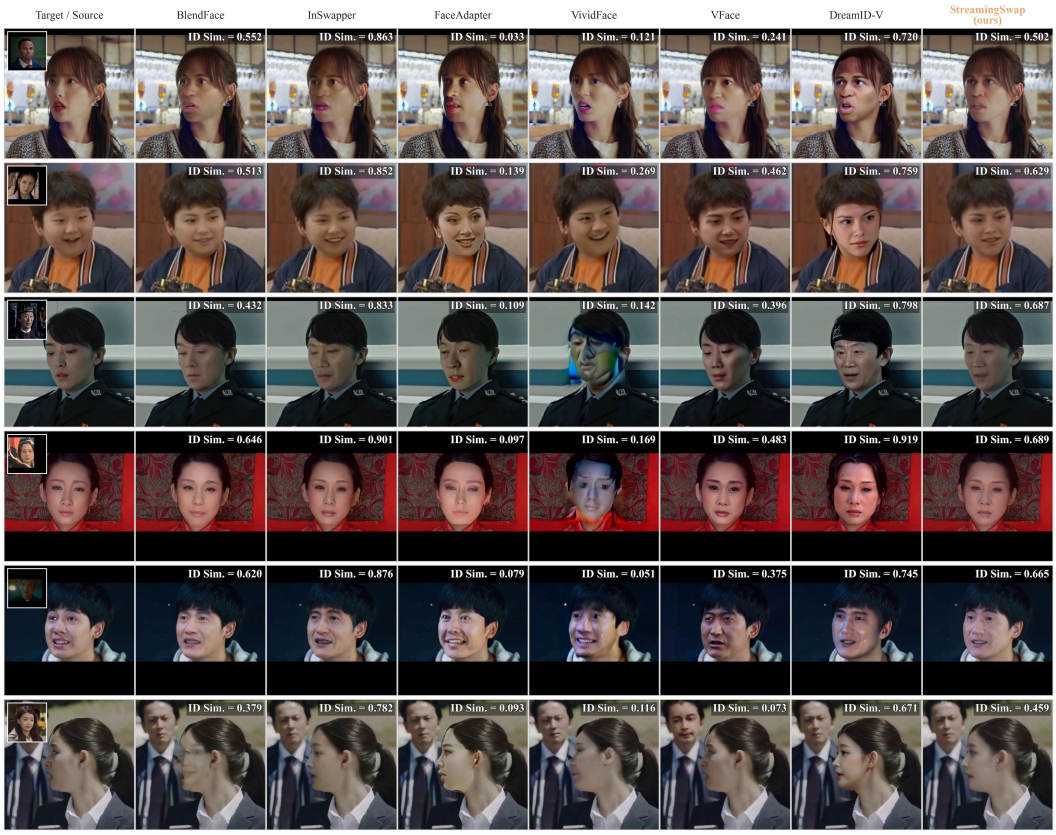

Figure 14: More qualitative results for method comparisons on short video bench.

frames conditioned on N_{ctx} clean composited history frames, using bidirectional attention within the window. Unlike the main results, which use a model trained with a fixed $(N_{\text{ctx}}, N_{\text{gen}}) = (3, 9)$ configuration, this ablation uses a model trained for 1,000 iterations with randomly sampled window sizes. Table 5 compares six configurations in terms of identity similarity, target-attribute preservation, and end-to-end throughput and latency.

Results. Increasing N_{gen} at a fixed history length improves throughput but increases latency, whereas increasing N_{ctx} at a fixed generation length adds computational overhead and reduces throughput. Among the evaluated configurations, $(N_{\text{ctx}}, N_{\text{gen}}) = (3, 9)$ achieves the highest FPS with lower latency than the larger-window alternatives. Target-attribute preservation generally improves with larger windows, while identity similarity peaks at $(3, 3)$ and then declines. We therefore adopt $(3, 9)$ as a practical balance between throughput and latency, while maintaining competitive identity transfer and target-attribute preservation.

H Additional Qualitative Comparisons on Benchmarks

We provide additional qualitative comparisons in Fig. 14, 15 on the short and long video benchmarks to supplement the results presented in the main paper. These examples further illustrate identity transfer and target-attribute preservation across diverse scenes.

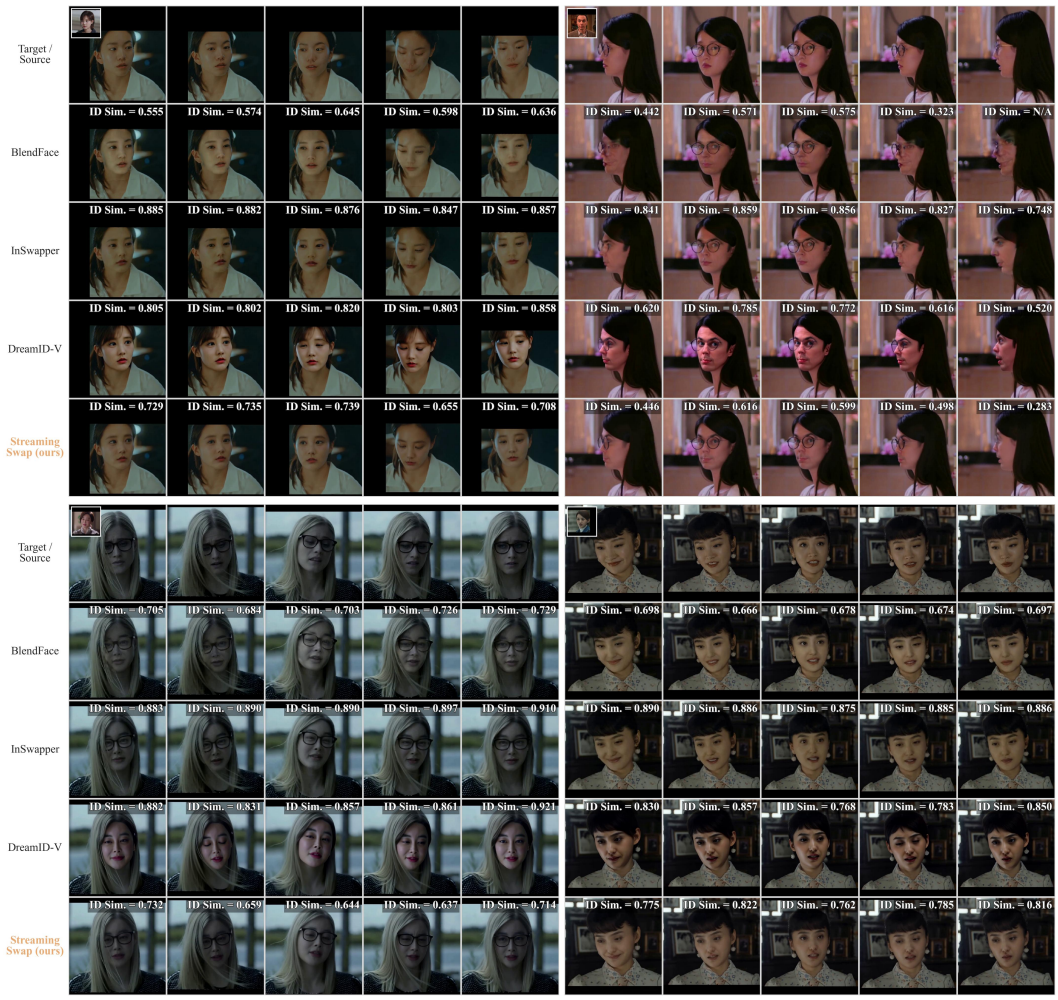

Figure 15: More qualitative results for method comparisons on long video bench.