STREAMING END2END AUDIO-VISUAL IDENTITY SWAPPING FOR TALKING VIDEOS

Anonymous authors
Paper under double-blind review



Figure 1: **UniSwap performs streaming real-time end-to-end joint audio-visual identity replacement in a single model.** Given a source talking video and a reference identity (an image and voice clip), UniSwap jointly transfers appearance and vocal timbre while preserving the source motion, background, and linguistic content. Blockwise autoregressive generation with 3-step sampling runs at 23.5 decoded FPS end-to-end and supports stable long-form generation.

ABSTRACT

Talking-video character replacement requires coordinated transfer of appearance and voice while preserving the source motion, scene, linguistic content, and audio-video timing. Existing methods use separately optimized models for the two modalities, making audio-visual consistency difficult to enforce. We present **UniSwap**, the first framework for streaming joint audio-visual identity replacement in talking videos. Given a source video, a reference image, and a reference voice clip, UniSwap transfers the reference appearance and vocal timbre within a single audio-visual diffusion transformer while preserving the source content and dynamics. To address the scarcity of aligned cross-identity training pairs, we introduce a swap-and-reconstruct pipeline that synthesizes identity-mismatched yet motion-preserving source videos with learned replacement models and reconstructs the original clips under reference guidance, exposing the model exclusively to natural video conditioning without estimated pose, landmark, or segmentation inputs. Starting from a bidirectional backbone, we progressively adapt the model through In-context Pretraining for joint replacement, Conditional Streaming Adaptation for block-causal KV-cached generation, and Efficient Self-forcing DMD for mitigating exposure bias while reducing sampling from 30 to 3 denoising steps per block; Efficient Multi-LoRA Switching lets the three DMD roles share a single frozen backbone. Feature-RoPE Decomposition keeps cached positions within the training range, supporting stable long-form inference. Experiments demonstrate strong audio-visual synchronization, competitive identity preservation, efficient streaming, and stable long-form generation.

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.

1 INTRODUCTION

Recent generative models have substantially advanced video synthesis Wan et al. (2025); HaCohen et al. (2026); Yang et al. (2025) as well as speech and audio generation Du et al. (2024); Défossez et al. (2022). These advances enable applications such as film post-production, content localization, and personalized media, where a central capability is audio-visual identity replacement: changing both the appearance and voice of a person in a talking video while preserving the source motion, background, and linguistic content.

Existing methods address only one side of this task. Video character-replacement methods Xu et al. (2026); Cheng et al. (2025); Jiang et al. (2025b); Hu et al. (2025) transfer appearance from a reference image but not the voice, whereas voice-conversion systems Liu (2024); Jiang et al. (2025a); Du et al. (2024) modify speaker timbre without visual context. Cascading such components leaves the two modalities independently optimized: no joint objective enforces consistency between the converted speech and lip motion, and errors in one modality cannot be corrected using evidence from the other. Moreover, most diffusion-based video replacement methods require the complete clip before generation, preventing blockwise streaming and limiting low-latency interactive applications.

We therefore formulate character replacement in talking videos as joint audio-video generation: a unified model conditions on both target appearance and vocal timbre, with cross-modal interactions coordinating the generated speech and lip motion. Two challenges arise. First, paired videos that differ in visual and vocal identity yet remain aligned in motion, scene, speech, and timing are impractical to collect at scale. Second, high-quality audio-video diffusion backbones operate bidirectionally over complete sequences and require many denoising steps, whereas low-latency streaming requires causal, blockwise generation with bounded per-block computation.

We introduce **UniSwap**, the first framework for *streaming joint audio-visual identity replacement in talking videos* (Fig. 1). Given a source talking video, a reference image, and a reference voice clip, UniSwap transfers the reference appearance and vocal timbre within a single audio-video diffusion transformer and generates the output autoregressively in blocks.

Aligned supervision is constructed by swap-and-reconstruct: each real clip serves as the target, a learned character-replacement model and voice conversion substitute a donor identity to form the source, and the original appearance and voice provide the references. Keeping the source a natural video preserves motion cues that structural proxies discard, avoids estimator errors, and aligns the training source distribution with real videos at inference, so ordinary talking videos become aligned training pairs without different people performing the same motion and speech.

UniSwap builds on LTX-2.3 HaCohen et al. (2026), an audio-video diffusion transformer with native cross-modal attention, and is progressively adapted through three training stages. Stage 1, In-context Pretraining, concatenates source, reference, and noisy target latents with shared temporal coordinates to learn joint visual and vocal identity replacement. Stage 2, Conditional Streaming Adaptation, injects the source block by block under a Decoupled Streaming Conditioning Mask that restricts each token region to its inference-time receptive field, converting the bidirectional model into a block-causal generator with bounded-memory KV-cached inference. Stage 3, Efficient Self-forcing DMD Huang et al. (2026), trains on the student's own rollouts while DMD Yin et al. (2024b) reduces sampling from 30 to 3 denoising steps per block, and Efficient Multi-LoRA Switching lets the teacher, generator, and critic share one frozen backbone. For long-form inference, Feature-RoPE Decomposition separates cached features from rotary coordinates and reapplies bounded positions, preserving cross-modal temporal alignment.

Our main contributions are threefold:

- We introduce UniSwap, the first framework for streaming joint audio-visual identity replacement in talking videos, transferring appearance and vocal timbre within a single diffusion transformer so that lip motion and converted speech are coherent by construction.
- We present a three-stage recipe that converts a bidirectional audio-video diffusion backbone into a causal three-step streaming generator: its swap-and-reconstruct pipeline supplies aligned supervision from ordinary talking videos, and Feature-RoPE Decomposition supports stable long-form generation from a model trained only on short clips.

- Extensive experiments show that UniSwap achieves the best audio-visual synchronization among the evaluated systems with competitive identity preservation and perceptual quality, runs end-to-end at 23.5 FPS and maintains stable identity over minute-scale generation.

2 Related Work

2.1 Video Character Replacement

Video character replacement substitutes a person's visual identity while preserving motion, expression, and background. Frame-based face swapping Chen et al. (2020); Li et al. (2020); Gao et al. (2021) often suffers from temporal inconsistency, so recent methods build on video diffusion models: MoCha Xu et al. (2026) casts replacement as in-context generation with condition-aware RoPE Su et al. (2024), while Wan-Animate Cheng et al. (2025), VACE Jiang et al. (2025b), and HunyuanCustom Hu et al. (2025) inject structural signals or masked-region conditioning into DiT-based generation Peebles & Xie (2023). In-context diffusion more generally places conditioning examples and noisy targets in a shared transformer context Huang et al. (2024); Fei et al. (2024); He et al. (2025). These methods address visual identity only, so dubbing still requires a separate audio system.

2.2 Voice Conversion

Voice conversion changes a speaker's timbre while preserving linguistic content and prosody. Recent zero-shot methods Liu (2024); Jiang et al. (2025a); Du et al. (2024) replace parallel data and explicit feature disentanglement with large-scale pretraining on self-supervised or supervised speech representations. These systems operate without visual context and therefore do not jointly optimize converted speech and lip motion.

2.3 Audio-Visual Generation

Recent generative models have begun to model audio and video jointly. MM-Diffusion Ruan et al. (2023) couples modality-specific denoisers through cross-modal attention, and more recent diffusion transformers strengthen joint modeling: JavisDiT Liu et al. (2026a) introduces hierarchical spatio-temporal priors for audio-visual synchronization, while LTX-2 HaCohen et al. (2026) generates synchronized audio and video from text via interacting modality streams. UniSwap builds on this capability for character replacement, transferring visual and vocal identity while retaining source motion and speech content.

2.4 Distillation for Real-Time Generation

Diffusion models typically require many denoising steps, limiting low-latency applications. Distribution Matching Distillation (DMD) Yin et al. (2024b;a) distills a multi-step teacher into a few-step student by matching output distributions, while consistency models Song et al. (2023) enforce consistency along the probability-flow ODE. For streaming generation, Diffusion Forcing Chen et al. (2024) diffuses independently noised tokens, Self-Forcing Huang et al. (2026) trains on the model's own outputs, CausVid Yin et al. (2025) combines DMD with causal generation, Rolling Forcing Liu et al. (2026b) adds rolling-window denoising with an initial-frame attention sink, and OmniForcing Su et al. (2026) distills a bidirectional audio-visual diffusion model into a streaming generator. UniSwap instead applies self-forcing DMD to source-driven joint appearance-and-voice replacement.

3 Method

Given a source video $V_s \in \mathbb{R}^{F\times 3\times H\times W}$ with audio waveform A_s, a reference image I_r, and a reference voice clip A_r, UniSwap generates a target video V_t and audio A_t. The target should match the reference appearance and vocal timbre while preserving the source motion, lip movements, background, and linguistic content. UniSwap builds on the frozen LTX-2.3 audio-video diffusion transformer HaCohen et al. (2026), which processes video and audio latents with channel dimensions C_v and C_a in separate streams connected by cross-modal attention; the token positions

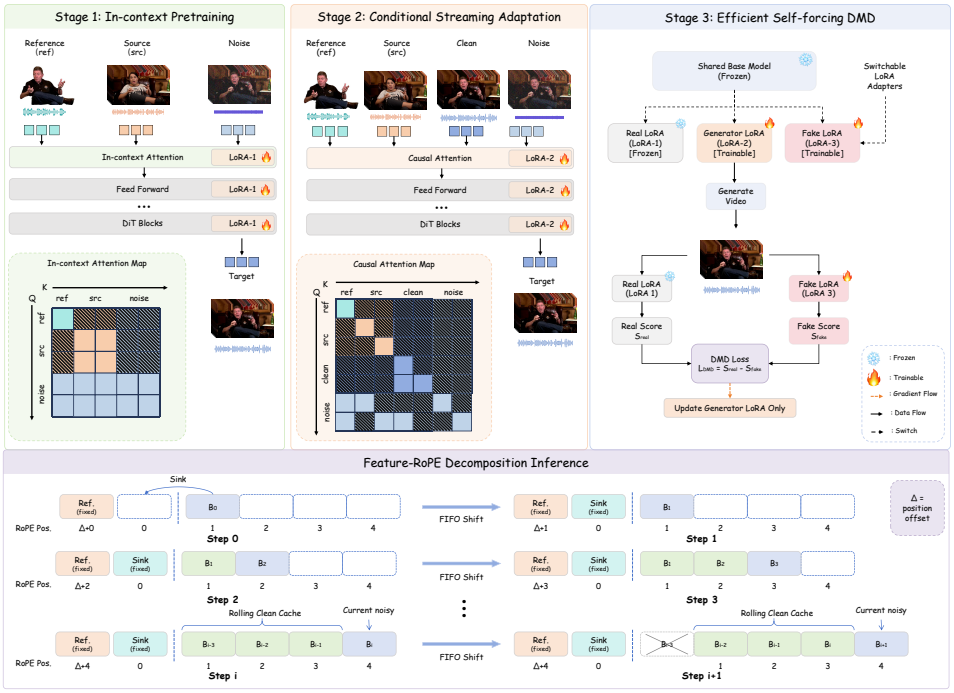


Figure 2: **Overview of UniSwap.** (a) In-Context Pretraining for joint audio-video replacement; (b) Conditional Streaming Adaptation with block-causal masking and KV-cached inference; (c) Efficient Self-Forcing DMD reduces denoising to 3 steps per block. Feature-RoPE Decomposition bounds cached positions while preserving cross-modal physical-time alignment.

of both modalities are assigned on a shared physical-time axis to maintain cross-modal alignment. Figure 2 illustrates the three-stage training pipeline built on this backbone.

3.1 Swap-and-Reconstruct Paired Data Synthesis

Paired data for joint audio-video identity replacement are not naturally available at scale. We therefore formulate training as self-reconstruction (Fig. 3): a real talking video provides the target (V_t, A_t), while an identity-altered version provides the source (V_s, A_s), and the original visual and vocal identities serve as the references. Specifically, a learned character-replacement model Xu et al. (2026) replaces the person in the real clip with a donor identity sampled from another clip of the same body-framing category to yield V_s, and an off-the-shelf voice conversion model Liu (2024) shifts the speaker timbre towards the same donor to yield A_s; a white-background portrait frame of the original person and a 2-second crop of the original audio form the reference image I_r and reference voice A_r. Because the source remains a natural video at all times, this construction preserves motion cues that structural proxies discard, such as hand articulation, gaze, and subtle facial dynamics, avoids estimator errors, and aligns the training source distribution with real videos at inference. All synthesized tuples pass two final automatic rounds of visual-quality and audio-visual synchronization filtering, which catch artifacts and desynchronization introduced by synthesis, and all three training stages use the resulting data. The full curation funnel, filtering counts, and quality-control protocol are described in Appendix F.

3.2 Stage 1: In-context Pretraining

Following the in-context conditioning paradigm for diffusion transformers Fei et al. (2024); Huang et al. (2024), the first stage adapts the model to joint audio-video character replacement by placing the conditioning and target latents in a unified attention context. As shown in Stage 1 of Fig. 2, all inputs are concatenated into a unified token sequence.

Input Formulation. For each training sample, we encode: (1) the source video latent $z_s^v \in \mathbb{R}^{L_s^v \times C_v}$ and source audio latent $z_s^a \in \mathbb{R}^{L_s^a \times C_a}$ representing the driving motion and content; (2) a reference

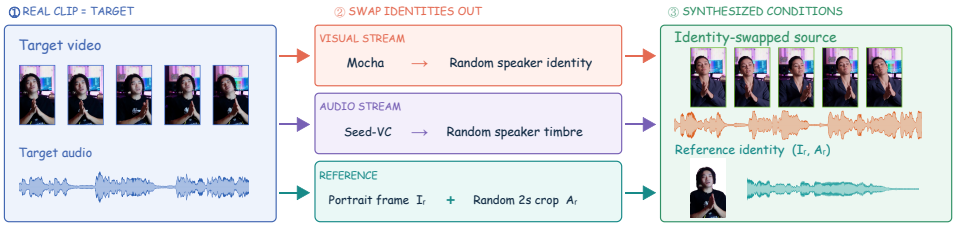


Figure 3: **Swap-and-reconstruct paired data synthesis.** Every real talking video serves as its own reconstruction target: a learned character-replacement model and voice conversion substitute a donor identity to form the source, while the original appearance and voice provide the references. Curation and filtering details are given in Appendix F.

image latent $z_r^v \in \mathbb{R}^{L_r^v \times C_v}$ and reference audio latent $z_r^a \in \mathbb{R}^{L_r^a \times C_a}$ providing the target identity; and (3) the target video latent z_t^v and target audio latent z_t^a to be denoised.

These are concatenated along the sequence dimension for each modality stream:

$$\text{Video:} \quad x^v = [z_r^v; z_s^v; z_t^v], \quad \text{Audio:} \quad x^a = [z_r^a; z_s^a; z_t^a] \tag{1}$$

During training, noise is added only to the target portions z_t^v and z_t^a; the source and reference tokens serve as clean conditioning context with $\sigma = 0$.

Condition Positional Encoding Offset. A naive application of 3D RoPE Su et al. (2024) would assign sequential temporal indices to all tokens, coupling the model to fixed input lengths. Instead, we employ condition positional encoding offsets: the source and target share identical temporal positions in both modalities because they represent the same time span, while the reference image and reference audio use fixed offsets Δ_r^v and Δ_r^a, respectively. This design enables variable-length inputs while keeping the positional semantics of all conditions consistent throughout adaptation and inference.

Training. The model is trained end-to-end with in-context attention across the entire concatenated sequence. For a clean latent z_0, Gaussian noise ϵ, and noise level σ, we define $z_\sigma = (1-\sigma)z_0 + \sigma\epsilon$ and the conditional flow-matching loss as

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{z_0,\sigma,\epsilon}\left[\|v_\theta(z_\sigma, \sigma) - (\epsilon - z_0)\|_2^2\right]. \tag{2}$$

Both video and audio target streams are jointly denoised:

$$\mathcal{L}_{\text{Stage1}} = \mathcal{L}_{\text{FM}}^{\text{video}} + \mathcal{L}_{\text{FM}}^{\text{audio}} \tag{3}$$

Text prompts are randomly replaced by a null-prompt embedding with probability 0.5, so that a single model supports both text-guided and text-free conditioning at inference.

3.3 Stage 2: Conditional Streaming Adaptation

Streaming replacement must also *consume* the source in a streaming fashion: the model cannot wait for the complete source clip, and its memory must stay bounded regardless of duration. Stage 2 therefore converts the bidirectional model into a blockwise autoregressive generator with blockwise source injection. We partition the target into blocks of K=3 video latent frames (an initial 4-frame block is required by the causal video VAE), with audio blocks defined over the same physical-time intervals, and denote the i-th target and source blocks by B_i and S_i. Conditioning is decoupled into three roles with distinct temporal scopes: the reference is global and persistent, each source block is local and temporary, and the clean target history is causal and growing. The model denoises B_i conditioned on the reference, its aligned S_i, and the preceding target blocks $B_{<i}$. Concretely, each source block S_i is encoded independently and injected only while B_i is denoised: the noisy tokens of B_i attend to S_i in both modality self-attention and cross-modal attention, and since the two blocks cover the same physical-time interval, they share identical RoPE coordinates under the condition offsets of Sec. 3.2. Unlike full-sequence conditioning, where every target token attends to the entire source, each target block here sees exactly the source tokens within its own time span, so the source can be consumed as it arrives with a temporary, duration-independent cache, and future source motion cannot leak into the current block.

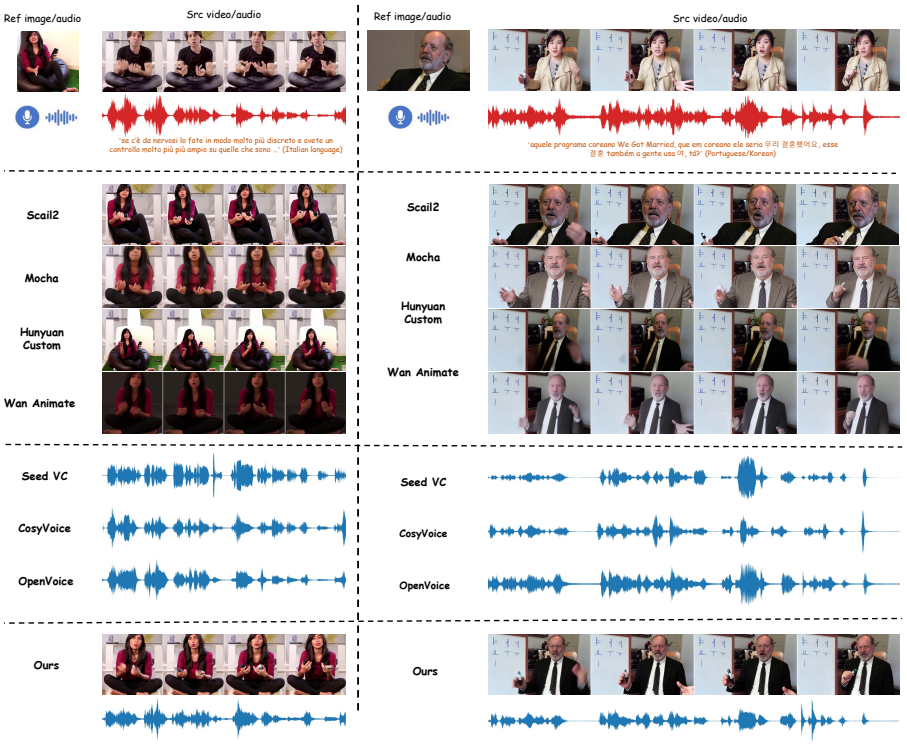


Figure 4: **Qualitative comparison on the short-video benchmark.** Video replacement methods (upper rows) transfer the reference appearance but keep the original voice; voice conversion methods (middle rows) modify only the waveform; UniSwap (bottom) jointly replaces appearance and voice with synchronized lip motion.

Decoupled Streaming Conditioning Mask. During training, the three conditioning roles above are realized by the Decoupled Streaming Conditioning Mask, which restricts each token region to its inference-time receptive field (Stage 2 of Fig. 2): the noisy target block B_i attends to the reference, its aligned S_i, the clean history $B_{<i}$, and itself, while future target and future source tokens are masked out. The same role-wise mask is applied to audio and video self-attention and to cross-modal attention, with modality-specific block sizes chosen to preserve physical-time alignment.

The clean stream is populated with ground-truth latents (teacher forcing Williams & Zipser (1989)), so the model learns to denoise each block B_i under ground-truth history. The loss is computed only on the noisy target tokens:

$$\mathcal{L}_{\text{Stage2}} = \mathbb{E}_{i,\sigma}\left[\mathcal{L}_{\text{FM}}(B_i \mid \text{ref}, S_i, B_0^{\text{clean}}, \ldots, B_{i-1}^{\text{clean}})\right] \tag{4}$$

3.4 Stage 3: Efficient Self-forcing DMD

Stage 2 is trained with teacher forcing on ground-truth histories, whereas at inference the model must condition on its own predictions, a mismatch known as exposure bias; it also requires $\sim$30 denoising steps per block. Stage 3 addresses both issues with Self-forcing Huang et al. (2026): the student is trained on its own autoregressive rollouts, and its Distribution Matching Distillation (DMD) Yin et al. (2024b;a) objective distills the multi-step teacher into a 3-step generator with noise levels $[0.999, 0.757, 0.522]$, supervised at a noise level $\sigma \sim \mathcal{U}[0.02, 0.98]$ by the CFG-enhanced teacher score (γ_{video}=3.0, γ_{audio}=5.0) and a critic trained on noised student samples.

Efficient Multi-LoRA Switching. Self-forcing DMD ordinarily requires separate teacher, generator, and critic models. Efficient Multi-LoRA Switching represents these roles with three adapters on a shared frozen backbone (Stage 3 of Fig. 2): frozen Stage-1 LoRA-1 serves as the teacher, Stage-2-initialized LoRA-2 as the generator, and randomly initialized LoRA-3 as the trainable critic. Activating one adapter at a time avoids maintaining three backbone copies, reducing peak GPU memory from over 80 GB to 65.34 GB and fitting training on a single 80-GB GPU. Only LoRA-2 is retained for inference.

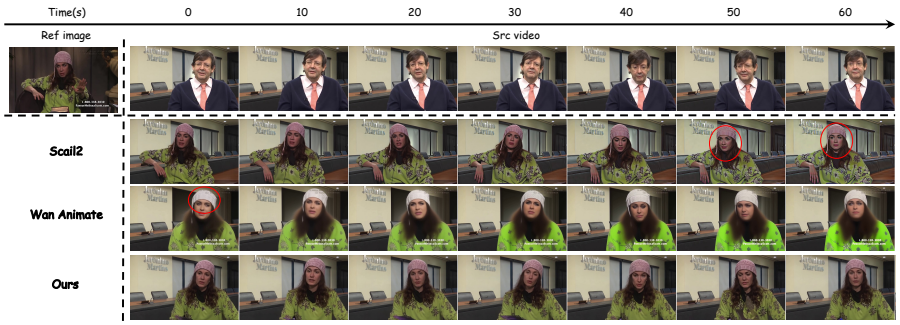


Figure 5: **Qualitative comparison on the long-video benchmark.** Frames are sampled every 10 seconds from 1-minute generations. The baselines exhibit identity drift and visual artifacts as generation progresses (highlighted in red), while UniSwap preserves the reference identity throughout the full duration. Zoom in for details.

Feature-RoPE Decomposition Inference. Absolute RoPE coordinates grow with an autoregressive stream and eventually exceed the range observed during fixed-length training. Feature-RoPE Decomposition instead stores unrotated keys and reapplies RoPE according to each block's current cache slot, without recomputing its features. The cache contains reference tokens, an initial sink block, and a rolling history of recent clean blocks, together with a scratch region for the current source block. At inference, the reference is cached once for the entire generation, each source block is prefilled as it arrives and discarded once its target block completes, and completed target blocks are committed as clean history. The per-block cost is therefore independent of the generated duration, and generation can start as soon as the first source block is available. The complete cache update procedure is given in Appendix C.

Adaptive Sink Block. We adapt the attention-sink mechanism Xiao et al. (2024) by retaining the first generated block B_0 at a fixed local position. Because B_0 establishes the initial appearance and voice, it provides a persistent identity anchor as the rolling history advances.

Reference Re-anchoring. Before denoising each block, the stored reference keys are re-rotated relative to the current generation slot, together with the modality-specific offsets Δ_r^v and Δ_r^a from Sec. 3.2. This keeps the reference at the same relative phase as during training.

Window-Bounded RoPE. Rolling blocks are remapped into a fixed window of W local slots. If $\tau(h_i)$ is the physical timestamp of the earliest retained rolling block, a token at time τ is assigned

$$\widetilde{p}_i(\tau) = \tau - \tau(h_i), \tag{5}$$

while B_0 remains fixed and the current block occupies the final slot. When the window is full, the oldest rolling block is evicted and the retained blocks shift locally. We use $W{=}4$: one sink block, two rolling blocks, and one current block. Video and audio coordinates are derived from the same physical-time axis, preserving cross-modal alignment after every shift.

4 Experiments

4.1 Experimental Setup

Implementation Details. All three stages share the frozen LTX-2.3 base model and train rank-128 LoRA adapters Hu et al. (2021) in bf16 mixed precision on 8 GPUs with FSDP, using AdamW throughout. Stage 1 adapts the attention projections (learning rate $1{\times}10^{-4}$, linear schedule, 50,000 steps). Stage 2 additionally adapts the feed-forward layers and cross-modal attention of both modalities (learning rate $1{\times}10^{-4}$, 50,000 steps). Stage 3 trains the generator and critic adapters (learning rate $1{\times}10^{-5}$, $\beta_1{=}0$, $\beta_2{=}0.999$, 20,000 steps), updating the critic five times per generator update. Inference runs end-to-end on three NVIDIA H100 GPUs, with the encoder, DiT, and decoder each on a separate card, using the cache configuration described in Sec. 3.4. We train on AVSpeech Ephrat et al. (2018) and web-crawled Chinese and English talking videos, and compare against video-replacement and voice-conversion baselines with each video-replacement output paired with Seed-VC Liu (2024) speech (Appendix G).

Evaluation Benchmarks and Metrics. We evaluate on two benchmarks: a short-video benchmark of 100 clips of about 10 seconds from AVSpeech speakers held out from training, and a long-video

Table 1: **Quantitative comparison on the short-video benchmark.** Each video replacement method is paired with the same Seed-VC audio backend, which matches the converter used to synthesize our training sources; audio-visual synchronization is measured on the resulting cascade, and video quality on the generated video. Because all video replacement baselines share the same Seed-VC output, their repeated voice-quality values are omitted. Voice-conversion methods retain the source video. The best result in each column is in **bold**, and the second best is underlined. "–" denotes a metric that is not separately reported for that row.

Method		A–V Sync		Video Quality			Voice Quality				
		Sync-C ↑	Sync-D ↓	ASE ↑	IQA ↑	DINO-S ↑	SIG ↑	BAK ↑	OVRL ↑	SECS ↑	SSIM ↑
Video	VACE Jiang et al. (2025b)	0.832±0.320	12.800±0.933	2.059±0.355	3.269±0.547	0.400±0.110	–	–	–	–	–
	Wan-Animate Cheng et al. (2025)	2.874±1.653	11.338±1.713	2.098±0.340	3.514±0.520	0.580±0.140	–	–	–	–	–
	SCAIL-2 Yan et al. (2026)	3.289±1.592	11.269±1.623	2.409±0.352	4.067±0.399	**0.630±0.150**	–	–	–	–	–
	MoCha Xu et al. (2026)	3.031±1.678	11.198±1.881	**2.534±0.340**	**4.249±0.304**	0.577±0.147	–	–	–	–	–
	HunyuanCustom Hu et al. (2025)	0.894±0.401	12.991±0.864	2.319±0.414	3.816±0.489	0.624±0.142	–	–	–	–	–
Voice	OpenVoice Qin et al. (2023)	–	–	–	–	–	3.458±0.415	3.438±0.663	2.910±0.486	0.755±0.066	**0.363±0.162**
	Seed-VC Liu (2024)	–	–	–	–	–	**3.489±0.335**	**3.750±0.571**	**3.074±0.445**	**0.829±0.047**	0.212±0.115
	CosyVoice Du et al. (2024)	–	–	–	–	–	3.461±0.249	3.738±0.456	3.041±0.366	0.802±0.051	0.137±0.106
Joint	**UniSwap**	**3.823±1.797**	**10.044±1.688**	2.175±0.300	3.792±0.354	0.596±0.141	3.275±0.565	3.385±0.829	2.784±0.604	0.808±0.049	0.302±0.061

Table 2: **Long-video comparison.** Metrics are computed independently on three 20-second segments of 1-minute generated videos. The best result in each segment is in **bold** and the second best is underlined. UniSwap maintains stable quality and identity across the full duration, while the baselines fluctuate or degrade over time.

Method	0–20 s			20–40 s			40–60 s		
	ASE ↑	IQA ↑	DINO-S ↑	ASE ↑	IQA ↑	DINO-S ↑	ASE ↑	IQA ↑	DINO-S ↑
SCAIL-2	**2.628±0.203**	**4.426±0.265**	**0.566±0.175**	**2.765±0.197**	**4.568±0.294**	0.538±0.154	**2.656±0.326**	**4.254±0.401**	0.517±0.135
Wan-Animate	2.241±0.397	3.766±0.635	0.554±0.119	2.271±0.413	3.741±0.701	0.533±0.115	2.238±0.364	3.628±0.714	0.528±0.114
UniSwap	2.429±0.370	4.096±0.424	0.562±0.115	2.418±0.306	4.083±0.386	**0.559±0.110**	2.318±0.304	3.902±0.557	**0.567±0.113**

benchmark of 30 web-crawled talking videos of 1 minute. Both span head, half-body, and full-body shots with randomly collected photorealistic reference portraits, and metrics on the long-video benchmark are computed independently per 20-second segment to expose temporal drift. All methods are evaluated on the same source clips, reference identities, resolutions, and frame rates; we report the mean ± standard deviation over the evaluation clips. Metrics cover four aspects: (1) audio-visual synchronization, Sync-C and Sync-D from SyncNet Chung & Zisserman (2016); (2) video quality, the aesthetic (ASE) and image-quality (IQA) scores of Q-Align Wu et al. (2023) and the DINO-S identity similarity Caron et al. (2021); (3) voice quality, following Seed-VC Liu (2024), SIG, BAK, and OVRL from DNSMOS Reddy et al. (2021), speaker encoder cosine similarity (SECS) to the reference voice Wan et al. (2018), and mel-spectrogram SSIM Wang et al. (2004) for content preservation; and (4) speech intelligibility, Whisper large-v3 Radford et al. (2023) word (English) and character (Chinese) error rates against the source transcript (Appendix E).

4.2 Qualitative Results

Figure 4 shows representative examples from the short-video benchmark; additional examples are provided in Appendix I. Video replacement baselines transfer the reference appearance but require a separate voice-conversion stage; voice conversion baselines modify only the waveform and cannot change the appearance. UniSwap jointly replaces both modalities, with the generated character following the source pose and background, matching the reference appearance and vocal timbre, and maintaining lip motion synchronized with the generated speech.

Fig. 5 further illustrates long-duration generation. SCAIL-2 and Wan-Animate exhibit increasing identity drift and visual artifacts in later segments, whereas UniSwap remains stable over the full minute, consistent with the per-segment metrics in Table 2.

4.3 Quantitative Results

Table 1 presents the short-video comparison. UniSwap obtains the highest Sync-C (3.823) and lowest Sync-D (10.044) among the evaluated replacement pipelines. Unlike cascades that generate lip motion without access to the converted speech, UniSwap generates both modalities jointly. Its identity similarity (DINO-S 0.596) trails SCAIL-2 (0.630) and HunyuanCustom (0.624), and its

aesthetic and image-quality scores remain below MoCha and SCAIL-2. For speech, UniSwap's SECS (0.808) and SSIM (0.302) rank second only to Seed-VC and OpenVoice respectively, while its DNSMOS scores remain below the dedicated voice converters. These results reflect the trade-off between unified joint replacement and separately optimized single-modality systems.

Table 2 compares one-minute generations by segment. UniSwap's identity similarity remains essentially flat across segments (DINO-S 0.562, 0.559, and 0.567): slightly below SCAIL-2 in the first segment (0.566) but clearly ahead by the last (0.567 versus 0.517), while SCAIL-2 retains higher aesthetic and image-quality scores throughout. Wan-Animate's IQA likewise decreases from 3.766 to 3.628 across segments. The results indicate that UniSwap preserves identity more consistently over the evaluated duration, but they do not establish superiority on every visual-quality dimension.

Streaming Efficiency. UniSwap pipelines input encoding, DiT generation, and decoding across three H100 GPUs, sustaining 23.5 FPS end-to-end at 512×512 (breakdown in Appendix D).

4.4 Ablation Studies

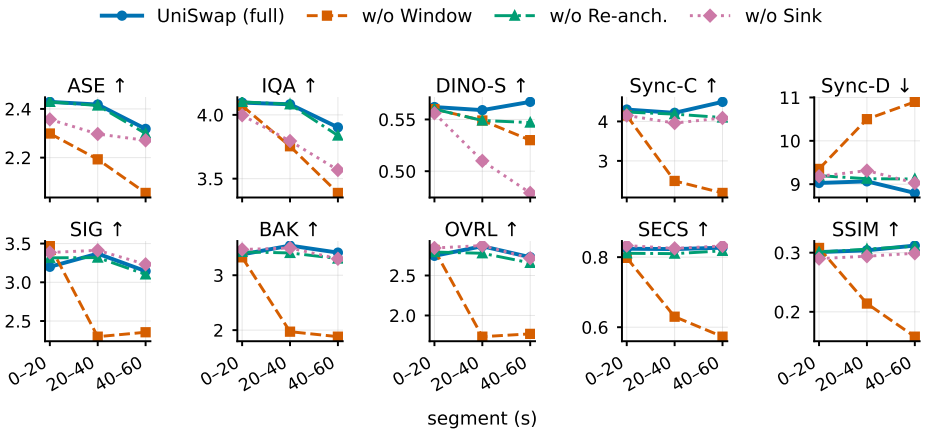


Figure 6: **Per-segment trends in the Feature-RoPE Decomposition ablation** (long-video benchmark). Without Window-Bounded RoPE both modalities collapse after the first segment; without the Adaptive Sink Block visual identity degrades while synchronization and voice similarity hold.

Figure 6 ablates the three positional components of Feature-RoPE Decomposition on the long-video benchmark. The full model remains stable across the minute in both modalities (per-segment DINO-S varies by less than 0.01; Sync-C stays above 4.2). Removing Window-Bounded RoPE progressively degrades both modalities: by the final segment, IQA drops from 4.071 to 3.388, SECS from 0.798 to 0.573, and Sync-C from 4.149 to 2.188; ASR error rates rise from 21.37% to 43.21% (English WER) and from 37.83% to 43.28% (Chinese CER). Removing the Adaptive Sink Block loses visual identity (DINO-S from 0.556 to 0.479) but keeps synchronization and voice identity intact; removing Reference Re-anchoring causes milder drift (English WER 29.73%). Per-segment statistics and the training-stage ablation are in Appendices A–B.

5 Conclusion

We presented UniSwap, the first framework for streaming joint audio-visual identity replacement in talking videos. UniSwap formulates appearance and voice transfer as a unified conditional generation task supervised by swap-and-reconstruct data synthesis, and its three-stage recipe with Feature-RoPE Decomposition converts a bidirectional diffusion backbone into a causal three-step streaming generator for bounded-cache long-form inference. Experiments show stronger synchronization than cascaded systems, competitive identity preservation and perceptual quality, and stable visual identity over one-minute sequences.

REFERENCES

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2024.

Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 2003–2011, 2020.

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.

Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pp. 251–263. Springer, 2016.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, et al. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*, 2024.

Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*, 2018.

Zhengcong Fei, Di Qiu, Debang Li, Changqian Yu, and Mingyuan Fan. Video diffusion transformers are in-context learners. *arXiv preprint arXiv:2412.10783*, 2024.

Gege Gao, Huaibo Huang, Chaoyou Fu, Zhaoyang Li, and Ran He. Information bottleneck disentanglement for identity swapping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3404–3413, 2021.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

Xuanhua He, Quande Liu, Zixuan Ye, Weicai Ye, Qiulin Wang, Xintao Wang, Qifeng Chen, Pengfei Wan, Di Zhang, and Kun Gai. Fulldit2: Efficient in-context conditioning for video diffusion transformers. *arXiv preprint arXiv:2506.04213*, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Teng Hu, Zhentao Yu, Zhengguang Zhou, Sen Liang, Yuan Zhou, Qin Lin, and Qinglin Lu. Hunyuancustom: A multimodal-driven architecture for customized video generation. *arXiv preprint arXiv:2505.04512*, 2025.

Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context lora for diffusion transformers. *arXiv preprint arXiv:2410.23775*, 2024.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *Advances in Neural Information Processing Systems*, 38:167283–167308, 2026.

Yuepeng Jiang, Ziqian Ning, Shuai Wang, Chengjia Wang, Mengxiao Bi, Pengcheng Zhu, Zhonghua Fu, and Lei Xie. Ref-vc: Robust, expressive and fast zero-shot voice conversion with diffusion transformers. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–6. IEEE, 2025a.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17191–17202. IEEE, 2025b.

Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Advancing high fidelity identity swapping for forgery detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5073–5082. IEEE, 2020.

Kai Liu, Wei Li, Lai Chen, Shengqiong Wu, Yanhao Zheng, Jiayi Ji, Fan Zhou, Jiebo Luo, Ziwei Liu, Hao Scofield Fei, et al. Javisdit: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. In *International Conference on Learning Representations*, volume 2026, pp. 139160–139194, 2026a.

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, 2026b.

Songting Liu. Zero-shot voice conversion with diffusion transformers. *arXiv preprint arXiv:2411.09943*, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Zengyi Qin, Wenliang Zhao, Xumin Yu, and Xin Sun. Openvoice: Versatile instant voice cloning. *arXiv preprint arXiv:2312.01479*, 2023.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

Chandan KA Reddy, Vishak Gopal, and Ross Cutler. Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6493–6497. IEEE, 2021.

Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10219–10228. IEEE, 2023.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Yaofeng Su, Yuming Li, Zeyue Xue, Jie Huang, Siming Fu, Haoran Li, Haoyang Huang, and Nan Duan. Omniforcing: Unleashing real-time joint audio-visual generation. In *European Conference on Computer Vision*, pp. 566–584. Springer, 2026.

Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. Generalized end-to-end loss for speaker verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4879–4883. IEEE, 2018.

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

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Ronald J Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280, 1989.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, volume 2024, pp. 21875–21895, 2024.

Zhengbo Xu, Jie Ma, Ziheng Wang, Zhan Peng, Jun Liang, and Jing Li. Mocha: End-to-end video character replacement without structural guidance. *arXiv preprint arXiv:2601.08587*, 2026.

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.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025.

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, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6613–6623. IEEE, 2024b.

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

Appendix

AI USE STATEMENT

We used generative AI tools solely to polish the writing of this paper, such as refining wording and improving readability. All research ideas, methods, experiments, and analyses were carried out by the authors. We have reviewed all AI-assisted work and take full responsibility for the final content of this work.

A QUANTITATIVE ABLATION ON FEATURE-ROPE DECOMPOSITION

Table 3 reports the full per-segment statistics of the Feature-RoPE Decomposition ablation on the long-video benchmark, covering audio-visual synchronization, video quality, and voice quality for each 20-second segment.

Table 3: **Ablation on Feature-RoPE Decomposition** (long-video benchmark, per-segment metrics). The best result in each segment is in **bold** and the second best is underlined.

	Full	w/o Window	w/o Re-anch.	w/o Sink
		0–20 s		
ASE ↑	**2.429±0.370**	2.299±0.052	**2.429±0.373**	2.357±0.386
IQA ↑	4.096±0.424	4.071±0.185	**4.102±0.430**	3.997±0.473
DINO-S ↑	**0.562±0.115**	0.560±0.089	0.560±0.115	0.556±0.116
Sync-C ↑	**4.301±0.950**	4.149±1.050	4.237±0.891	4.151±0.916
Sync-D ↓	**9.028±0.453**	9.352±0.363	9.197±0.286	9.182±0.408
SIG ↑	3.200±0.600	**3.474±0.107**	3.316±0.372	3.385±0.298
BAK ↑	3.379±0.854	3.326±0.588	3.426±0.636	**3.469±0.608**
OVRL ↑	2.746±0.666	2.820±0.354	2.793±0.491	**2.844±0.456**
SECS ↑	0.824±0.043	0.798±0.032	0.811±0.058	**0.833±0.037**
SSIM ↑	0.301±0.037	**0.308±0.039**	0.301±0.038	0.290±0.041
		20–40 s		
ASE ↑	**2.418±0.306**	2.193±0.021	2.414±0.319	2.297±0.468
IQA ↑	**4.083±0.386**	3.754±0.527	**4.083±0.443**	3.795±0.713
DINO-S ↑	**0.559±0.110**	0.549±0.091	0.549±0.114	0.510±0.135
Sync-C ↑	**4.218±0.933**	2.492±1.196	4.185±0.919	3.959±0.942
Sync-D ↓	**9.065±0.576**	10.500±0.781	9.126±0.518	9.313±0.735
SIG ↑	3.370±0.276	2.298±0.863	3.319±0.315	**3.413±0.276**
BAK ↑	**3.540±0.573**	1.972±0.729	3.406±0.579	3.496±0.509
OVRL ↑	2.866±0.436	1.735±0.583	2.778±0.437	**2.874±0.392**
SECS ↑	0.823±0.049	0.630±0.013	0.810±0.058	**0.825±0.036**
SSIM ↑	0.304±0.048	0.214±0.009	**0.306±0.049**	0.294±0.044
		40–60 s		
ASE ↑	**2.318±0.304**	2.055±0.094	2.299±0.335	2.271±0.427
IQA ↑	**3.902±0.557**	3.388±0.741	3.838±0.702	3.568±0.764
DINO-S ↑	**0.567±0.113**	0.530±0.097	0.547±0.112	0.479±0.138
Sync-C ↑	**4.496±1.179**	2.188±1.624	4.099±1.382	4.087±1.279
Sync-D ↓	**8.798±0.640**	10.899±2.165	9.120±0.714	9.028±0.665
SIG ↑	3.144±0.672	2.355±1.105	3.102±0.744	**3.232±0.508**
BAK ↑	**3.412±0.921**	1.881±0.668	3.304±0.901	3.300±0.645
OVRL ↑	**2.729±0.659**	1.770±0.623	2.658±0.698	2.716±0.476
SECS ↑	0.827±0.043	0.573±0.073	0.818±0.055	**0.832±0.035**
SSIM ↑	**0.312±0.051**	0.158±0.047	0.311±0.050	0.299±0.049

B ABLATION ON TRAINING STAGES

Table 4 compares the three training stages on the short-video benchmark. Stage 1 obtains the strongest synchronization, video quality, identity similarity, and speaker similarity. Stage 2 enables

Table 4: **Ablation on training stages and the condition positional encoding offset** (short-video benchmark). The best result in each column is in **bold** and the second best is underlined.

Setting	A–V Sync		Video Quality			Voice Quality				
	Sync-C ↑	Sync-D ↓	ASE ↑	IQA ↑	DINO-S ↑	SIG ↑	BAK ↑	OVRL ↑	SECS ↑	SSIM ↑
Stage 1 (In-context)	**4.155**±1.918	**9.853**±1.831	**2.305**±0.316	**3.963**±0.322	**0.679**±0.133	3.387±0.495	**3.575**±0.717	2.937±0.531	**0.822**±0.050	0.301±0.068
Stage 2 (Teacher forcing)	3.947±1.861	10.014±1.761	2.218±0.320	3.834±0.349	0.593±0.139	**3.522**±0.143	3.570±0.487	**3.002**±0.298	0.701±0.069	**0.308**±0.051
Stage 3 (Self-forcing DMD)	3.823±1.797	10.044±1.688	2.175±0.300	3.792±0.354	0.596±0.141	3.275±0.565	3.385±0.829	2.784±0.604	0.808±0.049	0.302±0.061
Stage 2 w/o condition PE offset	1.738±1.737	11.843±1.606	2.152±0.352	3.472±0.600	0.463±0.166	3.238±0.728	3.280±0.924	2.767±0.702	0.624±0.086	0.103±0.142

block-causal generation and attains the best DNSMOS scores (SIG 3.522 and OVRL 3.002), while synchronization, identity, and speaker similarity degrade. Stage 3 reduces denoising from 30 to 3 steps per block; compared with Stage 2, it recovers speaker similarity (SECS 0.701 to 0.808) and identity preservation (DINO-S 0.593 to 0.596) at a moderate cost in synchronization, aesthetic, and DNSMOS scores.

The condition positional encoding offset is similarly critical: removing it from Stage 2 reduces Sync-C from 3.947 to 1.738 and DINO-S from 0.593 to 0.463, supporting its role in audio-video synchronization and reference-identity preservation.

C KV-Cached Streaming Inference

Algorithm 1 details the blockwise inference procedure used by Stages 2 and 3. The reference cache persists throughout generation, source keys and values are temporary, and completed target blocks are committed to the clean-history cache.

Algorithm 1 KV-Cached Streaming Inference

Require: Reference tokens R; source blocks $\{S_i\}_{i=0}^{N-1}$; reference region $\mathcal{R}$; history $\mathcal{H}$ (sink + rolling)

1: **Prefill reference:** forward R once and write its key/value tensors into $\mathcal{R}$ {never evicted}
2: **for** $i = 0$ **to** $N - 1$ **do**
3: **Prefill source:** forward S_i at the current slot and temporarily cache its key/value tensors
4: **Denoise:** iteratively denoise B_i in read-only mode, attending to $\mathcal{R}$, $\mathcal{H}$, and S_i
5: **Commit:** forward the denoised B_i as clean context, and append its unrotated keys and values to the sink region if $i = 0$, or to the rolling region otherwise
6: **Shift:** once the rolling region is full, evict its oldest block, shift the retained blocks, and reapply RoPE to their cached keys at bounded local coordinates
7: **Rollback:** discard the temporary key/value tensors of S_i
8: **end for**
9: **return** Generated blocks $\{B_i\}_{i=0}^{N-1}$

D Streaming Efficiency

Figure 7 shows the stage-wise runtime at 512×512. Per 24-frame chunk, Input VAE encoding takes 570 ms; on the recurrent DiT stage, source KV prefill, three-step denoising, and clean KV-cache construction take 84, 723, and 179 ms, respectively; and video and audio decoding with muxing take 749 and 273 ms. Since all three stages run concurrently on consecutive chunks, the steady-state throughput is governed by the slowest stage. Decoding and muxing occupy the full 1023 ms chunk launch period, so UniSwap emits one 24-frame chunk every 1023 ms, yielding 23.5 FPS end-to-end. Generation therefore keeps pace with any source stream up to 23.5 fps; a 16-fps source, for instance, leaves about $1.5\times$ throughput headroom, so the pipeline runs in real time.

E Speech Intelligibility

To verify that the generated speech preserves the linguistic content of the source, we transcribe the generated audio with a local Whisper large-v3 model Radford et al. (2023) and measure corpus-micro error rates against the source-side transcript: the word error rate (WER) for English and the

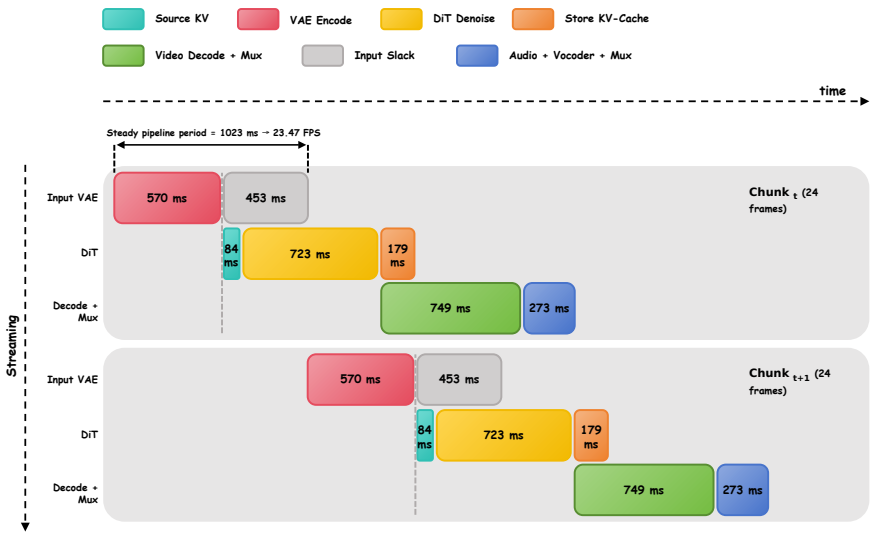


Figure 7: **Stage-wise runtime of UniSwap's real-time deployment pipeline on three H100 GPUs at** 512×512**.** The incoming stream is processed in 24-frame chunks across three pipelined stages: Input VAE encoding on one GPU, recurrent DiT generation on one GPU, and video/audio decoding and muxing on one GPU. Colored bars denote active computation, and gray bars denote input-stage slack. Decoding and muxing occupy a full 1023 ms period per chunk and set the steady-state chunk launch period, yielding 23.5 FPS end-to-end throughput.

character error rate (CER) for Chinese. The evaluation uses 100 additional talking-video clips of AVSpeech speakers held out from training, 50 Chinese and 50 English.

Table 5: **Speech intelligibility** on the supplementary Chinese–English benchmark (ASR error rates against the source transcript; lower is better). Generated audio is transcribed with Whisper large-v3 Radford et al. (2023), and errors are aggregated corpus-micro over the 50 clips of each language. The best result in each column is in **bold** and the second best is underlined.

Method	English WER ↓	Chinese CER ↓
UniSwap (Full)	**6.89**	36.55
w/o Window	30.79	40.98
w/o Re-anch.	<u>8.18</u>	33.63
w/o Sink	**6.89**	34.74
OpenVoice Qin et al. (2023)	8.35	52.15
Seed-VC Liu (2024)	8.91	<u>31.48</u>
CosyVoice Du et al. (2024)	12.64	**30.72**

Table 5 compares UniSwap, its ablated variants, and the voice-conversion baselines on this benchmark. UniSwap attains the lowest English WER (6.89%), below all three dedicated voice converters (8.35–12.64%), and its character-level error rate aggregated over both languages (11.91%) is close to the best converter (Seed-VC, 11.44%); its Chinese CER (36.55%) remains above the strongest converters (CosyVoice 30.72%, Seed-VC 31.48%) but well below OpenVoice (52.15%). The ablations mirror the per-segment trends in Fig. 6 of the main paper: removing Window-Bounded RoPE degrades content most sharply, with English WER rising from 6.89% to 30.79% and Chinese CER from 36.55% to 40.98%, whereas w/o Sink and w/o Re-anch. remain within a few points of the full model on both languages. These results confirm that the generated speech remains readily intelligible and that bounded positional encoding preserves content stability over long horizons.

F Data Curation and Quality Control

Source-pool curation. Figure 8 summarizes the curation funnel. Raw talking-video clips pass a sequence of automatic filters before swap synthesis: (i) near-duplicate removal with MD5 and perceptual hashes; (ii) face filtering with RetinaFace, discarding clips without a detected face or

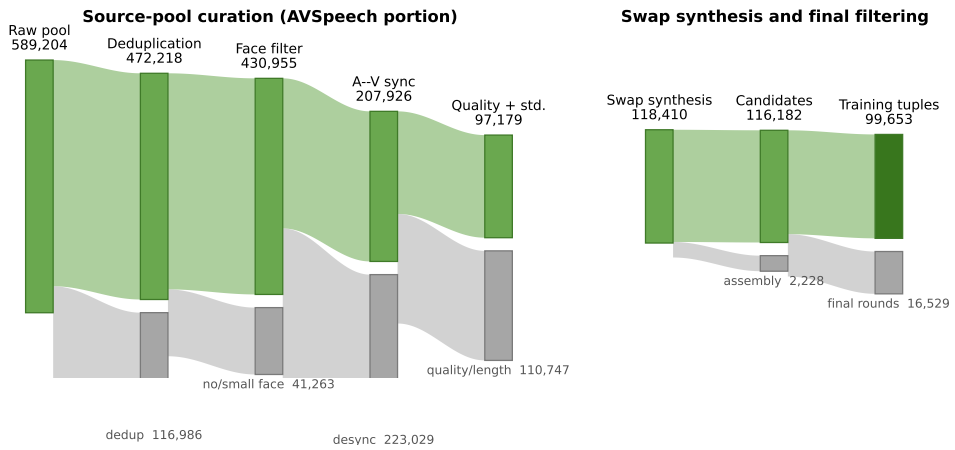


Figure 8: **Data curation and filtering funnel.** Ribbon widths follow a square-root scale for readability. Left: source-pool curation on the AVSpeech portion of the pool, comprising deduplication, face filtering, SyncNet synchronization, and Q-Align quality scoring with standardization. Right: swap synthesis and final assembly, where the candidate tuples pass two final automatic rounds, by visual quality and by audio-visual synchronization, retaining 99,653 training tuples.

whose largest face is too small (short side below 64 px or area below 1% of the frame); (iii) audio-visual synchronization with SyncNet Chung & Zisserman (2016), discarding clips with $|\text{offset}|>1$ frame and correcting ± 1-frame offsets by re-muxing; (iv) visual-quality scoring with Q-Align Wu et al. (2023), keeping clips scored at least 4.0; and (v) standardization to 25 fps, 16 kHz mono audio, 5–20 s duration, and fixed aspect-ratio buckets. Applied to the AVSpeech portion of the source pool (589,204 raw clips), these stages retain 472,218 clips after deduplication, 430,955 after face filtering, 207,926 after synchronization (71,235 already synchronized and 136,691 corrected), and 97,179 after quality scoring and standardization. The web-crawled Chinese and English portion of the pool (25,807 clips) is pre-filtered upstream with the same criteria.

Donor pool and matching. A donor reference is admitted only if its detected face is at least 120 px wide and normally exposed. For each source clip, the donor is sampled from the same body-framing category (head, waist, thigh, calf, or full body) as the source person.

Swap synthesis. Visual identity is swapped with a learned character-replacement model Xu et al. (2026) using an 8-step distilled LoRA sampler (guidance scale 1.0, shift 5.0) and a rectangular bounding-box mask; the donor reference is cropped, centered, and composited onto a gray background. Vocal timbre is converted with Seed-VC Liu (2024) using the donor's speech as the target voice.

Quality assurance. Production is audited by sampling 25 clips per 5,000 produced, combining identity-similarity checks against the source and donor with manual inspection. The audits identified low-quality donor references (back-lit, occluded, or off-domain faces) as the main cause of donor-identity drift; because the training direction reconstructs the *original* identity, such tuples still provide valid supervision and are retained. In a final audit of 100 random clips, 98% have identity similarity to the original person of at most 0.2, and the converted speech preserves linguistic content with a mean character error rate of 0.105 measured by automatic speech recognition. Finally, to catch artifacts and desynchronization introduced by synthesis, the 116,182 candidate tuples pass two further automatic rounds of visual-quality and audio-visual synchronization filtering, retaining 99,653 training tuples.

G Training Data and Baselines

Training Data. We train UniSwap on AVSpeech Ephrat et al. (2018) together with web-crawled supplementary talking videos in Chinese and English, providing clean speech from diverse speakers. Paired training tuples are synthesized from these unpaired clips using the swap-and-reconstruct pipeline described in the main paper, retaining 99,653 filtered clips. Both training and inference

Table 6: **User-study results.** Ratings are mean scores on a five-point Likert scale.

Method	Appearance ID ↑	Voice ID ↑	Lip Sync ↑	Naturalness ↑
Wan-Animate + Seed-VC	3.85	4.04	3.42	3.77
SCAIL-2 + Seed-VC	4.03	4.05	3.67	3.85
MoCha + Seed-VC	3.91	**4.08**	3.54	3.89
HunyuanCustom + Seed-VC	3.64	3.98	3.28	3.61
UniSwap	**4.16**	3.87	**4.11**	**3.96**

support two aspect-ratio buckets: 704×416 and 416×704 pixels. Training clips contain up to 241 frames (approximately 9.6 seconds at 25 fps), with shorter clips kept at their natural length.

Baselines. We compare against three categories of methods. (1) *Video replacement*: MoCha Xu et al. (2026), Wan-Animate Cheng et al. (2025), VACE Jiang et al. (2025b), HunyuanCustom Hu et al. (2025), and SCAIL-2 Yan et al. (2026), which replace the visual identity but not the voice. We pair every visual output with speech converted by Seed-VC—the strongest converter in our evaluation (Table 1 of the main paper) and the backend used in our training pipeline—and evaluate the resulting cascade as the direct competitor to joint replacement. Sharing a common audio backend across all video-only methods isolates differences in their visual components. (2) *Voice conversion*: Seed-VC Liu (2024), CosyVoice Du et al. (2024), and OpenVoice Qin et al. (2023), which convert the voice while keeping the original video. (3) *Ours*: UniSwap, the final streaming generator; the individual training stages are compared separately in Sec. B.

H User Study

We conducted a blinded user study with 30 participants. The study compared UniSwap with four video-replacement baselines, each paired with Seed-VC following the cascade protocol in Sec. G. Each participant evaluated anonymized outputs from all five methods on four source clips and rated appearance identity, voice identity, lip synchronization, and naturalness on a five-point Likert scale. Method order was randomized independently for each participant.

UniSwap receives the highest ratings for appearance identity, lip synchronization, and naturalness.

I Additional Qualitative Results

We provide additional qualitative results for both short and long source videos. In all figures, each example contains a reference image and reference voice clip, a source video and its audio, and the joint audio-video output produced by UniSwap. The red and blue waveforms denote the source and generated audio, respectively. These examples complement the quantitative evaluation in the main paper by covering diverse identities, poses, gestures, clothing, backgrounds, and recording conditions.

I.1 Short-Video Results

Figures 9 and 10 present 16 additional short-video examples that complement the qualitative comparison in the main paper. Across these cases, UniSwap transfers the appearance specified by the reference image while retaining the source composition, body motion, and facial activity. The examples include cross-gender replacement, varied camera framing, substantial hand and upper-body motion, and both simple and cluttered backgrounds. The corresponding generated-audio waveforms are shown alongside the visual outputs to illustrate that the two modalities are produced jointly for the full source sequence.

I.2 Long-Video Results

Figure 11 shows three additional one-minute generations sampled at 10-second intervals. The output identity remains visually consistent from the beginning to the end of each sequence despite continuous autoregressive generation. At the same time, the outputs preserve the source background,



Figure 9: **Additional qualitative results on short videos (Part I).** For each example, the left column provides the reference image and reference voice, while the right column shows sampled frames and the waveform of the source video/audio followed by the UniSwap output. Red waveforms indicate source audio and blue waveforms indicate generated audio. UniSwap changes the visual and vocal identity according to the references while preserving the source scene composition and motion.

camera framing, and time-varying facial and upper-body motion. These results provide further qualitative evidence that the bounded cache coordinates and persistent identity context used by UniSwap mitigate long-horizon identity drift.

J LIMITATIONS

UniSwap currently targets single-speaker talking videos; multi-speaker scenes, occlusions, and complex interactions remain challenging. Facial expressions and head motion follow the source video

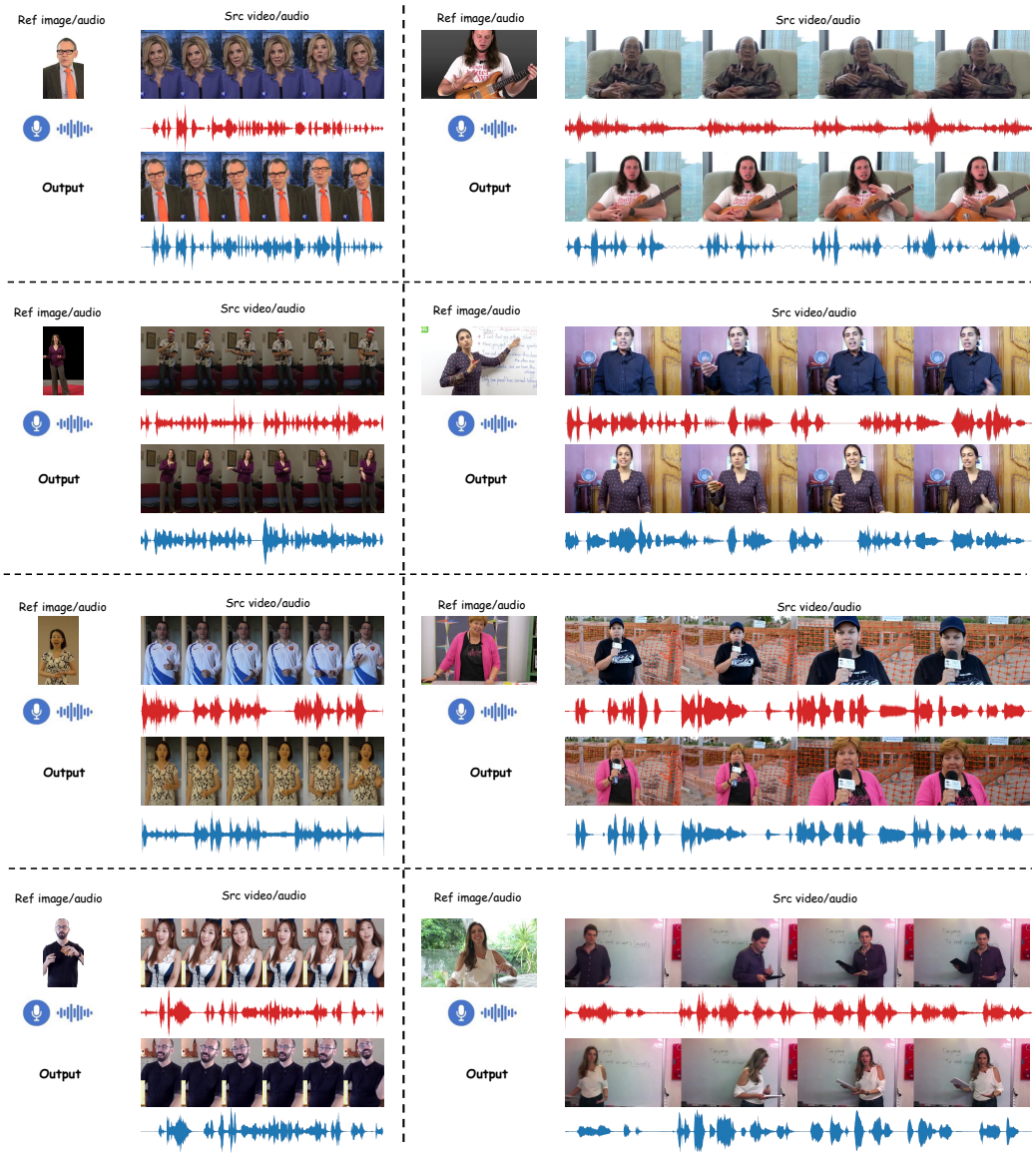


Figure 10: **Additional qualitative results on short videos (Part II).** The examples follow the layout of Fig. 9 and cover additional identities, viewpoints, gestures, and backgrounds. Red and blue denote the source- and generated-audio waveforms, respectively. The generated frames adopt the reference appearance while following the pose, expression, framing, and scene content of the source video.

and the driving audio, with only coarse prompt-level control; the model does not support explicit, fine-grained expression editing. On the deployment side, the end-to-end pipeline currently runs at 512×512 and distributes encoding, generation, and decoding across three GPUs; reducing the hardware footprint and extending to higher resolutions are left to future work.

K Broader Impact

Audio-video character replacement can support filmmaking, localization, and accessibility, but it also increases the risk of impersonation, non-consensual media, and misinformation. Deployment should require consent and provenance mechanisms, visible disclosure where appropriate, access controls, and compatibility with forensic detection tools.

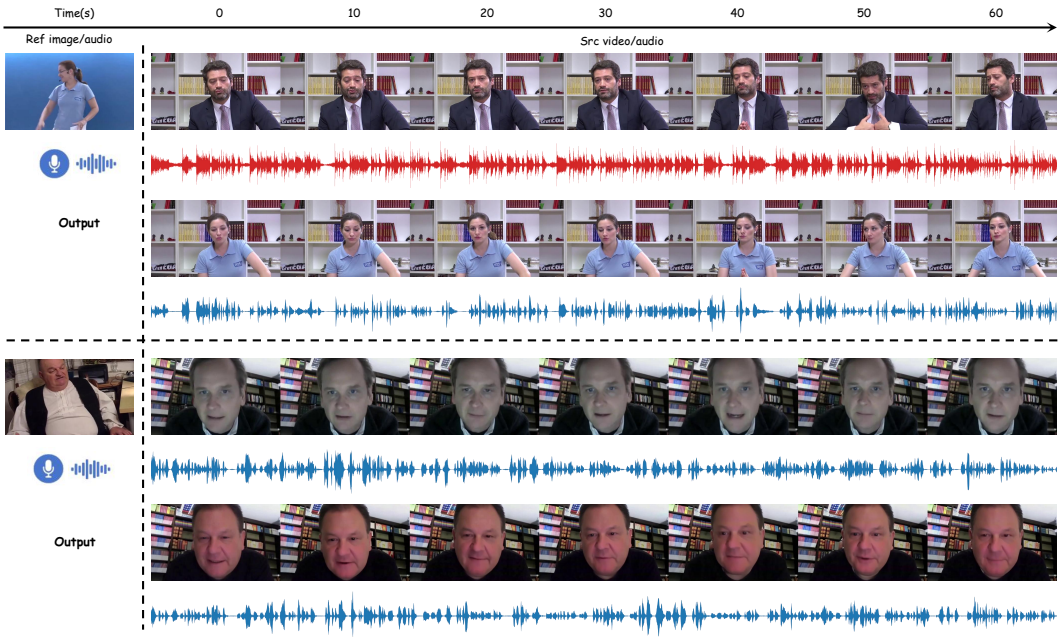


Figure 11: **Additional qualitative results on one-minute videos.** Frames are sampled every 10 seconds from 0 to 60 seconds. Each row group shows the reference image and voice, the source video/audio (red waveform), and the UniSwap output (blue waveform). Across all three sequences, the generated character remains consistent with the reference identity throughout the minute while retaining the source scene and temporal performance.