PIXREENACT: PIXEL-CONDITIONED CAUSAL VIDEO DIFFUSION FOR STREAMING HEAD-AVATAR REENACTMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Streaming head-avatar reenactment aims to animate a reference image according to a live driving video, requiring robust motion transfer, long-term identity stability, and low latency. Existing methods often rely on specialized identity or motion representations, which can discard useful visual information and inherit failure modes from external extractors. In addition, many recent diffusion-based reenactment methods use offline, clip-based generation, jointly processing and denoising an entire video clip before producing its output, making continuous low-latency streaming difficult. We introduce **PixReenact**, a pixel-conditioned streaming reenactment framework built on causal video diffusion. PixReenact conditions directly on VAE-encoded reference and driving frames, without specialized identity or motion representations. To separate reference identity from driver motion, we train with cross-identity pseudo supervision together with corrective objectives anchored to the original reference and driving inputs. Long self-rollouts reduce autoregressive drift, while state-aware dual-teacher distillation separately addresses cold-start and steady-state generation. Across three cross-identity benchmarks and a long-horizon streaming benchmark, PixReenact demonstrates robust cross-identity reenactment, particularly under challenging conditions such as extreme viewpoints, occlusions, and pronounced facial expressions, while maintaining the reference identity over long streams. A 4-NFE rolling student continuously emits four frames per update with a mean emission latency of 239 ms.

1 INTRODUCTION

A streaming head-avatar or portrait-animation continuously animates a reference image from a live driving video, transferring the driver's facial expressions and head motion while preserving the reference identity Li et al. (2026). Such avatars provide a visual interface for interactive digital agents, with applications in video conferencing, telepresence, and conversational interaction Ki et al. (2026); Grassal et al. (2022); Wang et al. (2021). Interactive use requires **low-latency, stable, and robust streaming reenactment**.

Many existing head-avatar methods condition generation on specialized identity or motion representations, such as implicit keypoints Guo et al. (2024); Li et al. (2026), 3DMM or pose/expression parameters Chu & Harada (2024), and learned identity or motion representations Zhao et al. (2025); Xu et al. (2025); Cheng et al. (2025). These representations provide effective and controllable conditioning. Moreover, motion representations that suppress appearance reduce identity information carried by the driving signal, supporting cross-identity reenactment. However, they can discard fine-grained visual information. In particular, methods that rely on external keypoint or motion extractors inherit their failure modes: under severe occlusions, challenging expressions, or large head-pose variations, inaccurate motion estimates can degrade the reenactment. This raises a natural question: can head-avatar reenactment instead operate directly on the visual representations of the reference and driving frames, *without specialized identity or motion representations for conditioning*?

Beyond robust conditioning, a second challenge is low-latency continuous generation. Many recent diffusion-based reenactment methods use offline, clip-based generation, jointly processing and denoising a full video clip before producing output Cheng et al. (2025); Xu et al. (2025); Yin et al.

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.

(2025a); Guo et al. (2025). Consequently, the first output frames become available only after denoising the full clip, making low-latency streaming difficult. Longer videos require processing successive clips, often using overlapping or historical frames to maintain continuity. Such clip-wise continuation may suffer from temporal inconsistency and identity drift. Causal autoregressive video diffusion therefore provides a natural basis for incremental streaming generation.

Motivated by recent advances in causal autoregressive video generation Huang et al. (2025); Liu et al. (2026); Samuel et al. (2026), we start from a compact bidirectional video-diffusion backbone and adapt it to causal generation using the CausVid Yin et al. (2025b) framework, conditioning directly on VAE-encoded reference and driving frames. This yields a streaming model without specialized identity or motion representations. However, this design introduces three challenges. First, under standard self-driving training, direct visual conditioning creates a trivial shortcut: since the reference and driver share the same identity, the model can copy the driver's appearance rather than learn to separate identity from motion. This motivates cross-identity supervision, for which real paired ground truth is generally unavailable. While pseudo targets from an offline reenactment model provide a practical alternative, they may contain identity, motion, or appearance inaccuracies. Second, autoregressive generation repeatedly conditions on its own previous outputs, allowing small errors to accumulate over long streams and cause temporal inconsistency or identity drift. Finally, stream initialization is challenging when the initial driver pose or expression differs substantially from the reference image, requiring a large motion transition while preserving the reference identity.

To address these challenges, we introduce **PixReenact**, a pixel-conditioned streaming head-avatar framework built on causal video diffusion. We train PixReenact in a teacher-student framework with cross-identity pseudo pairs generated offline by Wan-Animate Cheng et al. (2025), together with real self-driving samples for clean reconstruction supervision. Since pseudo targets may contain identity, motion, or appearance inaccuracies, we mitigate their effect using corrective objectives anchored to the reference and driving inputs. To reduce autoregressive drift, the student is trained on long self-rollouts, using its own generated history as at inference. We treat stream initialization separately using state-aware dual-teacher distillation, with cold-start and steady-state teachers. The resulting model continuously emits frames with low latency while conditioning directly on visual inputs.

We evaluate PixReenact across three cross-identity and a long-horizon streaming benchmarks. PixReenact achieves robust cross-identity motion transfer, particularly under challenging conditions such as extreme viewpoints and occlusions, while maintaining the reference identity over long streams. Using a 4-NFE rolling denoising schedule, PixReenact continuously emits four frames per update with a mean emission latency of 239 ms. Our main contributions are:

- We introduce a **pixel-driven streaming reenactment framework** that conditions directly on VAE-encoded reference and driving frames, without specialized identity or motion representations for conditioning.
- We develop a **training strategy for robust cross-identity streaming**, combining cross-identity pseudo supervision, corrective objectives, and long self-rollouts to mitigate driver-appearance leakage and long-term autoregressive drift.
- We propose state-aware dual-teacher distillation, using separate cold-start and steady-state teachers to distill a 4-NFE causal student for low-latency continuous generation.

2 Related Work

2.1 Head Avatar Reenactment and Conditioning

Head avatar reenactment animates a reference identity using motion from a driving video. A common strategy is to transform driving frames into specialized conditioning representations. LivePortrait Guo et al. (2024) and PersonaLive Li et al. (2026) use implicit keypoint-based motion conditioning, while HunyuanPortrait Xu et al. (2025) uses pretrained encoders for implicit motion features. Wan-Animate Cheng et al. (2025) encodes face images into frame-wise implicit latents injected into the diffusion backbone through a dedicated face-attention module. MegActor-Σ Yang et al. (2025) uses a learned driven encoder to transform the driving video into specialized conditioning features before the denoising transformer. X-NeMo Zhao et al. (2025) learns a 1D identity-agnostic latent

motion descriptor from the driving image without pretrained motion detectors, but compresses the driving signal into a specialized latent representation before conditioning the diffusion model.

Other methods explicitly decompose the driving signal: DeX-Portrait Shi et al. (2026) separates motion into geometric head pose and a learned facial-expression latent. PortraitDirector Ji et al. (2026) decomposes the driving signal into specialized pose, local facial-motion, and emotion representations before conditioning the diffusion model. FlashPortrait Tu et al. (2026) similarly relies on identity-agnostic facial expression features. These representations provide effective motion control and help suppress driver appearance leakage, but their compression may discard fine-grained visual information. Methods relying on external motion extractors can also inherit their estimation errors.

An alternative retains richer driving-pixel information. MegActor Yang et al. (2024) explores raw-video-driven portrait animation and identifies identity leakage as a key challenge of direct visual conditioning. Unlike PixReenact, it uses an image-diffusion backbone with temporal modules, whereas we use a native video-diffusion backbone adapted for causal generation. X-Portrait Xie et al. (2024) likewise learns motion control from driving RGB and reduces appearance leakage through cross-identity augmentation that modifies the conditioning signal. In contrast, our conditioning remains entirely real, and only the supervision target is synthesized, preserving the same real-driving interface used at inference. Concurrent Wan-Animate-2 Wang et al. (2026) similarly uses direct visual conditioning and Wan-Animate-generated cross-identity supervision, whose Lite variant supports autoregressive streaming. Unlike Wan-Animate-2 Lite, PixReenact combines pseudo supervision with reference- and driver-anchored corrective losses during student training, uses separate cold-start and steady-state score teachers, and emits four-frame rather than eight-frame updates.

2.2 Streaming and Causal Video Generation

Many recent diffusion-based portrait animation methods target offline generation rather than causal streaming. FlashPortrait Tu et al. (2026) supports long, potentially unbounded animation through overlapping sliding-window generation and accelerates window-wise denoising by up to $6\times$, but remains offline and relies on extracted facial-expression features. PersonaLive Li et al. (2026) introduces autoregressive micro-chunk generation for low-latency long-term streaming, with sliding training and historical keyframes, while retaining specialized facial-motion and implicit-keypoint conditioning. PortraitDirector Ji et al. (2026) supports real-time streaming through diffusion distillation and causal attention, but conditions on decomposed pose, local facial-motion, and emotion representations. In contrast, PixReenact combines low-latency causal streaming with direct visual conditioning rather than specialized identity or motion representations.

Beyond portrait animation, recent causal video-diffusion works Yin et al. (2025b); Huang et al. (2025); Liu et al. (2026); Zhu et al. (2026) show that video diffusion models can be adapted and distilled into few-step autoregressive generators for low-latency streaming. These works establish causal autoregressive video diffusion as a practical approach for streaming, but do not address pixel-conditioned cross-identity reenactment challenges, including driver-appearance leakage, imperfect cross-identity supervision, and large reference-to-driver transitions at stream initialization. PixReenact builds on this paradigm with training mechanisms tailored to these challenges.

3 Method

3.1 Overview

Head avatar reenactment animates a reference image I_{ref} using the facial expression, head pose, and motion of a driving video $D_{1:T}$. We consider the challenging setting of *streaming head-avatar reenactment*, where driving frames arrive sequentially and the model must continuously generate $Y_{1:T}$, preserving the reference identity while transferring the driver's motion with low latency. Unlike methods relying on specialized conditioning representations such as keypoints or learned motion codes that may discard fine-grained visual cues, PixReenact conditions directly on VAE-encoded reference and driving frames. Avoiding such bottlenecks is important in streaming, since lost information cannot be recovered and errors may propagate through the autoregressive context.

Direct visual conditioning introduces challenges for cross-identity reenactment and autoregressive streaming. We therefore combine cross-identity supervision with corrective objectives to reduce

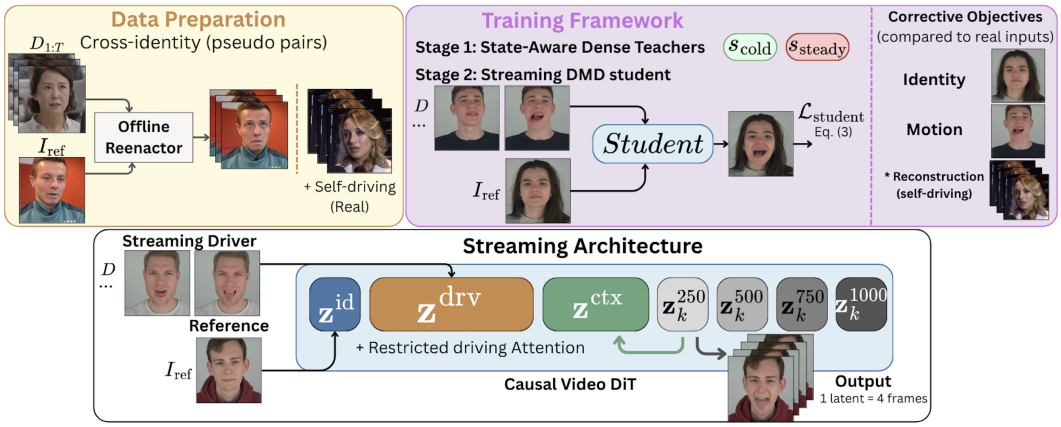

Figure 1: **PixReenact Overview.** Top-left: offline cross-identity pseudo targets from real reference-driver pairs, plus real self-driving data. Top-right: cold-start and steady-state teachers are distilled into a 4-NFE student with DMD and corrective objectives. Bottom: reference, driving, and context latents condition a causal video DiT with restricted driving attention and rolling four-frame emission.

appearance leakage, train on long self-rollouts to reduce autoregressive drift, and treat stream initialization separately from steady-state generation. At inference, a causal rolling denoising process continuously incorporates new driving frames and emits new video frames with low latency.

3.2 From Video Diffusion to Pixel-Driven Streaming Reenactment

A natural route to streaming reenactment is adapting a diffusion-based reenactment model for causal generation. However, such models typically target high-quality offline generation and are not naturally suited to low-latency streaming Li et al. (2026); Tu et al. (2026). We instead convert a compact bidirectional video-diffusion backbone into a causal autoregressive generator using the framework of Yin et al. (2025b). This yields an efficient streaming backbone, but not yet a reenactment model.

To obtain robust reenactment without specialized conditioning, we condition directly on VAE-encoded reference and driving frames. Figure 1 (bottom) shows the resulting streaming architecture. Let $\mathcal{E}$ and $\mathcal{D}$ denote the frozen VAE encoder and decoder. Given a reference image I_{ref} and driving frames $D_{1:T}$, we obtain an identity latent $\mathbf{z}^{\text{id}} = \mathcal{E}(I_{\text{ref}})$ and driving latents $\mathbf{z}^{\text{drv}}_{1:T'} = \mathcal{E}(D_{1:T})$. The causal VAE compresses time by four, so each latent spans four frames and $T' = \lceil T/4 \rceil$. All temporal indices below refer to this latent grid. The reference and driving latents become DiT tokens and are concatenated with the context and video tokens into a shared self-attention sequence.

At streaming iteration k, the model processes

$$[\mathbf{z}^{\text{id}} \mid \mathbf{z}^{\text{drv}} \mid \mathbf{z}^{\text{ctx}} \mid \mathbf{z}_k^{250}, \mathbf{z}_k^{500}, \mathbf{z}_k^{750}, \mathbf{z}_k^{1000}],$$

where $\mathbf{z}^{\text{drv}}$ denotes the local driving latents used for the current generation window, and $\mathbf{z}^{\text{ctx}}$ contains two slots filled by the latest emitted video latents. Before sufficient history is available, context slots are initialized from the reference. The term $\mathbf{z}_k^t$ denotes an in-flight video latent at diffusion timestep t; all in-flight latents are indexed by the current iteration for brevity. At each iteration, a new latent enters at $t = 1000$. The forward pass advances all in-flight latents by one level and emits the cleanest as $\hat{\mathbf{z}}_k$. A latent admitted at iteration k undergoes four passes and is emitted at $k{+}3$. It is decoded into four output frames, $Y_k = \mathcal{D}(\hat{\mathbf{z}}_k)$, using a causal streaming VAE decoder with a persistent temporal cache. The latest emitted latents update $\mathbf{z}^{\text{ctx}}$ for subsequent iterations. We obtain the causal model by applying causal attention masks to the bidirectional backbone and adapt it using self-attention LoRA. The rolling denoising and KV-cache design follows Yin et al. (2025b).

To preserve temporal correspondence, the video latent at position n attends only to a three-latent driving window $\mathbf{z}^{\text{drv}}_{n-1:n+1}$ through a restricted attention mask. Neighboring latents provide temporal context around the aligned driving latent while preventing attention to unrelated positions. Although this includes one-latent lookahead, it adds no latency beyond the rolling denoising pipeline: the

four-step staircase ensures $\mathbf{z}^{\mathrm{drv}}_{n+1}$ is available before latent n is emitted. Thus, the model gains local temporal context while retaining streaming generation. We evaluate this window in Sec. 4.4.

However, because the driving latents also contain the driver's appearance, standard self-driving training creates a strong shortcut. Since I_{ref} and $D_{1:T}$ share the same identity, the model is not explicitly required to preserve identity from I_{ref} while transferring pose, expression, and motion from $D_{1:T}$. It can instead copy appearance directly from the driving stream.

3.3 Learning Cross-Identity Reenactment

To remove this shortcut, we use **cross-identity supervision**, where reference and driving inputs come from different identities. Cross-identity training has also been explored in prior portrait-animation methods using synthetic, retargeted, or augmented pairs Xie et al. (2024); Luo et al. (2026); Yang et al. (2024); Yan et al. (2026). Since real paired cross-identity ground truth with matched motion is generally unavailable, we construct pseudo pairs using a strong offline reenactment model Cheng et al. (2025), as illustrated in Figure 1 (top-left). We sample a reference image I^A_{ref} and driving video $D^B_{1:T}$ from different identities ($A \neq B$), and use the offline model to generate a pseudo target $\tilde{Y}^{A\leftarrow B}_{1:T}$ intended to preserve the appearance of A while following the motion of B. Unlike prior strategies that modify or synthesize the cross-identity conditioning signal, our reference and driving inputs remain entirely real, while only the supervision target is synthesized. Thus, training uses the same real-driving input interface as inference.

Pseudo targets may contain identity, pose, expression, and visual-quality errors. We therefore filter low-quality targets using identity, motion, and emotion criteria (Appendix B.1). During student training, corrective objectives use the original inputs: identity is supervised against I^A_{ref}, while pose, expression, and emotion are supervised against the real driver $D^B_{1:T}$ (Sec. 3.5). We also mix real self-driving samples for clean reconstruction and reduced reliance on imperfect pseudo targets. Together, these signals retain the benefits of cross-identity supervision while mitigating pseudo-target errors.

3.4 State-Aware Streaming Distillation

While the above training enables cross-identity reenactment, we next describe how to achieve low-latency streaming. As shown in Figure 1 (top-right), we first train causal dense-timestep score teachers from the bidirectional backbone, then distill them into a causal 4-NFE streaming student using Distribution Matching Distillation (DMD) Yin et al. (2024), following Yin et al. (2025b). Teachers and student share the same causal architecture and pixel-conditioning interface.

Stage 1: State-aware dense teachers. We apply the causal attention masks of Sec. 3.2 to the bidirectional backbone and train self-attention LoRA over a dense range of noise levels using cross-identity supervision. The first streaming emission differs substantially from later ones. At $k = 1$, no generated history is available, so $\mathbf{z}^{\mathrm{ctx}}$ contains only reference-initialized latents. This creates a challenging **cold-start** state, especially when reference and driver have a large pose gap: the model must establish the driver's pose while preserving the reference identity. For $k > 1$, the context contains generated history already better aligned with the current driver state, so the model mainly follows incremental motion while maintaining temporal consistency and identity.

We therefore train two dense teachers specialized to these rollout states. The cold-start teacher uses reference-initialized context and emphasizes large reference-driver pose gaps, whereas the steady-state teacher uses preceding target latents as context, matching later rollouts. Both teachers retain the base video-diffusion model's native flow-velocity parameterization, but are supervised using MSE on the implied clean-latent estimate. For each clean target latent $\mathbf{z}^0$, we independently sample a diffusion timestep t, while keeping context latents clean ($t = 0$), construct $\mathbf{z}^t$, and obtain $\hat{\mathbf{z}}^0 = \mathbf{z}^t - \sigma_t \hat{\mathbf{v}}^t$. The predicted flow is also converted to the corresponding score for distillation. The real-score teacher used during distillation is

$$s^{(k)}_{\mathrm{real}} = s_{\mathrm{cold}} \text{ if } k = 1, \quad s_{\mathrm{steady}} \text{ if } k > 1. \tag{1}$$

Stage 2: Streaming DMD student. The student follows the 4-level denoising staircase as at inference (Sec. 3.2), and is trained with long self-rollouts. This exposes it to its own history, reducing training-inference mismatch and long-horizon error accumulation.

We re-noise an emitted latent $\hat{\mathbf{z}}_k$ at timestep t to obtain $\mathbf{z}_k^t$. The frozen state-dependent teacher $s_{\text{real}}^{(k)}$ estimates the target score, while the online s_{fake} estimates the student score. The DMD gradient is

$$\mathbf{g}_{\text{DMD}}^{(k)} = w(t)\Big[s_{\text{fake}}(\mathbf{z}_k^t, t) - s_{\text{real}}^{(k)}(\mathbf{z}_k^t, t)\Big], \tag{2}$$

where $w(t)$ is the standard DMD timestep weight and both scores condition on the identity, driving, and context tokens. We use the standard surrogate $\mathcal{L}_{\text{DMD}} = \frac{1}{2}\|\hat{\mathbf{z}}_k - \text{sg}(\hat{\mathbf{z}}_k - \mathbf{g}_{\text{DMD}}^{(k)})\|_2^2$, with gradient $\mathbf{g}_{\text{DMD}}^{(k)}$ w.r.t. $\hat{\mathbf{z}}_k$, backpropagated through the rollout. See Appendix B.4, B.5 for training details.

3.5 Corrective Training Objectives

In addition to DMD, we apply corrective objectives during student training. Cross-identity pseudo targets may contain errors in identity, motion, and appearance, so these objectives are computed from the original real inputs. At emission k, let Y_k denote the four emitted frames and D_k their aligned driving frames. Frame-wise losses are averaged over the four frames.

For **identity preservation**, we use $\mathcal{L}_{\text{id}} = \max\big(0, m - \cos(\phi(Y_k), \phi(I_{\text{ref}}))\big)$, where $\phi(\cdot)$ is a frozen identity embedding and m reflects typical same-identity similarity, avoiding unrealistic pressure. To suppress driver-identity leakage, we use $\mathcal{L}_{\text{leak}} = \max\big(0, \cos(\phi(Y_k), \phi(D_k)) - m_{\text{leak}}\big)$ on cross-identity samples only. For **motion supervision**, we apply ℓ_1 losses between estimated head-pose and expression parameters of Y_k and driver D_k ($\mathcal{L}_{\text{pose}}, \mathcal{L}_{\text{expr}}$), rather than the pseudo target. We also enforce emotion consistency using cosine similarity between extracted emotion features ($\mathcal{L}_{\text{emo}}$).

For real self-driving samples, exact reconstruction targets are available, so we apply LPIPS and adversarial losses for visual fidelity. All supervised samples, real and pseudo, also receive an ℓ_2 latent anchor $\mathcal{L}_{\text{anc}}$ toward their target latents. The full student objective is

$$\begin{aligned}\mathcal{L}_{\text{student}} = {} & \lambda_{\text{DMD}}\mathcal{L}_{\text{DMD}} + \lambda_{\text{anc}}\mathcal{L}_{\text{anc}} + \lambda_{\text{id}}\mathcal{L}_{\text{id}} + \lambda_{\text{leak}}\mathcal{L}_{\text{leak}} \\ & + \lambda_{\text{pose}}\mathcal{L}_{\text{pose}} + \lambda_{\text{expr}}\mathcal{L}_{\text{expr}} + \lambda_{\text{emo}}\mathcal{L}_{\text{emo}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}},\end{aligned} \tag{3}$$

where $\mathcal{L}_{\text{rec}}$ combines the LPIPS and adversarial losses on real self-driving samples. Estimator details, margins, loss weights, and implementation details are provided in Appendix B.6.

4 Evaluation

4.1 Experimental Setup

Training data. Following PersonaLive Li et al. (2026), we use the same three datasets for self-driving training: NeRSemble Kirschstein et al. (2023), VFHQ Xie et al. (2022), and DH-FaceVid-1K Di et al. (2025). PersonaLive does not report its number of training clips. Our self-driving set contains 1,548 clips: 402 from NeRSemble, 146 from VFHQ, and 1,000 from DH-FaceVid-1K. For cross-identity training, we reuse these clips with different-identity references and add 800 FFHQ Karras et al. (2019) references, each paired with three drivers. Wan-Animate Cheng et al. (2025) generates offline pseudo targets, yielding 13,561 pairs after label-quality filtering (Sec. B.1).

All clips are 4-second, 512^2 segments. For self-driving, the first frame is the reference and the clip both driver and ground truth. We call the Wan-Animate-supervised model **PixReenact**. To test pseudo-label sensitivity, we train an otherwise identical *PL-labels* variant using PersonaLive Li et al. (2026) with the same reference-driver pairings. Details and results are in the Appendix (Sec. B.1, C).

Training and evaluation sets are strictly disjoint. NeRSemble evaluation uses 15 held-out subjects, VFHQ uses VFHQ-Test, and the 50 FFHQ evaluation images are not used in training. TalkingHead-1KH Wang et al. (2021) and CelebV-HQ Zhu et al. (2022) are used only for evaluation.

Baselines and metrics. We compare PixReenact with image-based methods LivePortrait Guo et al. (2024) and X-Portrait Xie et al. (2024); offline video methods MegActor-Σ Yang et al. (2025), Wan-Animate Cheng et al. (2025), X-NeMo Zhao et al. (2025), Wan-Animate-2-Base Wang et al. (2026) (Wan-Animate-2 in all tables), and FlashPortrait Tu et al. (2026); and streaming method PersonaLive Li et al. (2026). All baselines are run using their official releases; implementation details are in Appendix B.10. We measure identity preservation by cosine similarity between AdaFace embeddings Kim et al. (2022) of generated frames and the reference image (ID). Motion transfer uses



Figure 2: **Qualitative comparison under challenging viewpoints and occlusions.** Compared with PersonaLive, PixReenact better preserves the reference identity while following the driver's motion.

average expression (AED) and pose (APD) distances from SMIRK Retsinas et al. (2024) estimates for generated and driving frames. Temporal consistency and video quality are measured by tLP Chu et al. (2020) and Fréchet Video Distance (FVD) Unterthiner et al. (2019), respectively. For streaming methods, we report mean steady-state emission latency per four-frame chunk and throughput in frames per second (FPS), computed as $4/\text{latency}$. The metrics capture complementary reenactment quality aspects and should be interpreted jointly; their limitations are discussed in Appendix B.12.

Evaluation benchmarks. We evaluate PixReenact on three cross-identity protocols and one long-horizon streaming benchmark. First, we introduce a 68-pair extreme-condition benchmark with CelebV-HQ drivers Zhu et al. (2022) and FFHQ references, targeting large viewpoint changes and challenging occlusions. We also evaluate on two standard benchmarks. Following LivePortrait Guo et al. (2024), we use 25 TalkingHead-1KH test clips Wang et al. (2021) still available on YouTube (out of 35 used in Guo et al. (2024)) and 50 VFHQ Xie et al. (2022) test clips as drivers, paired with the first 50 FFHQ images Karras et al. (2019) as references. We also construct a 70-pair benchmark from 15 held-out NeRSemble subjects Kirschstein et al. (2023), with five providing references and all 15 serving as drivers. For long-horizon streaming, we evaluate identity stability on 28 cross-identity videos over five minutes (details in Appendix B.9). We additionally report self-reenactment and single-window cross-reenactment results in the Appendix (Tables 10, 11, and 12). Supplementary videos covering the extreme-condition, standard cross-identity, long-horizon, and self-reenactment settings are included with the submission and described in Appendix A.

Implementation details. PixReenact uses Wan2.1-T2V-1.3B Team et al. (2025) as its bidirectional video-diffusion backbone and adopts the causal generation and DMD distillation framework of CausVid Yin et al. (2025b). We apply causal attention masks and use a 4-NFE rolling denoising schedule with timesteps $[1000, 750, 500, 250, 0]$, adapting the DiT with self-attention LoRA. All experiments use 512^2 resolution, with runtime measurements on a single NVIDIA H200 GPU. Additional training hyperparameters and implementation details are provided in the Appendix B.

4.2 Robust Streaming Reenactment

Extreme viewpoints and occlusions. We first evaluate the challenging conditions motivating our pixel-driven formulation, comparing streaming methods on difficult inputs. Figure 2 shows representative examples with large viewpoints and occlusions, including hands, face masks, and facial accessories. PixReenact remains effective while preserving the reference appearance.

Table 1 reports results on our 68-pair CelebV-HQ extreme-condition benchmark. PixReenact substantially improves identity preservation over PersonaLive (ID 0.727 vs. 0.472), lowers AED (1.135 vs. 1.295), with similar tLP (1.853 vs. 1.866). PersonaLive obtains lower APD and FVD, despite visible degradation in several qualitative examples. PixReenact's clearest advantage in this challenging setting is its substantially stronger reference-identity preservation. This highlights the importance of interpreting metrics jointly under extreme viewpoints and occlusions (see Appendix B.12).

Table 1: **Extreme-condition benchmark** (68 pairs). FFHQ references driven by CelebV-HQ clips with extreme poses and occlusions. Mean±std over clips; FVD is set-level.

Method	ID↑	AED↓	APD↓	tLP↓	FVD↓
PersonaLive	0.472 ±0.23	1.295 ±0.64	**0.135** ±0.10	1.866 ±1.57	**643.5**
PixReenact	**0.727** ±0.11	**1.135** ±0.30	0.169 ±0.12	**1.853** ±1.35	662.5

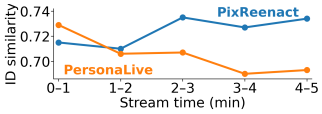

Figure 3: **Long-term identity stability.** AdaFace similarity over five minutes on 28 clips.

Long-horizon streaming. Autoregressive self-conditioning can accumulate errors, so we evaluate identity stability over 28 five-minute clips. Figure 3 reports mean AdaFace similarity per minute: PersonaLive declines over time, while PixReenact remains stable. Appendix Figure 7 provides qualitative examples, showing reference-identity stability over horizons much longer than training.

4.3 Standard Cross-Identity Benchmarks

We next evaluate PixReenact on the public and NeRSemble cross-identity benchmarks. Figure 4 shows qualitative results. PixReenact preserves reference identity while following the driver's facial motion, including pronounced expressions. Additional comparisons against all baselines are shown in Appendix Figure 5. Table 2 compares image-based, offline video-based, and streaming methods. On the public benchmark, PixReenact achieves the best FVD (377.1) and better identity preservation than PersonaLive (ID 0.709 vs. 0.662), with comparable AED (0.790 vs. 0.788). PersonaLive performs better in APD and tLP. On NeRSemble, PixReenact improves over PersonaLive in ID (0.753 vs. 0.734), AED (0.714 vs. 0.824), tLP (0.803 vs. 0.890), and FVD (264.8 vs. 310.0), while remaining close in APD (0.0426 vs. 0.0392). Overall, PixReenact remains competitive with offline methods while streaming continuously without specialized extractors at inference.

PixReenact emits four frames at 238.9 ms (16.74 FPS), versus 211.1 ms (18.95 FPS) for PersonaLive. The gap mainly reflects our video-native causal DiT and causal VAE decoder, compared with PersonaLive's image-based U-Net and VAE decoder. A detailed comparison is in Appendix Table 6.

4.4 Ablation Studies

We ablate PixReenact's main components on the 75-pair public cross-identity benchmark in Table 3. Each ablation continues the same checkpoint for 5K steps, with the full recipe as reference, so absolute values may differ from Table 2. Qualitative effects are shown in Appendix Figure 8.

Pixel-driven conditioning. We ablate the driving-attention window of Sec. 3.2. Replacing our ± 1 window with only the aligned latent ($w = 0$) hurts identity preservation: ID drops from 0.707 to 0.585, while driver-ID similarity (IDD) rises from 0.023 to 0.210. Meanwhile, AED, APD, and FVD improve, showing that better motion scores can reflect driver similarity rather than better reenactment. Local temporal context therefore better balances motion transfer and identity preservation.

Cross-identity learning and corrective objectives. Removing cross-identity supervision worsens AED ($0.816 \rightarrow 0.852$), APD ($0.071 \rightarrow 0.087$), and FVD ($389 \rightarrow 458$), supporting its role in learning cross-identity motion transfer. Without identity losses, ID drops to 0.353 and IDD rises to 0.245 while motion scores improve due to driver leakage. Without motion losses, ID rises to 0.795 but AED, APD, and FVD worsen. Both are needed to balance identity preservation and motion transfer.

State-aware distillation. Using the steady-state teacher for the first emission worsens AED ($0.816 \rightarrow 0.834$), APD ($0.071 \rightarrow 0.079$), and FVD ($389 \rightarrow 406$), despite slightly higher ID. This supports a dedicated first-emission teacher when motion-aligned history is unavailable.

Quality-latency trade-off. Reducing 4-NFE to 3/2-NFE cuts latency from 238.9 to 211.4/186.6 ms but worsens motion and distribution quality. Appendix Table 7 gives full PersonaLive comparisons.

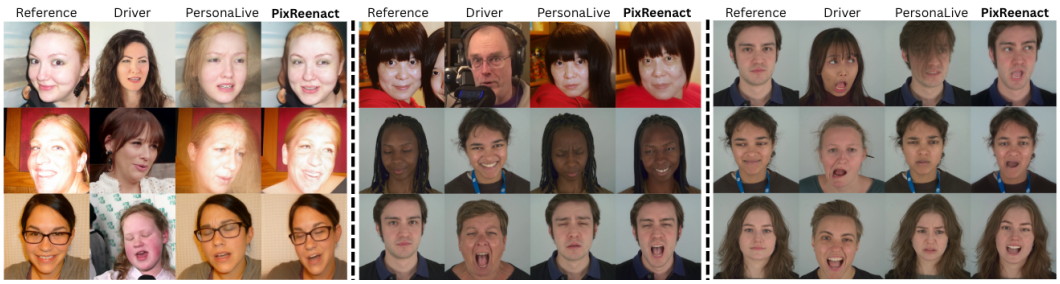

Figure 4: **Qualitative cross-identity reenactment.** PixReenact better preserves reference identity while transferring driver facial motion on the public and NeRSemble benchmarks.

Table 2: **Cross-identity reenactment** on public (75 pairs) and NeRSemble (70 pairs) benchmarks. Single-H200 runtime; latency is mean steady-state emission time. Better streaming results are bold.

		Public benchmark (75 pairs)					**NeRSemble (70 pairs)**					**Runtime**	
	Method	ID↑	AED↓	APD↓	tLP↓	FVD↓	ID↑	AED↓	APD↓	tLP↓	FVD↓	Lat.↓	FPS↑
Per-frame	LivePortrait	0.863	1.071	0.208	1.146	579.0	0.849	0.787	0.0840	0.763	302.4	–	18.90
	X-Portrait	0.701	0.887	0.097	1.613	595.2	0.774	0.694	0.0461	0.795	281.2	–	0.65
Offline	MegActor-Σ	0.554	0.841	0.106	1.910	582.7	0.729	0.802	0.0566	0.976	247.4	–	0.49
	Wan-Animate	0.570	0.966	0.089	1.903	579.2	0.632	0.789	0.0381	1.199	201.3	–	0.41
	X-NeMo	0.614	0.775	0.065	1.378	449.1	0.710	0.639	0.0291	0.733	276.5	–	1.37
	Wan-Animate-2	0.603	0.863	0.085	4.500	655.8	0.583	0.723	0.0451	1.831	276.5	–	0.29
	FlashPortrait	0.760	0.864	0.112	3.072	633.7	0.800	0.758	0.0598	1.527	303.2	–	3.03
Streaming	PersonaLive	0.662	**0.788**	**0.050**	**1.059**	407.1	0.734	0.824	**0.0392**	0.890	310.0	**211.1**	**18.95**
	PixReenact	**0.709**	0.790	0.066	1.234	**377.1**	**0.753**	**0.714**	0.0426	**0.803**	**264.8**	238.9	16.74

Appendix Table 14 ablates DMD, pseudo-target filtering, self-driving data, and reconstruction loss.

5 Conclusion

We introduce **PixReenact**, a pixel-conditioned framework for low-latency streaming head-avatar reenactment. Across three cross-identity benchmarks and a long-horizon streaming benchmark, PixReenact demonstrates robust cross-identity reenactment under extreme viewpoints, occlusions, and pronounced facial expressions, while maintaining reference identity over five-minute streams. Its 4-NFE rolling formulation continuously emits four frames at 239 ms mean latency.

Challenges remain. Extreme viewpoints and severe occlusions still challenge streaming reenactment, leaving room to improve motion fidelity and visual quality. Cross-identity training uses costly pseudo targets from an offline model, whose errors leave supervision imperfect despite corrective objectives. Reducing this reliance is important future work. More broadly, our results support direct visual conditioning with causal video generation for robust, low-latency streaming head avatars.

Table 3: **Component ablations** on the 75-pair public cross-identity benchmark. All variants continue from the same checkpoint for 5K steps. The gray row is the matched full recipe. IDD: driver-identity similarity. †Better motion scores may reflect driver copying with worse reference identity.

	Variant	ID↑	IDD↓	AED↓	APD↓	tLP↓	FVD↓
	Full recipe (continuation baseline)	0.707	0.023	0.816	0.071	1.254	389
Pixel cond.	Driver attention $w = 0$†	0.585	0.210	0.679	0.067	1.690	273
Cross-ID	w/o cross-identity supervision	0.768	0.020	0.852	0.087	1.156	458
	w/o identity losses†	0.353	0.245	0.629	0.057	1.638	293
	w/o motion losses	0.795	0.011	0.901	0.090	1.171	548
State-aware	w/o cold-start teacher	0.721	0.028	0.834	0.079	1.188	406

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on research methodology or experiments, but they were not used to determine the methodology and experimental design. Additionally, we used generative AI tools for editing and debugging software code, creating and refining scientific figures, drafting and editing parts of the manuscript to improve readability, summarizing and analyzing existing literature, searching for and identifying relevant literature, sourcing information, suggesting manuscript structure, and formatting references. All citations were carefully examined and added by the authors to avoid hallucinations.

We did not use LLMs or agentic AI tools to generate synthetic datasets. Our pseudo targets were generated by running the released Wan-Animate and PersonaLive implementations, as described in the paper. We have not used generative AI tools to develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine research hypotheses, implement methods, clean or reformat research datasets, or interpret results. Translation and qualitative or thematic data analysis were not applicable to this work.

We have reviewed all AI-assisted work. AI-assisted code was tested for correctness, scientific figures were checked against the underlying data and methodology, and literature summaries were verified against the original sources. We take responsibility for the final content of this work, including all text, claims, code, figures, results, and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our work studies head-avatar reenactment from images and videos of real people. We use established research-oriented datasets, including NeRSemble Kirschstein et al. (2023), VFHQ Xie et al. (2022), DH-FaceVid-1K Di et al. (2025), FFHQ Karras et al. (2019), TalkingHead-1KH Wang et al. (2021), HDTF Zhang et al. (2021), and CelebV-HQ Zhu et al. (2022), only for the research purposes described in this paper. For long-horizon evaluation, we additionally use 10 publicly available YouTube videos. We do not collect private personal data or infer sensitive attributes. Pretrained face-recognition models are used only to measure and supervise identity preservation, rather than to identify individuals. As with other face-video benchmarks, the demographic composition of the available data may be imbalanced, and we do not claim uniform performance across demographic groups or other population subgroups.

Head-avatar reenactment can support beneficial applications such as telepresence and interactive digital agents, but it can also be misused to create deceptive or non-consensual representations of real people. PixReenact is a research system and is not designed to verify consent, authenticity, or ownership of its inputs. Deployment should therefore require appropriate consent and authorization for the identities being animated, and generated content should be clearly disclosed as synthetic when it could otherwise be mistaken for authentic footage. We encourage future work on provenance, watermarking, and other safeguards for responsible deployment. The authors take responsibility for the research practices and experimental results reported in this work.

REPRODUCIBILITY STATEMENT

We provide extensive details to support reproducibility of PixReenact and its evaluation. The main paper (Sec. 3) describes the model architecture, conditioning scheme, rolling denoising procedure, training objectives, and teacher-student distillation framework, while Section 4.1 specifies the training datasets, evaluation benchmarks, baselines, metrics, and hardware. The Appendix provides further implementation details, including cross-identity pseudo-target construction and filtering B.1, model and LoRA configuration B.2, streaming-state construction B.3, dense-teacher B.4 and student training procedures B.5 B.6, optimization settings B.7, runtime measurement protocol B.8, benchmark construction B.9, baseline configurations B.10, and evaluation procedures B.11. We additionally provide the exact clips used for the extreme-condition benchmark (Tab. 8) and detailed supplementary ablations and evaluation protocols (Sec. C). We plan to release code to facilitate reproduction of the reported results.

REFERENCES

Gang Cheng, Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Ju Li, Dechao Meng, Jinwei Qi, Penchong Qiao, et al. Wan-animate: Unified character animation and replacement with holistic replication. *arXiv preprint arXiv:2509.14055*, 2025.

Mengyu Chu, You Xie, Jonas Mayer, Laura Leal-Taixé, and Nils Thuerey. Learning temporal coherence via self-supervision for gan-based video generation. *ACM Transactions on Graphics (TOG)*, 39(4):75–1, 2020.

Xuangeng Chu and Tatsuya Harada. Generalizable and animatable gaussian head avatar. *Advances in Neural Information Processing Systems*, 37:57642–57670, 2024.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.

Donglin Di, He Feng, Wenzhang Sun, Yongjia Ma, Hao Li, Wei Chen, Lei Fan, Tonghua Su, and Xun Yang. Dh-facevid-1k: A large-scale high-quality dataset for face video generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12124–12134. IEEE, 2025.

Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021.

Songwei Ge, Aniruddha Mahapatra, Gaurav Parmar, Jun-Yan Zhu, and Jia-Bin Huang. On the content bias in fréchet video distance. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7277–7288. IEEE, 2024.

Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18632–18643. IEEE, 2022.

Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. Liveportrait: Efficient portrait animation with stitching and retargeting control. *arXiv preprint arXiv:2407.03168*, 2024.

Mingtao Guo, Guanyu Xing, and Yanli Liu. High-fidelity relightable monocular portrait animation with lighting-controllable video diffusion model. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 228–238. IEEE, 2025.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 167283–167308. Curran Associates, Inc., 2025. doi: 10.52202/085713-5576.

Chaonan Ji, Jinwei Qi, Sheng Xu, Peng Zhang, and Bang Zhang. Portraitdirector: A hierarchical disentanglement framework for controllable and real-time facial reenactment. *arXiv preprint arXiv:2604.19129*, 2026.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 4396–4405. IEEE, 2019.

Taekyung Ki, Sangwon Jang, Jaehyeong Jo, Jaehong Yoon, and Sung Ju Hwang. Avatar forcing: Real-time interactive head avatar generation for natural conversation. *arXiv preprint arXiv:2601.00664*, 2026.

Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 18729–18738. IEEE, 2022.

Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Transactions on Graphics (TOG)*, 42(4):1–14, 2023.

Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017.

Zhiyuan Li, Chi-Man Pun, Chen Fang, Jue Wang, and Xiaodong Cun. Personalive! expressive portrait image animation for live streaming. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18118–18128, 2026.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. In *International Conference on Learning Representations*, volume 2026, pp. 91177–91196, 2026.

Mingshuang Luo, Shuang Liang, Zhengkun Rong, Yuxuan Luo, Tianshu Hu, Ruibing Hou, Hong Chang, Yong Li, Yuan Zhang, and Mingyuan Gao. Dreamactor-m2: Universal character image animation via spatiotemporal in-context learning. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–11, 2026.

George Retsinas, Panagiotis P Filntisis, Radek Danecek, Victoria F Abrevaya, Anastasios Roussos, Timo Bolkart, and Petros Maragos. 3d facial expressions through analysis-by-neural-synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2490–2501, 2024.

Dvir Samuel, Issar Tzachor, Matan Levy, Michael Green, Gal Chechik, and Rami Ben-Ari. Fast autoregressive video diffusion and world models with temporal cache compression and sparse attention. *arXiv preprint arXiv:2602.01801*, 2026.

Yuxiang Shi, Zhe Li, Yanwen Wang, Hao Zhu, Xun Cao, and Ligang Liu. Dex-portrait: Disentangled and expressive portrait animation via explicit and latent motion representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40009–40019, 2026.

Wan Team, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2(3):4, 2025.

Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence*, 3(1):42–50, 2021.

Shuyuan Tu, Yueming Pan, Yinming Huang, Xintong Han, Zhen Xing, Qi Dai, Kai Qiu, Chong Luo, and Zuxuan Wu. Flashportrait: 6x faster infinite portrait animation with adaptive latent prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25163–25173, 2026.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. In *ICLR 2019 Workshops*, 2019.

Guangyuan Wang, Li Hu, Dechao Meng, Zhongyi Zhang, Peng Zhang, Mingyang Huang, Ruoshi Zhang, Ke Sun, Zhe Zhang, Xingjun Wang, et al. Wan-animate-2: Pushing the application boundaries of character animation. *arXiv preprint arXiv:2608.06009*, 2026.

Jianfeng Wang, Ye Yuan, and Gang Yu. Face attention network: An effective face detector for the occluded faces. *arXiv preprint arXiv:1711.07246*, 2017.

Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10034–10044. IEEE, 2021.

Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 656–665. IEEE, 2022.

You Xie, Hongyi Xu, Guoxian Song, Chao Wang, Yichun Shi, and Linjie Luo. X-portrait: Expressive portrait animation with hierarchical motion attention. In *ACM SIGGRAPH 2024 conference papers*, pp. 1–11, 2024.

Zunnan Xu, Zhentao Yu, Zixiang Zhou, Jun Zhou, Xiaoyu Jin, Fa-Ting Hong, Xiaozhong Ji, Junwei Zhu, Chengfei Cai, Shiyu Tang, et al. Hunyuanportrait: Implicit condition control for enhanced portrait animation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15909–15919. IEEE, 2025.

Wenhao Yan, Fengjia Guo, Zhuoyi Yang, and Jie Tang. Scail-2: Unifying controlled character animation with end-to-end in-context conditioning. *arXiv preprint arXiv:2606.10804*, 2026.

Shurong Yang, Huadong Li, Juhao Wu, Minhao Jing, Linze Li, Renhe Ji, Jiajun Liang, and Haoqiang Fan. Megactor: Harness the power of raw video for vivid portrait animation. *arXiv preprint arXiv:2405.20851*, 2024.

Shurong Yang, Huadong Li, Juhao Wu, Minhao Jing, Linze Li, Renhe Ji, Jiajun Liang, Haoqiang Fan, and Jin Wang. Megactor-sigma: Unlocking flexible mixed-modal control in portrait animation with diffusion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 9256–9264, 2025.

Fei Yin, Vikram Voleti, Nikita Drobyshev, Maksim Lapin, Aaryaman Vasishta, Varun Jampani, et al. Stable video-driven portraits. *arXiv preprint arXiv:2509.17476*, 2025a.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024.

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, 2025b.

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

Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3660–3669. IEEE, 2021.

Xiaochen Zhao, Hongyi Xu, Guoxian Song, You Xie, Chenxu Zhang, Xiu Li, Linjie Luo, Jinli Suo, and Yebin Liu. X-nemo: Expressive neural motion reenactment via disentangled latent attention. In *International Conference on Learning Representations*, volume 2025, pp. 77070–77088, 2025.

Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. Celebv-hq: A large-scale video facial attributes dataset. In *European conference on computer vision*, pp. 650–667. Springer, 2022.

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 Supplementary Videos

We provide a supplementary `.zip` archive containing video results that complement the qualitative evaluations in the main paper and Appendix. The archive is organized into four folders corresponding to the evaluation settings used throughout the paper.

Contents.

- `extreme_conditions`: side-by-side comparisons of PixReenact and PersonaLive under large viewpoint changes and challenging occlusions, corresponding to the extreme-condition benchmark in the main paper (Sec. 4.2).

- `long_videos`: long PixReenact streaming outputs illustrating appearance and identity stability over extended autoregressive generation (Sec. 4.2).
- `standard`: side-by-side comparisons of PixReenact and PersonaLive on the standard public and NeRSemble cross-identity benchmarks (Sec. 4.3).
- `self_reenactment`: side-by-side comparisons of PixReenact and PersonaLive in the self-reenactment setting. These results complement the self-reenactment evaluation reported later in the Appendix C.

Video encoding. To meet the supplementary size limit, videos are re-encoded with H.264 using CRF 26-27. Relative to the original renders, the compressed videos retain PSNR ≥ 39.6 dB and SSIM ≥ 0.96. The same encoding settings are applied to all method outputs, and all quantitative results are computed on the original outputs.

Usage. Extract the supplementary `.zip` archive and open any of the included `.mp4` files with a standard video player or modern web browser. The folder names follow the evaluation settings described above, and each comparison video identifies the corresponding inputs and method outputs.

B Implementation and Evaluation Details

Unless stated otherwise, all details refer to the released *wan-labels* model. The *PL-labels* variant differs only in the pseudo-target generator, PersonaLive Li et al. (2026) instead of Wan-Animate Cheng et al. (2025), and in the filtering thresholds noted below.

B.1 Training Data Construction

Pseudo-pair construction. Each cross-identity training example pairs a reference image with a 4-second, 512^2 driver clip of a different identity. For NeRSemble Kirschstein et al. (2023), VFHQ training set Xie et al. (2022), and DH-FaceVid-1K Di et al. (2025), references and drivers are drawn from the corresponding training source. For FFHQ, we evenly sample 800 references from indices 00050-69999 and pair each with three drivers, yielding 2,400 candidate pairs before filtering. The pseudo target is generated offline by Wan-Animate Cheng et al. (2025) (20 denoising steps) at 512^2. The PL-labels variant instead generates targets with PersonaLive Li et al. (2026) in its streaming mode on the same pairings.

Pseudo-target filtering. Pseudo targets are filtered for *label quality* using identity, motion, and emotion consistency criteria: expression and pose agreement with the driver (AED ≤ 0.25, APD ≤ 0.18), emotion agreement (EmoSim ≥ 0.7), and an identity-routing margin $\mathrm{ID}(\tilde{Y}, I_{\mathrm{ref}}) - \mathrm{ID}(\tilde{Y}, D) \geq 0$ that rejects pseudo targets that copied the driver's identity. We never filter on absolute identity similarity alone, which would remove mainly hard pairs in which the reference image and driver pose differ. PL labels exhibit a stronger driver-copy failure mode, so their thresholds are stricter on emotion (EmoSim ≥ 0.84) and margin (≥ 0.1), while the AED/APD thresholds are identical. Table 4 reports per-dataset counts. Filtering metrics are computed with the training-time estimators (Sec. B.6): DECA expression (50-D) and head-rotation ℓ_1 for AED/APD, EmoNet feature cosine for EmoSim, and ArcFace cosine for the identity terms, each evaluated per frame on FAN face crops Wang et al. (2017) and averaged over the clip (full-frame resize when the detector misses a frame). If no face is detected in the pseudo target at all, its identity terms are undefined and the pair fails the margin gate, i.e., it is discarded.

Real self-driving clips. 1,548 real 4-second clips (402 NeRSemble, 146 VFHQ, 1,000 DH-FaceVid-1K). The first frame is the reference image and the clip is both driver and ground truth. During student training, each step draws a self-driving sample with probability of 60% and a cross-identity pseudo-pair otherwise.

B.2 Model Architecture

We use Wan2.1-T2V-1.3B Team et al. (2025) as the bidirectional video-diffusion backbone (30 transformer blocks, hidden dimension 1,536) and adopt CausVid's Yin et al. (2025b) causal gener-

Table 4: Pseudo-pair counts before (scored) and after quality filtering.

	wan-labels		PL-labels	
Source	scored	kept	scored	kept
NeRSemble	4,020	3,182	4,012	3,061
VFHQ	1,450	1,106	1,447	1,084
DH-FaceVid-1K	10,000	7,990	9,994	7,686
FFHQ	2,400	1,283	2,400	1,642
Total	17,870	13,561	17,853	13,473

ation framework. The model uses the Wan-VAE ($8\times$ spatial, $4\times$ temporal compression; one latent spans four frames), and all experiments run at 512^2. The pretrained backbone weights and VAE are frozen. We train LoRA adapters (rank 64, $\alpha = 64$) in the self-attention projections of all 30 blocks, together with the conditioning interface. Conditioning follows Sec. 3.2: one identity latent, window-aligned driving latents, and two context latents, with the restricted attention mask described there (driving window w=1). Text prompts are not used.

B.3 Streaming Configuration

The students' rolling staircase uses four noise levels $t = [1000, 750, 500, 250]$. Each forward pass admits one fresh noise latent at t=1000 (seed 42), then advances all in-flight latents by one level, and emits the t=250 latent to t=0 (four frames). Context is the two most recent emitted latents. Conditioning keys/values (identity and driving tokens) are KV-cached across passes. Decoding of emitted latents uses the streaming VAE decoder with a persistent temporal cache. Streams start with a ramp state: the context slots hold causal-VAE latents of a short still clip of the reference image, and the first emission is aligned to driver frame 0, so outputs are frame-aligned to the driver over the full stream. Driver frames are face-aligned 512^2 crops encoded by the causal VAE. At inference time, encoding of the next driving chunk overlaps with denoising of the current chunk.

B.4 Dense Score Teachers

Steady-state teacher. We train rank-64 LoRA adapters on the pretrained bidirectional Wan2.1-T2V-1.3B backbone under the causal attention mask, using the deployed rolling geometry (6-latent windows: 2 context + 4 video latents). The adapters are trained for 27,500 steps from initialization (lr 2×10^{-4}), followed by 10,000 adaptation steps (lr 10^{-4}) on the filtered pseudo-pairs with 10% self-driving mixing. Batch size is 4 per GPU on 4 NVIDIA H200 GPUs. Noise states mix 60% staircase-patterned windows ($t = [0, 0, 250, 500, 750, 1000]$) and 40% dense timesteps $t \sim \mathcal{U}(0.02, 1.0)$, so the teacher covers both inference-like mixed-noise states and arbitrary re-noise levels queried during distillation. Both teachers retain the backbone's native flow-velocity parameterization but are supervised with MSE on the implied clean-latent estimate $\hat{z}^0 = z^t - \sigma_t \hat{v}^t$ against the clean target z^0.

Cold-start teacher. Initialized from the steady-state teacher and fine-tuned for 3,500 steps (lr 10^{-4}, batch 4) exclusively on first-emission states: windows whose context is the reference-initialized still clip, over cross pairs with large reference-driver pose gaps (plus 10% self-driving transport, where the reference and context come from a still of the same person as the real driver clip). Timesteps emphasize high noise ($t \geq 0.8$ with probability of 90%), matching the states at which this teacher is queried (first emission from high-noise).

B.5 Student Distillation

The student is trained through long self-rollouts spanning the full 4-second driver clip (30 latents, 120 frames), giving ≈28 consecutive emissions from a single cold start. Rollouts are always opened from the deployment cold-start state (reference-initialized context). Losses are applied to emitted latents only. For DMD, the emitted latent is re-noised to $t \sim \mathcal{U}(0.02, 0.98)$ with high-noise emphasis ($t \geq 0.6$ for steady-state emissions; $t \geq 0.8$ with probability of 85% for the first emission). The

real score is the cold-start teacher at the first emission and the steady-state teacher afterwards. The DMD term is scaled $\times 3$ on first-emission states. The online fake-score critic is initialized from the same checkpoint as the student and updated three times per generator update. The released model is obtained by successive DMD stages: an initial 30,000-update distillation from the dense teacher followed by shorter refinement stages of the same objective family. The final wan-labels stage (1,000 updates) uses exactly the configuration reported here.

B.6 Training Objectives

Table 5 lists all loss terms, their weights, and the samples they apply to during student training. Pixel-space losses are computed on frames decoded jointly with the preceding latent (matching deployment decoding statistics). **Training-time estimators:** ArcFace R100 (Glint360k) Deng et al. (2019) for the identity and anti-leakage terms; DECA Feng et al. (2021) (FLAME Li et al. (2017) coefficients) on FAN-cropped faces for expression and head pose Wang et al. (2017); EmoNet Toisoul et al. (2021) (8-class, 128-D features) for emotion. On first emission (cold-start) states, the identity/anti-leak/anchor group is scaled $\times 1.5$ and the expression/emotion group $\times 0.6$.

Table 5: Student loss terms. *self* refers to real self-driving samples, *cross* to cross-identity pseudo-pairs.

Term	Definition	Weight	Applies to
$\mathcal{L}_{\text{DMD}}$	surrogate of Eq. 2 (main paper)	1.0 ($\times 3$ cold)	all
$\mathcal{L}_{\text{anc}}$	ℓ_2 to target latent	0.3	all
$\mathcal{L}_{\text{id}}$	$\max(0,\, 0.8 - \cos(\phi(Y_k), \phi(I_{\text{ref}})))$	1.0	all[a]
$\mathcal{L}_{\text{leak}}$	$\max(0,\, \cos(\phi(Y_k), \phi(D_k)) - 0.25)$	1.0	cross
$\mathcal{L}_{\text{pose}}$	ℓ_1 DECA global rotation vs. driver	1.2	all
$\mathcal{L}_{\text{expr}}$	ℓ_1 DECA expression+jaw vs. driver	0.8	all
$\mathcal{L}_{\text{emo}}$	$1 - \cos$ EmoNet features vs. driver	0.5	all
$\mathcal{L}_{\text{rec,LPIPS}}$	vs. ground-truth frames	1.0	self
$\mathcal{L}_{\text{rec,Adversarial}}$	PatchGAN hinge (disc. lr 10^{-4})	0.1	self

[a] For cross-identity samples, the identity anchor is the reference image I_{ref}, while for self-driving samples it is the aligned real ground-truth frame.

B.7 Optimization

We train using the Adam optimizer with β_1=0, β_2=0.999, and weight decay of 10^{-4}. The generator and critic learning rates are 10^{-5} and the discriminator uses AdamW with learning rate 10^{-4}. All learning rates are constant, with no warm-up or decay schedule and no gradient accumulation. The critic takes three updates per generator update, and the discriminator takes one update per generator step on which the adversarial loss is active (real self-driving steps). No global training seed is fixed, while all inference and evaluation use seed 42. The trainable parameters are the LoRA adapters and conditioning interface, 17.7M parameters in total (≈1.3% of the 1.3B backbone). The final student stage uses batch size 1 per GPU on 4 NVIDIA H200 GPUs, bfloat16 mixed precision, and gradient checkpointing.

B.8 Runtime Measurement Protocol

Runtime measurement protocol. All runtime measurements use a single NVIDIA H200 GPU. Per-stage latencies (Tab. 6) are measured on the first 10 drivers of the public cross-reenactment benchmark. Before timing, we perform one untimed warm-up invocation to stabilize GPU clocks and caches. Each PixReenact clip is then measured in a fresh process. Therefore, we additionally exclude the startup ramp and first steady-state chunk, which absorbs one-time kernel compilation costs. For each video, we report the median latency over the remaining steady-state chunks and aggregate these values as mean ± std across the 10 videos.

Emission latency measures the time from admitting a chunk's noise latent until its four output frames are decoded. Throughput is the corresponding steady-state rate, computed as $\text{FPS} = 4/\text{latency}$. Driver-VAE encoding is excluded because, in streaming use, encoding of the next driving chunk

Table 6: **Per-stage runtime of the two streaming pipelines** (per emitted 4-frame chunk, 512^2, steady state, single H200; mean $\pm$ std over 10 drivers). FPS is computed as $4/$total emission latency. Startup and one-time kernel warm-up are excluded. Driver VAE encode (*italic*) is excluded from our emission latency: benchmark drivers enter as precomputed latents, and in live use the next chunk's encode overlaps the current chunk's denoising. PersonaLive's reference-UNet refresh fires at most three times per stream and is amortized.

Pipeline	Stage	Lat. ↓
PixReenact	Window assembly (ctx + cond concat)	0.1 ±0.0
	Denoising DiT forward (4-level staircase, 1 pass)	155.1 ±0.2
	Flow→ x_0 + re-noise (scheduler)	0.7 ±0.1
	Streaming VAE decode (4 frames, cached)	82.9 ±0.5
	Buffer bookkeeping (shift context/buffers)	0.1 ±0.0
	Total emission latency	**238.9** ±0.7
	Generation throughput (FPS)	**16.74**
	Driver VAE encode (streaming, amortized)	*45.5* ±0.0
PersonaLive	Keypoint extraction	5.3 ±0.1
	Face crop (4 frames)	0.3 ±0.0
	Motion encoder	6.0 ±0.1
	Pose guider	1.1 ±0.0
	Denoising UNet (16-frame window, 1 step)	133.6 ±3.8
	Scheduler step	0.3 ±0.0
	Reference-UNet keyframe refresh (amortized)	0.1 ±0.1
	VAE decode (4 frames)	61.2 ±0.0
	Host ops (resize, kps render, GPU→CPU)	2.8 ±0.0
	Total emission latency	**211.1** ±3.6
	Generation throughput (FPS)	**18.95**

Table 7: **Quality-latency trade-off.** Quality metrics are measured on the 68-pair extreme-condition benchmark, while latency and FPS use the first 10 public-benchmark drivers under the same runtime protocol for all methods. Fewer NFEs reduce latency but generally degrade motion and distribution quality. Best and second-best results are shown in **bold** and underline, respectively.

Method / NFE	ID↑	AED↓	APD↓	tLP↓	FVD↓	Lat. ↓	FPS↑
PersonaLive	0.472	1.295	**0.135**	1.866	**643.5**	211.1	18.95
PixReenact, 4 (1000,750,500,250)	0.727	**1.135**	0.169	**1.853**	662.5	238.9	16.74
PixReenact, 3 (1000,500,250)	0.731	1.145	0.174	2.008	682.9	211.4	18.92
PixReenact, 2 (1000,250)	**0.770**	1.191	0.187	1.986	753.6	**186.6**	**21.44**

overlaps with denoising of the current chunk. Its cost is reported separately in Tab. 6. PersonaLive is evaluated under the same steady-state timing definition using its official streaming pipeline.

Quality-latency trade-off. We additionally vary the number of denoising steps while keeping all other inference settings fixed. Table 7 compares these schedules with PersonaLive under the same runtime protocol on the 68-pair extreme-condition benchmark. Reducing PixReenact from 4-NFE to 3-NFE and 2-NFE lowers latency from 238.9 ms to 211.4 ms and 186.6 ms, respectively, but generally degrades motion and distribution quality. Notably, 3-NFE has nearly identical latency to PersonaLive (211.4 vs. 211.1 ms), while the 4-NFE setting provides the strongest overall PixReenact quality. We therefore use 4-NFE throughout the main experiments.

B.9 Benchmark Construction

Public cross-reenactment (75 pairs, Table 2). Drivers are the 75 self-reenactment clips in fixed order (clips 1-25 = TalkingHead-1KH, 26-75 = VFHQ-Test). The reference image for pair i is FFHQ image $(i-1) \bmod 50$, so each of the first 50 FFHQ images appears once or twice.

Table 8: **Driver clips of the extreme-condition benchmark.** The 68 CelebV-HQ clip identifiers by group.

Group	CelebV-HQ clip IDs
Extreme yaw (25)	-3xE76Q0yLs_6_3, G4f_opQO7MA_4_0, I66kXy4rvME_3, Lrpy2CCnJR0_17, NkjnK6aSrqM_4, QgedeMGTh5E_9, RyevycwXzn4_7, SnBObwk5JF0_37, VCFAT8Hj7Qo_0_0, WGvhuhHYhiA_14, YXetIu8JOzI_3, cWTK6ttZUw0_7, dhSpE_2OF40_9_0, ecsatgCN6Y8_1, hGsR1fkUJqY_10, hOB4Qm1IiOY_0, i1UEj_6T1RE_1, iViIy5ci5yU_24, mpVMM7eqTKs_8_0, sb6b4r8no1w_14_1, t7BPFTnDsJc_1, x99FBmR6OrI_0_0, xGoRE2NsRGI_9, yprYw3FQUpQ_0, zW2T9TkSIcM_5
Extreme pitch (17)	-kFJKQvCrkY_1, 7NHQrVt1yOU_2, 8w0b64kOPss_0, 9DTvF8iBli8_1_3, AcvWJ8bgA8w_3, Fn40RI8ving_1, Jipg0KQ4id0_3, N4XB-3O7UTk_0_0, PFBrK7pbqCU_18, RLbry-3z8yQ_2, Yg6sZ2htZfc_1, dildzih2RPI_3_0, jeycN3BAkt8_38_0, mLh4ZKG50mk_0_0, od6IxUWPMcs_0, qAHkYcmcFfc_6, tnxXQgmHwrk_2
Accessory occlusion (16)	7FyjCUDR0IM_17, CPKMurScLSE_15_0, Fn40RI8ving_7, HK6TmhLEd2E_12, LjJ276Dh1jA_44, Tn0Ge0GIoGA_10, U8PnHiRRg-s_6, XDmfBTgIRUg_12, XDmfBTgIRUg_15, _wiIxAcoTF4_4, i4ivf-ZmorE_28, qs7FL5DVDMA_10, unmu4yKfBg0_9, xuphB-zbmPk_26, yLElO7gxXPo_43, z7VS7DVHTbc_1_0
Hand occlusion (10)	2Ke4u7vbLE0_0, M9FAInQwCm0_3, NgtK-eesI7g_0, PP9l4LP0WPI_4, Q9B0pFDCviw_102, dfrJhivMJJY_0, hSiFAK-p-Lg_20, v60fsgGSoDU_0, xu_UgfGriXA_3, yth5JoZKnGQ_1

NeRSemble (75 total pairs: 70 cross + 5 self, Tables 2 and 11). The test pool contains 15 held-out subjects (NeRSemble IDs 82, 89, 94, 99, 100, 133, 145, 186, 371, 383, 388, 422, 443, 445, 475), identity-disjoint from training. Five subjects (388, 422, 443, 445, 475) provide the reference stills, and each reference is driven by a clip of every one of the 15 subjects: the five driver=reference combinations form the self-reenactment set and the remaining 70 the cross set. Clips are center-square cropped, temporally resampled to 30 fps, and evaluated at 512^2.

Extreme-condition (68 pairs, Table 1). Candidate CelebV-HQ clips are ranked using the dataset's attribute annotations, scoring appearance attributes that occlude the face (mask, sunglasses, hat) and actions that produce large or fast head motion or hand/object occlusion (turn, look around, head wagging, head shake, nod, drink, eat, smoke, play instrument, make a face). Ranked candidates are then verified automatically (a head-pose probe records the median |yaw|/|pitch| over each clip's hardest span, and a MediaPipe FaceMesh scan records per-frame detection misses) and reviewed manually. The final 68 drivers are grouped into four groups: extreme yaw (25), extreme pitch (17), accessory occlusion (16), and hand occlusion (10). References are FFHQ images assigned in the public cross-reenactment benchmark order, and drivers are center-square cropped and resized to 512^2. Table 8 lists all 68 driver clips.

Long-horizon streaming (28 pairs, Figure 3). To evaluate identity stability over long streams, we construct a benchmark of 28 continuous 5-minute cross-identity clips. The drivers comprise 10 long-form Barack Obama videos collected from YouTube, 15 videos from the WDA subset of HDTF Zhang et al. (2021), and 3 long clips extracted from the raw TalkingHead-1KH validation videos Wang et al. (2021). Each driver is paired with a fixed FFHQ reference image, following the reference order of the public cross-reenactment benchmark, with driver and reference identities always different. Each method processes the full 5-minute sequence continuously, without resetting its temporal state. We measure AdaFace similarity to the reference throughout the stream and aggregate the results over five one-minute intervals. For each clip, scores are first averaged within each minute and then averaged across clips.

B.10 Baseline Implementation Details

All baselines run from their official repositories, released checkpoints, and default configurations, with their own native preprocessing. We list only the deviations required for full-length, frame-aligned evaluation. **LivePortrait**: official release; per-frame animation with its whole-clip tracking pass; no modifications. **Wan-Animate**: official 14B release at 20 denoising steps; native 77-frame windows with the repository's sliding-window continuation for full-length clips. **Wan-Animate-2-Base**: official release (37 frames per pass); full-length outputs are produced by sequential client-side chunking, as the release provides no long-video continuation. **X-NeMo**: official release with its native long-video context scheduler (24-frame windows, overlap 4); evaluated in its native crop space against identically cropped references. **PersonaLive**: official streaming (online) pipeline, unmodified. **X-Portrait**: official checkpoint; the required per-pair best-frame anchor is selected automatically by matching driver head pose to the reference via FAN landmarks. **MegActor-Σ**: official weights and flags (25 steps, guidance 2); the default 2.5-second generation limit is raised to the full clip length. **FlashPortrait**: official checkpoint on Wan2.1-I2V-14B with a fixed generic text prompt for all clips; its scheduler requires $\geq$51 latents, so shorter clips cap the context accordingly, and the official output drops the first frame and truncates to $4k+1$ frames, which we restore by duplicating $\leq$3 boundary frames. For every method, outputs are fitted to exactly the driver frame count by frame-indexed rewriting when the difference is $\leq$3 frames (float-fps container rounding); larger mismatches abort the sample rather than re-time the video.

B.11 Evaluation Protocol

All methods are scored on the same evaluation clips at 512^2, full length, with outputs frame-aligned to the driver (frame i of every method corresponds to driver frame i). **ID / IDD:** cosine similarity of AdaFace Kim et al. (2022) (CVLface IR-101, WebFace12M) embeddings between generated frames and the reference image (ID) or the corresponding driver frame (IDD). Faces are detected, aligned and cropped to 112^2, as in the standard AdaFace pipeline. For speed, we do not embed every frame of the clip, but use $\sim$60 uniformly sampled frames. **AED / APD:** SMIRK Retsinas et al. (2024) codes on face crops (FAN bounding box, DECA-convention crop, 224^2) computed for every frame. AED is the ℓ_1 distance over expression, eyelid, and jaw parameters (55-D) and APD over the 3 head-pose parameters, between generated and driver frames. **tLP:** Chu et al. (2020) mean over consecutive-frame pairs of $|d(Y_t, Y_{t+1}) - d(R_t, R_{t+1})| \times 100$, where d is LPIPS Zhang et al. (2018) (AlexNet) and R is the reference sequence (ground truth for self-reenactment, driver for cross). **FVD:** Unterthiner et al. (2019) set-level with an I3D backbone, where each clip contributes up to eight 16-frame windows (stride 16) spread uniformly over the clip, resized to 224^2. Real statistics come from the ground-truth (self) or driver (cross) clips. **PSNR / SSIM / LPIPS / ℓ_1:** full-frame at 512^2 against ground truth, self-reenactment only. X-NeMo Zhao et al. (2025) is evaluated in its native crop space against identically cropped references. Quality tables report the per-clip mean ($\pm$ std where shown), while FVD is a single set-level value.

B.12 Limitations of Evaluation Metrics

The metrics used in the main paper capture complementary aspects of reenactment quality, but no individual metric fully measures successful cross-identity reenactment. In particular, identity preservation and motion transfer must be considered jointly: reproducing the driver's face more directly can improve agreement with the driving motion while failing to preserve the reference identity.

AED and APD measure distances between expression and pose parameters estimated by SMIRK Retsinas et al. (2024), rather than distances to ground-truth facial motion. This distinction is especially relevant for our extreme-condition benchmark in Table 1. SMIRK reports sensitivity to occlusions and shows degraded reconstructions under occlusions and extreme head poses. Consequently, APD and AED estimates may become less reliable precisely on some of the most challenging frames in this benchmark. FVD Unterthiner et al. (2019) also has known limitations as a measure of temporal video quality. Prior work has shown that FVD can be strongly influenced by individual-frame appearance and can be relatively insensitive to temporal corruptions Ge et al. (2024). Thus, a lower FVD does not necessarily imply better motion transfer or temporal realism in isolation.

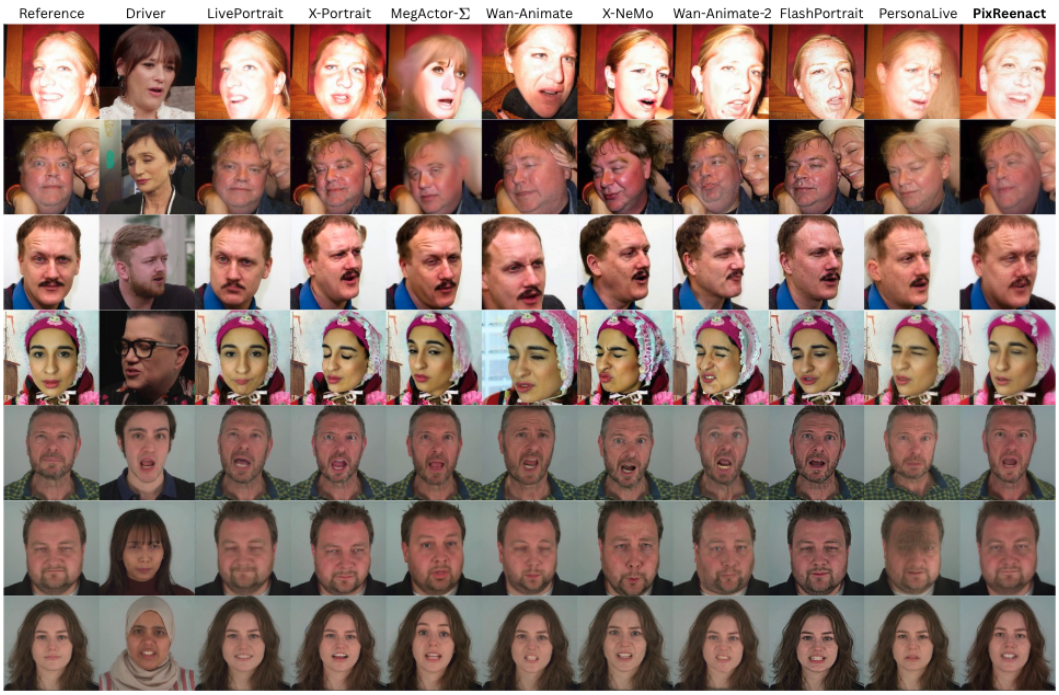

Figure 5: **Additional cross-identity reenactment examples.** Qualitative comparisons on the public (rows 1-4) and NeRSemble (rows 5-7) benchmarks against all evaluated baselines. PixReenact preserves the reference appearance while transferring the driver's facial motion.

Our component ablations provide an additional example specific to cross-identity reenactment. Removing the identity losses, or restricting the driving attention to only the aligned latent, increases driver-identity leakage while improving several motion or distribution metrics (Table 3). These improvements do not reflect better cross-identity reenactment: the generated video becomes more similar to the driver at the expense of preserving the reference identity. We therefore interpret ID, AED/APD, tLP, and FVD jointly and complement the quantitative results with qualitative comparisons.

C Additional Results

Cross-identity reenactment. Figure 5 provides additional qualitative comparisons on the public and NeRSemble benchmarks. Across varied identities, poses, and expressions, PixReenact preserves the reference appearance while following the driver's facial motion. The figure also compares against all per-frame, offline, and streaming baselines evaluated in the main paper.

Sensitivity to the pseudo-target generator. The main paper reports the *wan-labels* model as PixReenact. To test whether the method depends strongly on the particular model used to construct the cross-identity pseudo targets, Table 9 compares it with an otherwise equivalent model trained using PersonaLive-generated pseudo targets. Neither source consistently dominates across benchmarks or metrics: *PL-labels* performs slightly better on several motion and temporal metrics, while *wan-labels* provides stronger identity preservation on the public and NeRSemble benchmarks and lower FVD on the public benchmark. The two models also perform nearly identically on the extreme-condition benchmark. Overall, the results indicate that the proposed training framework is not specific to a single pseudo-target generator.

Self-reenactment. Tables 10 and 11 report self-reenactment results on TalkingHead-1KH, VFHQ-Test, and a small 5-sample NeRSemble benchmark, with qualitative comparisons shown in Figure 6. In this setting, the driving video has the same identity as the reference and therefore provides a

Table 9: **Sensitivity to the pseudo-target generator.** We compare PixReenact trained with Wan-Animate pseudo targets (wan-labels) against the same framework trained with PersonaLive pseudo targets (PL-labels) on cross-reenactment benchmarks. Best result within each benchmark is shown in **bold**.

Benchmark	Pseudo targets	ID↑	AED↓	APD↓	tLP↓	FVD↓
Public (75)	Wan-Animate	**0.709**	0.790	**0.066**	1.234	**377.1**
	PersonaLive	0.704	**0.777**	0.070	**1.126**	402.9
NeRSemble (70)	Wan-Animate	**0.753**	0.714	0.0426	0.803	264.8
	PersonaLive	0.735	**0.689**	**0.0386**	**0.786**	**250.8**
Extreme (68)	Wan-Animate	0.727	**1.135**	**0.169**	**1.853**	662.5
	PersonaLive	**0.729**	1.139	0.175	1.876	**653.7**

Table 10: **Self-reenactment on the public benchmark** (75 clips: TalkingHead-1KH + VFHQ-Test). All methods are re-run at 512^2 on full clips at native FPS with frame-aligned evaluation; PersonaLive uses streaming mode, and X-NeMo is evaluated in its native crop space. Offline methods have no emission latency. *VAE round-trip* gives the reconstruction ceiling and is excluded from comparison. For streaming methods, the better result between PersonaLive and PixReenact is shown in **bold**.

	Method	PSNR↑	SSIM↑	LPIPS↓	L1↓	ID↑	AED↓	APD↓	tLP↓	FVD↓	Lat.↓	FPS↑
TalkingHead-1KH (25 clips) Wang et al. (2021)												
	VAE round-trip	*38.25*	*0.971*	*0.015*	*0.009*	*0.857*	*0.084*	*0.0019*	*0.184*	*5.7*	–	–
Per-frame	LivePortrait	19.90	0.717	0.238	0.069	0.881	0.371	0.0256	0.856	154.2	–	18.90
	X-Portrait	20.32	0.722	0.203	0.063	0.839	0.366	0.0226	0.832	209.1	–	0.65
Offline	MegActor-Σ	17.47	0.664	0.284	0.086	0.825	0.460	0.0471	1.138	119.0	–	0.49
	Wan-Animate	21.22	0.750	0.184	0.058	0.763	0.393	0.0159	1.530	106.4	–	0.41
	X-NeMo	17.40	0.628	0.294	0.088	0.816	0.387	0.0291	1.067	189.2	–	1.37
	Wan-Animate-2	16.96	0.634	0.338	0.107	0.655	0.600	0.0462	3.990	381.7	–	0.29
	FlashPortrait	17.19	0.646	0.287	0.095	0.839	0.476	0.0566	1.621	203.4	–	3.03
Streaming	PersonaLive	19.55	0.696	0.222	0.070	**0.869**	0.363	0.0210	0.938	199.4	**211.1**	**18.95**
	PixReenact	**37.48**	**0.965**	**0.019**	**0.009**	0.855	**0.096**	**0.0028**	**0.237**	**8.2**	238.9	16.74
VFHQ-Test (50 clips) Xie et al. (2022)												
	VAE round-trip	*35.73*	*0.950*	*0.020*	*0.011*	*0.870*	*0.090*	*0.0022*	*0.243*	*1.8*	–	–
Per-frame	LivePortrait	18.97	0.639	0.284	0.075	0.889	0.318	0.0218	1.481	95.1	–	18.90
	X-Portrait	19.43	0.660	0.221	0.065	0.855	0.330	0.0270	0.852	112.7	–	0.65
Offline	MegActor-Σ	17.15	0.581	0.341	0.093	0.793	0.430	0.0449	1.084	111.4	–	0.49
	Wan-Animate	21.45	0.710	0.174	0.051	0.825	0.311	0.0139	0.997	57.1	–	0.41
	X-NeMo	17.09	0.567	0.318	0.092	0.839	0.339	0.0255	1.264	100.5	–	1.37
	Wan-Animate-2	18.49	0.618	0.267	0.074	0.780	0.397	0.0299	2.348	122.9	–	0.29
	FlashPortrait	17.32	0.587	0.313	0.089	0.844	0.395	0.0446	1.655	115.2	–	3.03
Streaming	PersonaLive	19.04	0.631	0.248	0.070	**0.878**	0.332	0.0212	1.317	83.8	**211.1**	**18.95**
	PixReenact	**34.03**	**0.930**	**0.034**	**0.014**	0.860	**0.117**	**0.0045**	**0.377**	**5.5**	238.9	16.74

strong appearance cue. PixReenact achieves reconstruction quality close to the VAE reconstruction baseline, obtained by directly encoding and decoding the ground-truth video with the frozen Wan-VAE. Importantly, its strong cross-identity performance shows that this does not arise simply from copying the driver: the same model transfers motion from a different identity while preserving the reference appearance.

Long-horizon streaming. Figure 7 shows PixReenact outputs sampled at one-minute intervals from two five-minute cross-identity streams. The generated appearance remains stable throughout the sequence while following the changing driver motion, complementing the quantitative identity-stability analysis in Figure 3.

Table 11: **Self-reenactment on the NeRSemble Kirschstein et al. (2023) test set** (5 clips of held-out subjects). *VAE round-trip* marks the shared-VAE reconstruction ceiling and is excluded from comparison. For streaming methods, the better result is shown in **bold**.

	Method	PSNR↑	SSIM↑	LPIPS↓	L1↓	ID↑	AED↓	APD↓	tLP↓	FVD↓
	VAE round-trip	*38.19*	*0.959*	*0.023*	*0.007*	*0.815*	*0.111*	*0.0024*	*0.259*	*8.8*
Per-frame	LivePortrait	25.88	0.810	0.168	0.030	0.848	0.341	0.0105	0.943	168.4
	X-Portrait	24.31	0.739	0.157	0.035	0.822	0.430	0.0168	0.841	139.1
Offline	MegActor-Σ	20.07	0.704	0.279	0.062	0.743	0.642	0.0297	1.190	319.3
	Wan-Animate	24.99	0.797	0.161	0.033	0.752	0.510	0.0151	1.679	209.3
	X-NeMo	23.36	0.765	0.186	0.043	0.784	0.368	0.0136	0.842	133.8
	Wan-Animate-2	23.10	0.767	0.212	0.042	0.667	0.508	0.0222	2.181	176.0
	FlashPortrait	21.30	0.695	0.234	0.054	0.811	0.538	0.0330	1.794	318.3
Streaming	PersonaLive	23.95	0.770	0.200	0.042	0.794	0.697	0.0223	1.219	251.7
	PixReenact	**37.53**	**0.949**	**0.031**	**0.009**	**0.814**	**0.129**	**0.0033**	**0.302**	**13.9**

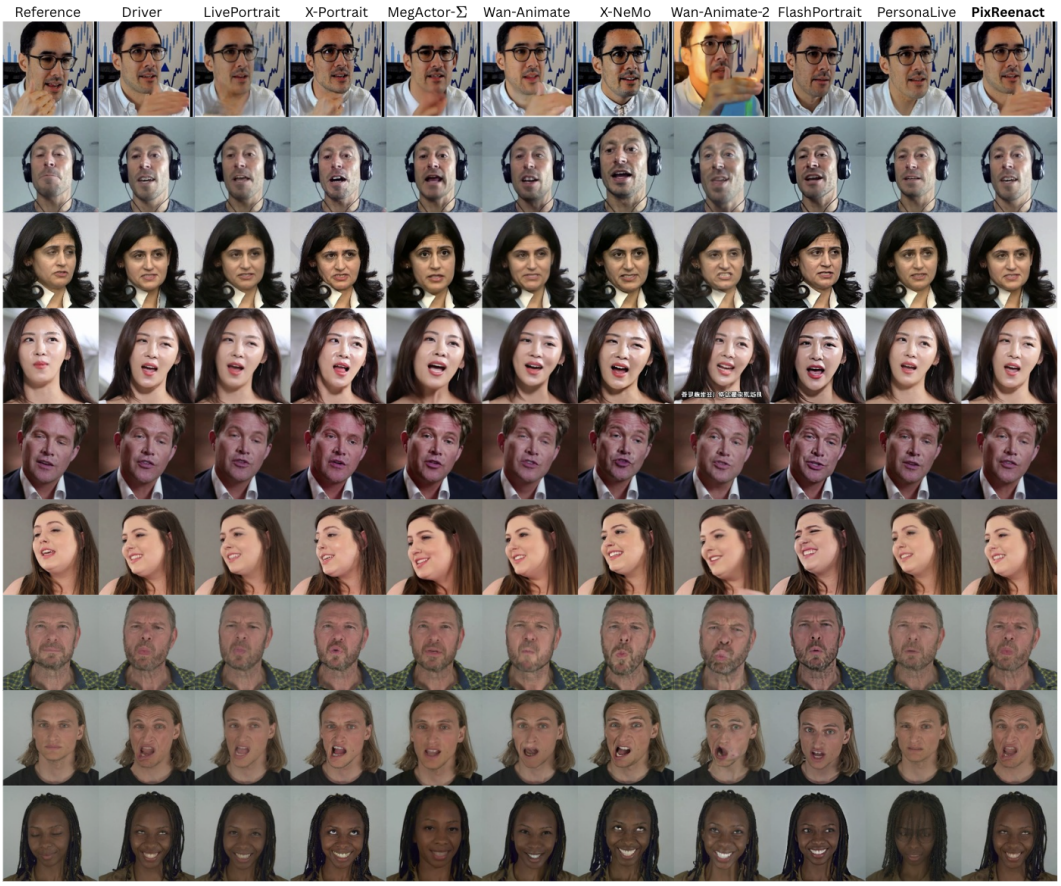

Figure 6: **Qualitative self-reenactment comparisons.** Rows 1-3 show TalkingHead-1KH, rows 4-6 VFHQ-Test, and rows 7-9 NeRSemble. PixReenact closely reconstructs the driving motion and appearance while maintaining high visual fidelity across all three datasets.

Single-window cross-reenactment. To separate cross-reenactment quality from the difficulty of long-video continuation, Table 12 evaluates the public benchmark using only the first 77 frames of each driver, corresponding to a single Wan-Animate Cheng et al. (2025) generation window. The overall trends remain similar to the full-length evaluation: PixReenact maintains strong identity preservation and video quality, while the relative differences in motion metrics remain. This shows



Figure 7: **Long-horizon streaming examples.** PixReenact outputs from two five-minute cross-identity videos, shown at one-minute intervals. The small images show the corresponding driver frames. PixReenact keeps a stable appearance while following the driver over time.

Table 12: **Single-window cross-reenactment.** We use the same 75 reference-driver pairs as the public benchmark (Table 2), but trim each driver to its first 77 frames (≈3 s), corresponding to one Wan-Animate generation window. Thus, offline baselines require no long-video continuation or stitching. Better streaming results are shown in bold. FVD is set-level.

	Method	ID↑	AED↓	APD↓	tLP↓	FVD↓
Per-frame	LivePortrait	0.892	1.075	0.207	1.151	636.6
	X-Portrait	0.716	0.911	0.106	1.546	639.5
Offline	MegActor-Σ	0.570	0.829	0.100	1.911	620.0
	Wan-Animate	0.586	0.935	0.085	1.609	690.0
	X-NeMo	0.619	0.765	0.064	1.373	510.8
	Wan-Animate-2	0.655	0.865	0.090	3.633	690.7
	FlashPortrait	0.775	0.858	0.107	3.697	715.9
Streaming	PersonaLive	0.681	**0.782**	**0.051**	**1.056**	439.3
	PixReenact	**0.728**	0.802	0.074	1.318	**432.2**

that the main cross-reenactment results are not explained solely by the need to stitch or continue long clips.

Extreme-condition breakdown. Table 13 breaks down the extreme-condition benchmark (Sec. 4.2) into extreme yaw, extreme pitch, accessory occlusion, and hand occlusion. PixReenact consistently preserves the reference identity better than PersonaLive across all four groups, while remaining competitive in motion and temporal consistency.

Qualitative ablations. Figure 8 complements the quantitative component ablations in Table 3. Weaker identity constraints and narrower driving attention can increase driver-appearance leakage, while removing motion supervision reduces expression and pose transfer. Removing cross-identity supervision or the cold-start teacher can also introduce local artifacts and weaker motion transfer.

Additional ablations. Table 14 reports three additional sets of controlled experiments. Their protocol differs from the main-paper ablations, where all variants start from the same full-model checkpoint and are continued for 5K steps with one component removed. Here, each group addresses a separate training-design question and therefore uses its own matched setup. Results are directly comparable only within the same group. For the **DMD ablation**, the variants with and without DMD are initialized from the same dense teacher and trained for 20K steps. For the **pseudo-target filtering ablation**, the filtered and unfiltered variants start from the same pre-filter checkpoint and are trained for an additional 5K steps, differing only in whether the cross-identity pseudo targets

Table 13: **Extreme-condition benchmark by group.** Per-group results of the streaming methods on the 68-clip extreme benchmark (Tab. 1); mean ±std over clips. FVD is omitted per group (set-level statistic, unstable at these set sizes).

Method	ID↑	AED↓	APD↓	tLP↓
Extreme yaw (25 clips)				
PersonaLive	0.334 ±0.20	1.416 ±0.46	**0.178** ±0.11	2.482 ±2.20
PixReenact	**0.708** ±0.09	**1.302** ±0.31	0.251 ±0.12	**2.289** ±1.89
Extreme pitch (17 clips)				
PersonaLive	0.566 ±0.18	1.342 ±0.64	**0.115** ±0.07	**1.661** ±1.00
PixReenact	**0.733** ±0.08	**1.057** ±0.15	0.126 ±0.06	1.954 ±0.90
Accessory occlusion (16 clips)				
PersonaLive	0.570 ±0.27	**0.999** ±0.22	0.096 ±0.07	1.416 ±0.76
PixReenact	**0.728** ±0.18	1.011 ±0.25	**0.096** ±0.05	**1.396** ±0.68
Hand occlusion (10 clips)				
PersonaLive	0.503 ±0.17	1.383 ±1.15	**0.125** ±0.11	1.395 ±0.66
PixReenact	**0.761** ±0.04	**1.050** ±0.32	0.152 ±0.11	**1.321** ±0.47
All (68 clips)				
PersonaLive	0.472 ±0.23	1.295 ±0.64	**0.135** ±0.10	1.866 ±1.57
PixReenact	**0.727** ±0.11	**1.135** ±0.30	0.169 ±0.12	**1.853** ±1.35

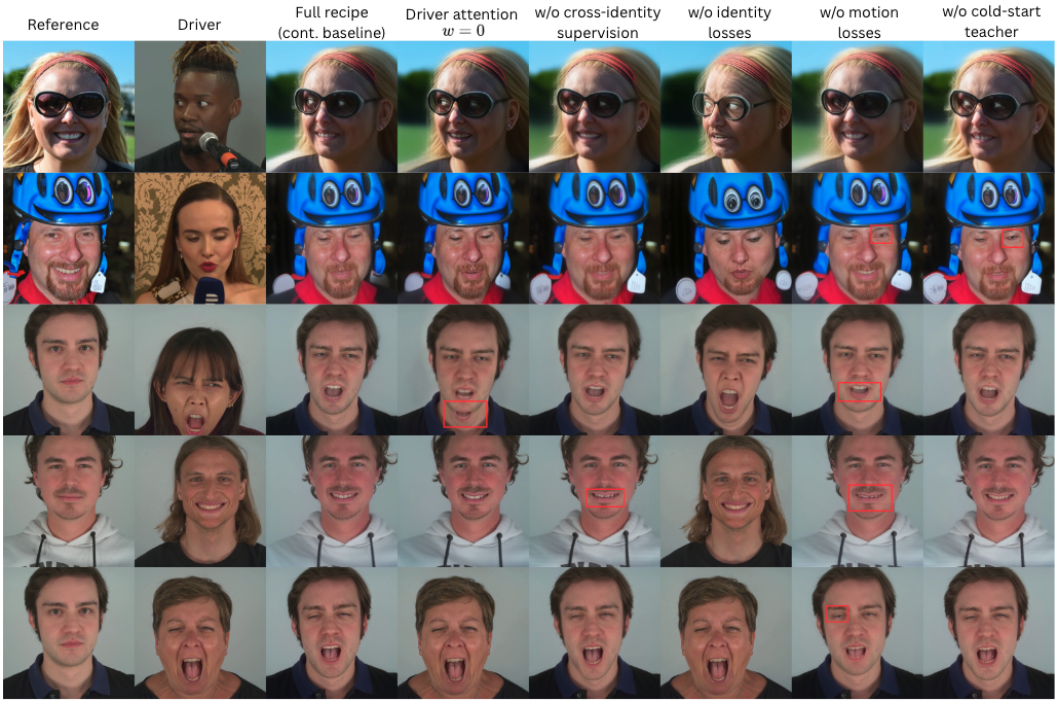

Figure 8: **Qualitative component ablations.** Representative cross-identity examples for the full recipe and its main ablations. Red rectangles highlight representative failure modes.

are filtered. Finally, the **self-reenactment ablations** follow the same continuation protocol as the main-paper ablations: starting from the full model, we continue training for 5K steps after removing either the reconstruction losses or the real self-driving data. The results show that DMD improves

Table 14: **Additional ablations.** DMD variants are trained for 20K steps from the same dense teacher, and filtering variants for 5K steps from the same pre-filter checkpoint. Self-reenactment variants continue the full model for 5K steps. Values are comparable only within each block.

	Variant	ID↑	IDD↓	AED↓	APD↓	tLP↓	FVD↓
DMD	with DMD	0.618	0.029	0.944	0.155	0.967	508
	w/o DMD	0.643	0.026	0.966	0.157	1.178	553
Filtering	filtered	0.841	0.018	0.977	0.161	0.995	630
	unfiltered	0.819	0.019	0.997	0.167	1.001	609

(a) Cross-identity training ablations

	TalkingHead-1KH			VFHQ-Test		
Variant	PSNR↑	LPIPS↓	FVD↓	PSNR↑	LPIPS↓	FVD↓
Full recipe	37.33	0.019	8.3	33.94	0.032	5.6
w/o reconstruction losses	35.40	0.043	45.6	33.32	0.036	6.2
w/o self-driving data	36.63	0.035	29.9	34.04	0.035	5.8

(b) Self-reenactment fidelity

temporal consistency and FVD, pseudo-target filtering provides modest gains in identity and motion accuracy, and reconstruction losses and real self-driving samples improve self-reenactment fidelity.