TEMPOTRACK: TEMPORALLY GROUNDED MULTI-TRACK CONTROL FOR STREAMING JOINT AUDIO-VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Interactive joint audio-video generation requires users to schedule independently timed events and introduce new instructions as generation proceeds. Existing global conditioning and segment-level prompt switching offer limited support for dynamically arriving controls with asynchronous, overlapping intervals. We introduce TEMPOTRACK, a framework for online, temporally grounded multi-track control in streaming joint audio-video generation. Its core mechanism, Hierarchical Context Routing (HCR), connects event-level instructions to blockwise generation through coarse-to-fine conditioning. An online prompt queue admits newly received instructions at block boundaries and selects events relevant to each causal window, while Temporal Attention Bias (TAB) aligns instruction intervals with audio-video query positions on a shared physical timeline through timestamp-dependent cross-attention biases. To enable few-step streaming generation, we directly compose the temporally controllable causal model with a pretrained distillation LoRA and apply self-forcing distribution matching under self-generated histories, without an additional ODE-based initialization stage. We further introduce MTBENCH to evaluate multi-track instruction adherence, temporal alignment, and joint audiovisual quality. Experiments demonstrate asynchronous, overlapping, and dynamically arriving controls alongside strong identity consistency and audiovisual synchronization, enabling streaming audiovisual content to be incrementally directed through independently scheduled events.

1 INTRODUCTION

Joint audio-video generation has made substantial progress in visual quality, audio synthesis, and cross-modal synchronization (HaCohen et al., 2026; Low et al., 2025; Wan et al., 2025; Seedance et al., 2026). Advances in causal modeling and few-step generation further enable streaming audio-visual synthesis (Gan et al., 2026; Su et al., 2026; Cui et al., 2025; 2026), opening up opportunities for interactive applications. In these settings, content unfolds through interaction: future events and their control requirements may not be fully determined before generation begins, and new instructions may arrive after part of the output has already been generated and presented. Streaming generation therefore requires not only temporal continuity, but also the ability to incorporate evolving user intent into the remaining, ungenerated content.

Interactive generation demands finer-grained control than overall prompt switching. For example, as illustrated in Fig. 1, a user may ask a character to smile from 1 to 4 seconds, raise her right hand from 3 to 6 seconds, and begin speaking at 5 seconds. When generation reaches the third second, the user may additionally submit an instruction intended to take effect at the seventh second. These controls govern facial expression, body movement, and speech, with independently specified and potentially overlapping intervals. The system must execute new events at their designated times while preserving ongoing behaviors, character identity, motion continuity, and speech–lip synchronization. The goal is thus not merely to generate a continuous stream, but to allow users to schedule and update individual events as generation proceeds.

Existing streaming generation paradigms do not fully address this requirement. Global conditioning typically assumes that a description of the intended sequence is available before generation, whereas

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.

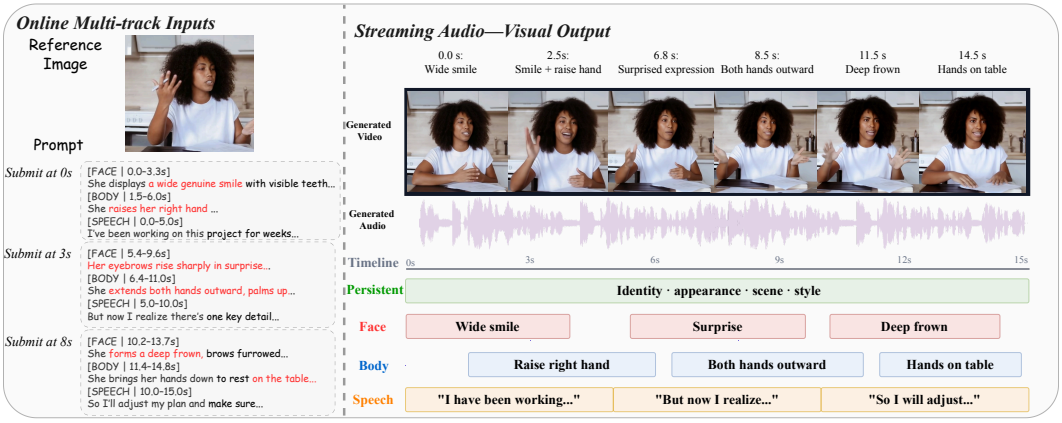

Figure 1: TempoTrack enables temporally grounded multi-track directing of streaming audiovisual performances. Users independently schedule events on multiple tracks along a shared timeline, composing overlapping events and introducing new instructions as generation proceeds.

segment-level prompt switching updates the prompt as a whole at generation boundaries (Yang et al., 2026; Gan et al., 2026; Luo et al., 2026). The former is not naturally suited to future intentions that have yet to arrive. For the latter, a naive approach is to assign a separate prompt to each chunk and switch prompts as generation proceeds. Related audio-driven methods (Gao et al., 2025b), such as SoulX-LiveAct (Zhen et al., 2026), support text-conditioned expression and action control given an audio stream. Extending such control to joint audio–video generation is challenging, as each chunk's prompt must be grounded in the corresponding audio and visual content while preserving cross-modal synchronization. Even when event timings are specified in the prompt, they remain textual descriptions rather than explicit associations.

Motivated by this observation, we introduce TEMPOTRACK, a framework for temporally grounded multi-track control in streaming joint audio-video generation. At the core of TempoTrack is *Hierarchical Context Routing* (HCR), which connects event-level control to blockwise generation through a coarse-to-fine mechanism. An online prompt queue selects events relevant to each causal window, providing coarse routing from event intervals to generation blocks and admitting newly received instructions at subsequent block boundaries. Temporal Attention Bias (TAB) then places instruction intervals and audiovisual query positions on a shared physical timeline and introduces a continuous, timestamp-dependent bias into text cross-attention. This provides fine-grained routing from individual instructions to their corresponding audio and video positions. Together, these mechanisms accommodate overlapping control intervals while assigning temporally differentiated influence to instructions within the same generation block.

To support few-step streaming generation, we directly compose a temporally controllable causal model with a pretrained distillation LoRA and subsequently refine the composed model under self-generated histories. The causal model is obtained through bidirectional control learning followed by teacher-forcing adaptation (Williams & Zipser, 1989). Loading the pretrained distillation LoRA provides few-step initialization without an additional ODE-based initialization stage. We then freeze the causal and distillation adapters and train an additional DMD LoRA adapter through self-forcing distribution matching (Huang et al., 2026; Yin et al., 2024). This design reuses pretrained few-step generation capabilities while adapting the composed model to the histories encountered during autoregressive inference.

Conventional metrics of perceptual quality, identity consistency, and audio–visual synchronization do not fully establish whether an instruction is realized on the intended control track and within its target interval. We therefore introduce MTBENCH to evaluate multi-track instruction adherence, event-level temporal alignment, and joint audiovisual quality. Experiments demonstrate that TempoTrack supports asynchronous, overlapping, and dynamically arriving controls while maintaining strong identity consistency and audiovisual synchronization through few-step causal generation. These results support a formulation of streaming audiovisual generation in which content can be incrementally directed through independently scheduled events.

Our main contributions are as follows:

1. We formulate online, temporally grounded multi-track control for streaming joint audio-video generation, where control instructions have independent, potentially overlapping intervals and may arrive dynamically during generation.
2. We propose Hierarchical Context Routing, which combines block-level prompt admission with token-level Temporal Attention Bias on a shared physical timeline to bridge the mismatch between blockwise generation and event-level control.
3. We develop a direct distillation-LoRA composition and self-forcing correction strategy for few-step streaming generation. A pretrained distillation LoRA provides initialization without an additional ODE-based stage, and an additional DMD LoRA adapter is trained under self-generated histories while the existing adapters remain frozen.
4. We introduce MTBench to evaluate multi-track instruction adherence and temporal alignment, and assess TempoTrack's audiovisual synchronization, identity preservation, and controllability under streaming inference.

2 Related Work

2.1 Video and audio-video generation

Audio-conditioned video generation synthesizes visual content from a given audio signal (Gao et al., 2025a; Huang et al., 2025; Gan et al., 2025), whereas joint audio-video generation models both modalities within a coupled generative process (HaCohen et al., 2026; Low et al., 2025). Recent methods differ in how they align and exchange information across modalities. JavisDiT++ (Liu et al., 2026) introduces fine-grained spatiotemporal alignment, Ovi (Low et al., 2025) uses a twin-backbone architecture, and LTX-2 (HaCohen et al., 2026) connects modality-specific streams through cross-modal attention. Building on LTX-2.3, TempoTrack studies temporally grounded multi-track control during streaming joint audio-video generation.

2.2 Temporally Controlled Video Generation

Recent work explores temporal control over multiple events in video generation. MinT (Wu et al., 2025b) aligns event descriptions with video tokens using rescaled rotary embeddings, TS-Attn (Zhang et al., 2026) uses region-aware event-specific attention biases without additional training, and TIE (Shu et al., 2026) encodes event intervals through duration-normalized rotary integration. These approaches primarily study temporal conditioning in video generation. LongLive (Yang et al., 2026) supports streaming prompt switching and KV recaching for improved continuity, but segment-level switching does not explicitly assign independent temporal intervals to multiple control tracks. TempoTrack combines online instruction admission with fine-grained temporal conditioning for streaming joint audio–video synthesis.

2.3 Streaming generation and distillation

DMD (Yin et al., 2024) enables few-step synthesis, and CausVid (Yin et al., 2025) extends distribution matching to causal video generation. Self-Forcing (Huang et al., 2026) trains causal models on self-generated histories to reduce the train–inference discrepancy, while Causal Forcing (Zhu et al., 2026) combines teacher forcing with ODE initialization. For joint audio-video streaming, Omni-Forcing (Su et al., 2026) introduces asymmetric block-causal alignment to mitigate synchronization drift, and TaoMate (Gan et al., 2026) uses anchor-guided persistent memory for few-step generation. TempoTrack builds on causal adaptation and self-forcing distillation for streaming multi-track control, and directly transfers a pretrained distillation LoRA to the causally adapted model without an additional ODE-based initialization stage.

3 Method

3.1 Problem Formulation

We consider online multi-track control for streaming joint audio–video generation. Given a persistent identity condition p^{per} and a reference first frame, users submit instructions for control tracks.

$$\mathcal{S} = \{s_i = (a_i, \{(c_i^r, \tau_i^r)\}_{r\in\mathcal{R}_i})\}_{i=1}^{N}, \qquad \tau_i^r = [t_{i,r}^{\mathrm{start}}, t_{i,r}^{\mathrm{end}}], \tag{1}$$



Figure 2: Overview of TempoTrack. Hierarchical Context Routing combines block-level admission with token-level temporal bias. Bidirectional and causal training establish the streaming generator, which is composed with a pretrained distillation LoRA before self-forcing DMD optimizes an additional residual on self-generated histories.

where a_i is the arrival time of submission s_i, $\mathcal{R}_i \subseteq \mathcal{R}$ identifies its control tracks, and c_i^r specifies the instruction for track r. Each clause has its own target interval τ_i^r on the generated audiovisual timeline. These intervals may overlap and need not align with generation block boundaries.

The two time variables have different meanings: a_i specifies when an instruction becomes available to the running system, while τ_i^r specifies when its action should appear in the generated sequence. When block k starts, the generator can use only submissions received before that time. An available clause is active when its target interval overlaps the block's output window; a clause for a later window waits until that window is reached, and a clause whose interval has passed is evicted. An instruction received during block k therefore takes effect from a later block and cannot change completed output. This online process aims to realize the scheduled events while preserving ongoing behaviors, identity consistency, and audiovisual coordination.

3.2 Hierarchical Context Routing

HCR separates instruction availability from temporal influence: block-level admission determines which instructions enter the current text context, and token-level temporal bias determines how that context conditions different positions within the block.

Block-level prompt admission. HCR forms the conditioning context independently at each block boundary. A clause is eligible for block k only when its submission is already available on the execution timeline and its target interval intersects the block's output window. Let g_k denote the execution time at which block k begins and $\mathcal{W}_k$ its output-time window. The eligible clause set is

$$\mathcal{A}_k = \{(i, r) \mid a_i \le g_k,\ r \in \mathcal{R}_i,\ \tau_i^r \cap \mathcal{W}_k \neq \varnothing\}\,. \tag{2}$$

The active conditioning sequence concatenates the text tokens of clauses in $\mathcal{A}_k$ with the persistent conditions. Clauses whose intervals lie in future output windows remain pending; once a clause no longer intersects the current window, it is evicted from the active context. Submissions received during block k are therefore considered from the next block boundary onward, and cannot alter completed blocks. This admission rule determines which clauses are available to a block, while the token-level temporal bias below determines how the admitted clauses are weighted within it.

Token-level temporal attention bias. TAB incorporates clause timing into text cross-attention. Let t_u^m denote the physical timestamp of query u in modality $m \in \{v, a\}$. Audio and video query positions are expressed in seconds relative to a shared time origin. For clause c_i^r, define

$$\mu_i^r = \frac{t_{i,r}^{\text{start}} + t_{i,r}^{\text{end}}}{2}, \qquad \sigma_i^r = \max\left(\frac{t_{i,r}^{\text{end}} - t_{i,r}^{\text{start}}}{2}, \epsilon\right), \tag{3}$$

where $\epsilon > 0$ prevents division by zero. Text token j belonging to this clause receives the bias

$$B_{uj}^m = -\lambda \frac{(t_u^m - \mu_i^r)^2}{2(\sigma_i^r)^2}, \tag{4}$$

where $\lambda > 0$ controls its strength. Tokens without a local target interval receive zero temporal bias. Text cross-attention is then

$$\text{Attn}^m = \text{softmax}\left(\frac{Q^m (K^m)^\top}{\sqrt{d_m}} + B^m\right) V^m, \tag{5}$$

where Q^m contains modality-specific queries, K^m and V^m are text keys and values, and d_m is the attention-head dimension.

As shown in Fig 2, the bias softly favors instructions near their target interval centers, with temporal extent scaled by interval duration. Consequently, queries at different times within the same block receive different preferences over the available instructions, while overlapping clauses can condition the same audiovisual positions. TAB is a soft Gaussian temporal prior, which shapes relative attention preference but does not impose exact interval boundaries. Exact block-level availability is determined separately by the admission and eviction rule above, and TAB introduces no additional learnable parameters.

3.3 Causal Adaptation for Streaming Generation

Both bidirectional control learning and causal adaptation are performed using LoRA, with the pretrained backbone weights kept frozen.

Bidirectional control learning. We train the model with full-sequence bidirectional context to learn temporally grounded multi-track generation. Let z_0^v and z_0^a denote the video and audio latents, and let $\mathcal{C}$ comprise the identity condition, reference first frame, and timed text instructions. For each modality $m \in \{v, a\}$, the noisy latent is $z_\sigma^m = (1-\sigma) z_0^m + \sigma \varepsilon^m$, where $\sigma \in [0, 1]$ and $\varepsilon^m \sim \mathcal{N}(0, I)$. Let $\mathbf{z}_\sigma = (z_\sigma^v, z_\sigma^a)$ denote the joint noisy audio–video latent. The model is optimized with the joint audio–video flow-matching (Lipman et al., 2022) objective:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}\left[\sum_{m \in \{v,a\}} \| v_\theta^m(\mathbf{z}_\sigma, \sigma, \mathcal{C}) - (\varepsilon^m - z_0^m) \|_2^2\right]. \tag{6}$$

The resulting bidirectional model provides the initialization for causal adaptation and is subsequently frozen as the target-distribution teacher for distribution matching.

Blockwise causal adaptation. We adapt the bidirectional model into a blockwise autoregressive generator by partitioning audio and video latents into blocks covering corresponding temporal intervals. Let B_k denote the k-th joint audio–video block and $\mathcal{C}_k$ its conditioning inputs. The model denoises B_k conditioned on $\mathcal{C}_k$ and preceding blocks $B_{<k}$.

To align the training context with causal inference, we apply a blockwise mask to modality-specific self-attention and audio–video cross-attention. As illustrated in Fig. 2, each noisy block attends to itself and preceding clean blocks, but not to its clean counterpart, other noisy blocks, or future

content. Clean blocks likewise follow block-causal attention, preventing future information leakage through the conditioning history. TAB is retained in text cross-attention to preserve temporally grounded conditioning.

During training, preceding blocks are populated with ground-truth latents through teacher forcing (Williams & Zipser, 1989), and the flow-matching loss is evaluated only on noisy target tokens:

$$\mathcal{L}_{\text{causal}} = \mathbb{E}_{k,\sigma}\left[\mathcal{L}_{\text{FM}}\left(B_k \mid \mathcal{C}_k, B_{<k}^{\text{clean}}\right)\right]. \tag{7}$$

The resulting model, parameterized by θ_{causal}, supports parallel blockwise training and autoregressive inference with cached history.

3.4 Self-Forcing Distribution Matching

Direct distillation-LoRA initialization. We initialize the few-step student by composing the causally adapted model with a pretrained distillation LoRA and an additional trainable DMD LoRA:

$$\theta_S = \underbrace{\theta_{\text{causal}} + \Delta\theta_{\text{distill}}}_{\theta_{\text{CD}}\ \text{(frozen)}} + \Delta\theta_{\text{DMD}}. \tag{8}$$

Here, θ_{causal} denotes the complete model after causal adaptation, and $\Delta\theta_{\text{distill}}$ is the pretrained adapter for few-step denoising. This composition avoids an additional ODE-based initialization stage. Only $\Delta\theta_{\text{DMD}}$ is optimized within the student.

Self-forcing distribution matching. Following self-forcing (Huang et al., 2026), the student autoregressively generates an audio–video sequence, conditioning each block on its preceding outputs:

$$\widehat{z}_k = F_{\theta_S}\left(\xi_k, \widehat{z}_{<k}, \mathcal{C}_k\right), \tag{9}$$

where F_{θ_S} denotes the few-step block sampler, ξ_k is the initial noise, and $\mathcal{C}_k$ contains the conditions available to block k. This rollout exposes the student to self-generated histories rather than the ground-truth context used during causal adaptation. We then apply DMD (Yin et al., 2024) to align the student distribution with the temporally controllable distribution represented by the frozen bidirectional teacher. A trainable bidirectional critic estimates the student distribution, and the DMD adapter is optimized from the difference between the critic and teacher predictions. The teacher and critic use matching temporal conditions; the critic is trained on detached student samples while the backbone remains frozen.

Sliding-Window Inference with Sink Repositioning. During inference, the learned DMD LoRA adapter is composed with the causal model and the pretrained distillation LoRA. Besides, we restrict attention to a sliding window of recent blocks to bound the context length. To preserve identity consistency, we additionally retain the first block as an attention sink, which remains accessible as the window advances. As the temporal distance between the sink and the active window grows, we apply an offset to the sink block's RoPE positions so that they immediately precede those of the current window. This repositioning keeps the sink contiguous with the active context in rotary positional coordinates.

4 Experiments

4.1 Implementation Details

Our models are initialized from LTX-2.3 (HaCohen et al., 2026) and adapted using LoRA (Hu et al., 2022) throughout the three training stages. Unless otherwise specified, both the LoRA rank and scaling factor are set to 128. The bidirectional and causal adaptation stages each comprise 6,750 optimization steps with a learning rate of 10^{-4}. The self-forcing DMD stage comprises 1,000 total optimization updates (200 student updates and 800 critic updates), with learning rates of 5×10^{-6} and 10^{-5} for the student and critic, respectively. All stages use AdamW (Loshchilov & Hutter, 2019). Training is conducted on eight NVIDIA H100 GPUs.

4.2 Data and Evaluation Protocol

Training data and MTBench. We construct training samples from AVSpeech (Ephrat et al., 2018) through hierarchical segmentation and annotation. We use Whisper (Radford et al., 2023) to obtain

Table 1: Quantitative comparison of generation quality, audio–visual synchronization, controllability, and inference efficiency.

Method	ASE ↑	IQA ↑	DINO-S ↑	Sync-C ↑	Sync-D ↓	WER ↓	Body Event ↑	Face Success ↑	Speech Accuracy ↑	End-to-end FPS ↑
OVI (Low et al., 2025)	2.8947	4.5025	0.9002	1.8842	13.3545	1.2755	0.4000	0.6019	0.1327	1.650
MOVA (Team et al., 2026)	3.0379	**4.7237**	0.8174	4.0395	8.5334	0.3048	0.4951	0.7727	0.9176	0.267
JavisDiT++ (Liu et al., 2026)	2.8577	4.3030	–	0.4416	11.9864	2.5271	0.1743	0.5172	0.0225	1.316
LTX-2.3 (HaCohen et al., 2026)	3.1314	4.6761	0.8701	2.6330	11.7135	**0.0011**	0.6972	0.8214	0.9069	0.704
OmniForcing (Su et al., 2026)	2.9768	4.5008	–	0.6627	13.1318	1.0546	0.3298	0.6500	0.2695	13.289
TaoMate (Gan et al., 2026)	**3.1625**	4.6891	0.8980	3.2873	10.0910	0.0871	0.3714	**0.9700**	0.9421	14.250
Ours	3.0338	4.6854	**0.9442**	**6.3495**	**8.4674**	0.1448	**0.7174**	0.9250	**0.9973**	**15.367**

speech transcripts and sentence-level timestamps. Clips shorter than two seconds are extended with adjacent utterances or nearby silent intervals. Consecutive clips are grouped into samples containing two to five clips with a total duration of 5–15 seconds. Gemini (Google DeepMind, 2026) annotates facial expressions, body movements, speech, and background music for each clip. These annotations and their temporal intervals are combined chronologically into structured multi-track prompts.

To evaluate temporally grounded multi-track control in joint audio–video generation, we introduce MTBENCH, a benchmark comprising 56 cases. Each case pairs a reference image generated by Qwen-Image (Wu et al., 2025a) with a structured prompt generated by Gemini-3.1-Pro (Google DeepMind, 2026), specifying the control tracks over explicit temporal intervals. The benchmark covers diverse human appearances and visual compositions, including frontal and profile views as well as half-body and full-body shots. The benchmark construction prompt is provided in appendix B. All evaluated methods use the same reference image and underlying control instructions for each case, with prompts adapted to their respective conditioning interfaces.

Evaluation metrics. We assess perceptual quality (IQA) and aesthetic quality (ASE) using Q-Align (Wu et al., 2023). Audiovisual synchronization is measured by Sync-C and Sync-D (Chung & Zisserman, 2016), while identity consistency is evaluated using DINOv2 (Oquab et al., 2024) cosine similarity between generated frames and the reference image (DINO-S). We use `gemini-3-flash-preview` to evaluate timed body and facial-expression instructions, yielding Body Event and Face Success, respectively. Separately, `faster-whisper large-v3` provides ASR transcripts and word timestamps for computing word error rate (WER) and Speech Accuracy. WER measures global transcription errors, while Speech Accuracy measures the fraction of reference words correctly recognized within their assigned time intervals, with a tolerance of ± 0.25 seconds applied to word midpoints. The detailed evaluation protocol is provided in appendix C.

4.3 MAIN COMPARISON

We compare TempoTrack with representative joint audio–video generators: OVI (Low et al., 2025), LTX-2.3 (HaCohen et al., 2026), MOVA (Team et al., 2026), JavisDiT++ (Liu et al., 2026), OmniForcing (Su et al., 2026), and TaoMate (Gan et al., 2026). To accommodate their conditioning interfaces, each structured benchmark prompt is converted into the corresponding model-specific format while preserving its semantic content, control attributes, and temporal ordering as closely as possible. For methods without native support for structured multi-track inputs, control clauses are serialized into temporally ordered natural-language instructions. All methods generate 321-frame videos at 24 fps for quality and control evaluation. During streaming inference, each block contains three video latent frames.

Multi-track controllability. As shown in Table 1, TempoTrack achieves the highest Body Event and Speech Accuracy scores, at 0.7174 and 0.9973, respectively. These results indicate improved adherence on the evaluated body and speech tracks. TaoMate obtains a higher Face Success score of 0.9700, but its lower DINO-S and weaker synchronization scores indicate a trade-off between highly exaggerated facial expressions and identity or audiovisual consistency. TempoTrack reaches a Face Success score of 0.9250, ranking second and exceeding LTX-2.3 and MOVA. The results therefore show strong gains on body and speech control together with competitive facial-expression control, rather than uniform superiority on every metric.

Audiovisual quality and efficiency. TempoTrack obtains the highest reported DINO-S and Sync-C scores and the lowest Sync-D, indicating strong identity consistency and audiovisual synchroniza-

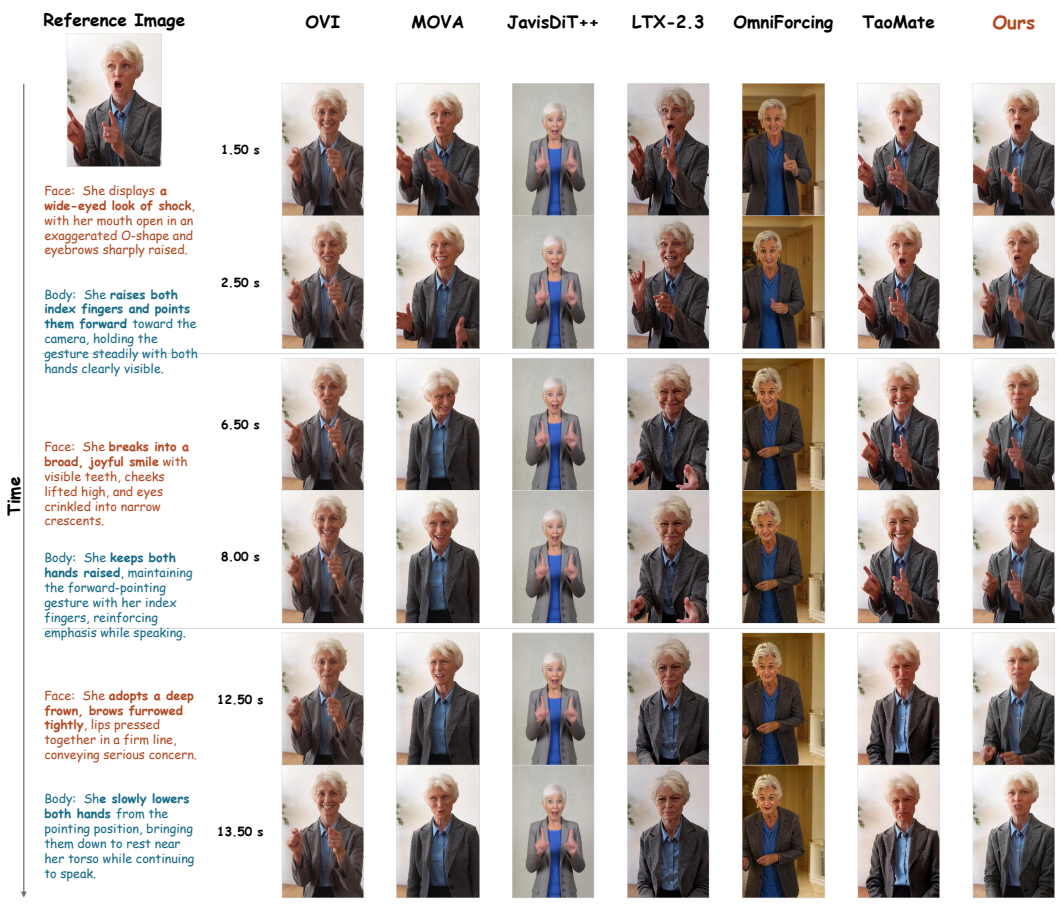

Figure 3: Qualitative comparison. Our method achieves the best identity preservation and instruction adherence.

tion. Its IQA and ASE scores are competitive, although neither ranks first, and its WER remains higher than those of LTX-2.3 and TaoMate. We define end-to-end (E2E) latency as the interval from request submission through prompt processing, conditioning, generator setup, recurrent DiT denoising, parallel video/audio decoding, and MP4 encoding. Under this definition, TempoTrack reaches 15.367 E2E FPS, the highest among the listed configurations.

Qualitative Results. As shown in Fig. 3, TempoTrack preserves the subject's identity most consistently among the compared methods, maintaining stable facial appearance and clothing across the generated sequence. It also produces body actions that closely match the specified prompt events and remain aligned with their intended temporal segments. Compared with the baselines, the generated actions exhibit fewer visible identity drifts and less ambiguity in event execution. These qualitative observations are consistent with the quantitative gains in identity consistency and multi-track controllability reported above.

4.4 Ablation Studies

Temporal Conditioning. We compare our method with TIE (Shu et al., 2026) and a hard-mask variant, which restricts audio and video queries to temporally corresponding text keys in their cross-attention layers. Table 2 shows that HCR achieves the highest Body Event and Speech Accuracy scores, together with the strongest audiovisual synchronization. Compared with the hard mask, HCR improves Body Event from 0.6713 to 0.7174 and Speech Accuracy from 0.9830 to 0.9973, at the cost of increasing WER from 0.0197 to 0.1448. The hard mask retains the highest Face Success score, with HCR ranking second. These results support the effectiveness of soft temporal conditioning for body and speech control, while revealing a trade-off in facial-expression and speech-quality control.

Table 2: Ablation study of different temporal conditioning mechanisms. The best and second-best results are highlighted in bold and underlined, respectively.

Method	ASE ↑	IQA ↑	DINO-S ↑	Sync-C ↑	Sync-D ↓	WER ↓	Body Event ↑	Face Success ↑	Speech Accuracy ↑
LTX-2.3	**3.1314**	4.6761	0.8701	2.6330	11.7135	**0.0011**	0.6972	0.8214	0.9069
TIE (Shu et al., 2026)	2.9420	4.5284	0.9007	3.3761	11.5564	0.0255	0.5357	0.6857	0.9707
Hard Mask	2.9516	4.5181	0.8750	3.2038	11.4832	0.0197	0.6713	**0.9722**	0.9830
Ours	3.0338	**4.6854**	**0.9442**	**6.3495**	**8.4674**	0.1448	**0.7174**	0.9250	**0.9973**

Training and Distillation-LoRA Composition. Table 3 compares the bidirectional model, causal adaptation with and without the pretrained distillation LoRA, and the final self-forcing DMD model. Causal adaptation improves visual quality, identity consistency, and synchronization relative to the bidirectional model, while introducing a moderate speech-recognition trade-off. Direct composition with the distillation LoRA further improves perceptual quality and identity preservation, but weakens synchronization relative to the non-distilled causal model. Self-forcing DMD then recovers and improves identity and audiovisual synchronization under self-generated histories, with only modest reductions in perceptual quality. These results support direct distillation-LoRA composition followed by self-forcing correction as a practical approach to few-step streaming generation.

Table 3: Ablation study across different training stages and causal/distilled variants. The best and second-best results are highlighted in bold and underlined, respectively.

Method	ASE ↑	IQA ↑	DINO-S ↑	Sync-C ↑	Sync-D ↓	WER ↓
Stage 1: Bidirectional	2.8854	4.5033	0.8789	3.7635	11.2714	**0.0048**
Stage 2 (w/o Distillation)	3.0000	4.5835	0.9008	5.9151	8.6672	0.0902
Stage 2 (w/ Distillation)	**3.1071**	**4.7261**	0.9298	5.4736	9.0372	0.1586
Stage 3 (Self-forcing DMD)	3.0338	4.6854	**0.9442**	**6.3495**	**8.4674**	0.1448

Sink Blocks and Repositioning. Table 4 evaluates the effects of retaining the first block as an attention sink and repositioning its RoPE (Su et al., 2023)coordinates. Adding the sink improves DINO-S and audiovisual synchronization while reducing WER from 0.2282 to 0.1825. Repositioning provides further improvements in identity consistency and synchronization and lowers WER to 0.1448, with slight reductions in ASE and IQA relative to the sink-only configuration. These results support retaining and repositioning the sink for identity preservation and audiovisual coordination. Additional qualitative comparisons for the ablation study are provided in Appendix D.

Table 4: Ablation study of Sink and Sink Reposition. The best and second-best results are highlighted in bold and underlined, respectively.

Sink	Sink Reposition	ASE ↑	IQA ↑	DINO-S ↑	Sync-C ↑	Sync-D ↓	WER ↓
×	×	**3.0535**	4.6910	0.9310	5.8951	8.5983	0.2282
✓	×	3.0438	**4.6950**	0.9411	6.2764	8.5021	0.1825
✓	✓	3.0338	4.6854	**0.9442**	**6.3495**	**8.4674**	**0.1448**

5 CONCLUSION

We proposed TEMPOTRACK, a framework for streaming joint audio–video generation with independently timed, potentially overlapping control instructions that can arrive during generation. Hierarchical Context Routing combines block-level prompt admission with token-level temporal attention bias to align event instructions with audio–video positions on a shared timeline. Direct composition with a pretrained distillation LoRA, followed by self-forcing correction, enables few-step causal generation without additional ODE-based initialization. We also introduced MTBENCH to evaluate multi-track instruction adherence and temporal alignment. Experiments demonstrate the framework's controllability, identity consistency, and audio–visual synchronization, supporting incremental, event-level directing of streaming audiovisual content.

AI USE STATEMENT

We used generative AI tools to assist with training-data annotation, benchmark construction, experimental evaluation, and manuscript polishing. These uses include generating benchmark reference images and multi-track instructions, as well as evaluating instruction adherence. The authors have reviewed all AI-assisted work and take responsibility for the entire content of this work, including all text, data, and analyses produced with the aid of generative AI.

REFERENCES

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.

Jiahao Cui, Hui Li, Yun Zhan, Hanlin Shang, Kaihui Cheng, Yuqi Ma, Shan Mu, Hang Zhou, Jingdong Wang, and Siyu Zhu. Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21086–21095. IEEE, 2025.

Liyuan Cui, Wentao Hu, Wenyuan Zhang, Zesong Yang, Fan Shi, and Xiaoqiang Liu. Avatarforcing: One-step streaming talking avatars via local-future sliding-window denoising, 2026. URL https://arxiv.org/abs/2603.14331.

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.

Qijun Gan, Ruizi Yang, Jianke Zhu, Shaofei Xue, and Steven Hoi. Omniavatar: Efficient audio-driven avatar video generation with adaptive body animation, 2025. URL https://arxiv.org/abs/2506.18866.

Qijun Gan, Chenwei Zhang, Meiguang Jin, Junfeng Ma, and Qiu Shen. Taomate: Anchor-guided memory bridging evolving and reference states for real-time audio-video digital human generation, 2026. URL https://arxiv.org/abs/2607.24359.

Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Dechao Meng, Jinwei Qi, Penchong Qiao, Zhen Shen, Yafei Song, Ke Sun, Linrui Tian, Guangyuan Wang, Qi Wang, Zhongjian Wang, Jiayu Xiao, Sheng Xu, Bang Zhang, Peng Zhang, Xindi Zhang, Zhe Zhang, Jingren Zhou, and Lian Zhuo. Wan-s2v: Audio-driven cinematic video generation, 2025a. URL https://arxiv.org/abs/2508.18621.

Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Dechao Meng, Jinwei Qi, Penchong Qiao, Zhen Shen, Yafei Song, Ke Sun, Linrui Tian, Guangyuan Wang, Qi Wang, Zhongjian Wang, Jiayu Xiao, Sheng Xu, Bang Zhang, Peng Zhang, Xindi Zhang, Zhe Zhang, Jingren Zhou, and Lian Zhuo. Wan-s2v: Audio-driven cinematic video generation, 2025b. URL https://arxiv.org/abs/2508.18621.

Google DeepMind. Gemini 3.1 Pro Model Card, 2026. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, Eitan Richardson, Guy Shiran, Itay Chachy, Jonathan Chetboun, Michael Finkelson, Michael Kupchick, Nir Zabari, Nitzan Guetta, Noa Kotler, Ofir Bibi, Ori Gordon, Poriya Panet, Roi Benita, Shahar Armon, Victor Kulikov, Yaron Inger, Yonatan Shiftan, Zeev Melumian, and Zeev Farbman. Ltx-2: Efficient joint audio-visual foundation model, 2026. URL https://arxiv.org/abs/2601.03233.

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. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

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.

Yubo Huang, Hailong Guo, Fangtai Wu, Shifeng Zhang, Shijie Huang, Qijun Gan, Lin Liu, Sirui Zhao, Enhong Chen, Jiaming Liu, and Steven Hoi. Live avatar: Streaming real-time audio-driven avatar generation with infinite length, 2025. URL https://arxiv.org/abs/2512.04677.

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

Kai Liu, Yanhao Zheng, Kai Wang, Shengqiong Wu, Rongjunchen Zhang, Jiebo Luo, Dimitrios Hatzinakos, Ziwei Liu, Hao Scofield Fei, and Tat-Seng Chua. Javisdit++: Unified modeling and optimization for joint audio-video generation. In *International Conference on Learning Representations*, volume 2026, pp. 150592–150618, 2026.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Chetwin Low, Weimin Wang, and Calder Katyal. Ovi: Twin backbone cross-modal fusion for audio-video generation, 2025. URL https://arxiv.org/abs/2510.01284.

Yawen Luo, Xiaoyu Shi, Junhao Zhuang, Yutian Chen, Quande Liu, Xintao Wang, Pengfei Wan, and Tianfan Xue. Shotstream: Streaming multi-shot video generation for interactive storytelling. In *European Conference on Computer Vision*, pp. 193–209. Springer, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

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.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, Mojie Chi, Xuyan Chi, Jian Cong, Qinpeng Cui, Fei Ding, Qide Dong, Yujiao Du, Haojie Duanmu, Junliang Fan, Jiarui Fang, Jing Fang, Zetao Fang, Chengjian Feng, Yu Gao, Diandian Gu, Dong Guo, Hanzhong Guo, Qiushan Guo, Boyang Hao, Hongxiang Hao, Haoxun He, Jiaao He, Qian He, Tuyen Hoang, Heng Hu, Ruoqing Hu, Yuxiang Hu, Jiancheng Huang, Weilin Huang, Zhaoyang Huang, Zhongyi Huang, Jishuo Jin, Ming Jing, Ashley Kim, Shanshan Lao, Yichong Leng, Bingchuan Li, Gen Li, Haifeng Li, Huixia Li, Jiashi Li, Ming Li, Xiaojie Li, Xingxing Li, Yameng Li, Yiying Li, Yu Li, Yueyan Li, Chao Liang, Han Liang, Jianzhong Liang, Ying Liang, Wang Liao, J. H. Lien, Shanchuan Lin, Xi Lin, Feng Ling, Yue Ling, Fangfang Liu, Jiawei Liu, Jihao Liu, Jingtuo Liu, Shu Liu, Sichao Liu, Wei Liu, Xue Liu, Zuxi Liu, Ruijie Lu, Lecheng Lyu, Jingting Ma, Tianxiang Ma, Xiaonan Nie, Jingzhe Ning, Junjie Pan, Xitong Pan, Ronggui Peng, Xueqiong Qu, Yuxi Ren, Yuchen Shen, Guang Shi, Lei Shi, Yinglong Song, Fan Sun, Li Sun, Renfei Sun, Wenjing Tang, Boyang Tao, Zirui Tao, Dongliang Wang, Feng Wang, Hulin Wang, Ke Wang, Qingyi Wang, Rui Wang, Shuai Wang, Shulei Wang, Weichen Wang, Xuanda Wang, Yanhui Wang, Yue Wang, Yuping Wang, Yuxuan Wang, Zijie Wang, Ziyu Wang, Guoqiang Wei, Meng Wei, Di Wu, Guohong Wu, Hanjie Wu, Huachao Wu, Jian Wu, Jie Wu, Ruolan Wu, Shaojin Wu, Xiaohu Wu, Xinglong Wu, Yonghui Wu, Ruiqi Xia, Xin Xia, Xuefeng Xiao, Shuang Xu, Bangbang Yang, Jiaqi Yang, Runkai Yang, Tao Yang, Yihang Yang, Zhixian Yang, Ziyan Yang, Fulong Ye, Bingqian Yi, Xing Yin, Yongbin You, Linxiao Yuan, Weihong Zeng, Xuejiao Zeng, Yan Zeng, Siyu Zhai, Zhonghua Zhai, Bowen Zhang, Chenlin Zhang, Heng Zhang, Jun Zhang, Manlin Zhang, Peiyuan Zhang, Shuo Zhang, Xiaohe Zhang, Xiaoying Zhang, Xinyan Zhang, Xinyi Zhang, Yichi Zhang, Zixiang Zhang, Haiyu Zhao, Huating Zhao, Liming Zhao, Yian Zhao, Guangcong Zheng, Jianbin Zheng, Xiaozheng

Zheng, Zerong Zheng, Kuan Zhu, and Feilong Zuo. Seedance 2.0: Advancing video generation for world complexity, 2026. URL https://arxiv.org/abs/2604.14148.

Zhilei Shu, Shangwen Zhu, Zihang Liang, Xiaofan Li, Qianyu Peng, Xinyu Cui, Bo Ye, Yiming Li, Fan Cheng, Jian Zhao, Yang Cao, Zheng-Jun Zha, and Ruili Feng. Tie: Time interval encoding for video generation over events, 2026. URL https://arxiv.org/abs/2605.10543.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023. URL https://arxiv.org/abs/2104.09864.

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.

OpenMOSS Team, Donghua Yu, Mingshu Chen, Qi Chen, Qi Luo, Qianyi Wu, Qinyuan Cheng, Ruixiao Li, Tianyi Liang, Wenbo Zhang, Wenming Tu, Xiangyu Peng, Yang Gao, Yanru Huo, Ying Zhu, Yinze Luo, Yiyang Zhang, Yuerong Song, Zhe Xu, Zhiyu Zhang, Chenchen Yang, Cheng Chang, Chushu Zhou, Hanfu Chen, Hongnan Ma, Jiaxi Li, Jingqi Tong, Junxi Liu, Ke Chen, Shimin Li, Shiqi Jiang, Songlin Wang, Wei Jiang, Zhaoye Fei, Zhiyuan Ning, Chunguo Li, Chenhui Li, Ziwei He, Zengfeng Huang, Xie Chen, and Xipeng Qiu. Mova: Towards scalable and synchronized video-audio generation, 2026. URL https://arxiv.org/abs/2602.08794.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models, 2025. URL https://arxiv.org/abs/2503.20314.

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

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025a. URL https://arxiv.org/abs/2508.02324.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-align: Teaching lmms for visual scoring via discrete text-defined levels, 2023. URL https://arxiv.org/abs/2312.17090.

Ziyi Wu, Aliaksandr Siarohin, Willi Menapace, Ivan Skorokhodov, Yuwei Fang, Varnith Chordia, Igor Gilitschenski, and Sergey Tulyakov. Mind the time: Temporally-controlled multi-event video generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23989–24000. IEEE, 2025b.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, YINGCONG CHEN, Yao Lu, Song Han, and Yukang Chen. Longlive: Real-time interactive long video generation. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 90037–90059, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/91a1610c6ed9e02d33f826b46f472b92-Paper-Conference.pdf.

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, 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, 2025.

Hongyu Zhang, Yufan Deng, Zilin Pan, Peng-Tao Jiang, Bo Li, Qibin Hou, Zhen Dong, Zhiyang Dou, and Daquan Zhou. TS-attn: Temporal-wise separable attention for multi-event video generation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=QixNhagZ9t.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li. Pytorch fsdp: Experiences on scaling fully sharded data parallel, 2023. URL https://arxiv.org/abs/2304.11277.

Dingcheng Zhen, Xu Zheng, Ruixin Zhang, Zhiqi Jiang, Yichao Yan, Ming Tao, and Shunshun Yin. Soulx-liveact: Towards hour-scale real-time human animation with neighbor forcing and convkv memory, 2026. URL https://arxiv.org/abs/2603.11746.

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. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=BYInOck3gr.

APPENDIX

A TRAINING DETAILS

We initialize TempoTrack from LTX-2.3 and perform all three training stages using LoRA with the backbone frozen. Distributed training uses Hugging Face Accelerate with Fully Sharded Data Parallel (FSDP) (Zhao et al., 2023). All stages use AdamW, BF16 mixed precision, gradient checkpointing, and a gradient clipping threshold of 1.0. The temporal attention bias follows Eq. 4, with strength controlled by λ and temporal width determined by each instruction's target interval. Table 5 summarizes the main hyperparameters.

For four-step self-forcing DMD, the student composes the causal LoRA and the pretrained distillation LoRA with strengths 1.0 and 0.5, respectively. Both adapters remain frozen while an additional rank-128 DMD LoRA is optimized under self-generated histories. A separate rank-128 critic LoRA is trained with flow matching. Training comprises 1,000 total optimization updates: 200 student updates and 800 critic updates, scheduled as one student update and four critic updates per five-update cycle. The nominal denoising timestep indices are $[999, 909, 725, 421]$, and the teacher uses video/audio guidance scales of 5.0/9.0.

Table 5: Training hyperparameters. Stage-3 learning rates are listed as student/critic.

Setting	Stage 1	Stage 2	Stage 3
Learning rate	10^{-4}	10^{-4}	5×10^{-6} / 10^{-5}
LR schedule	Cosine	Linear	Constant
Warmup steps	500	0	0
Trainable LoRA rank / alpha	128 / 128	128 / 128	128 / 128
Batch size per GPU	1	1	1
Distributed training	FSDP	FSDP	FSDP
Video latent frames per block	—	3	3

B DETAILS OF BENCHMARK

We construct the benchmark using Qwen-Image to generate reference images and Gemini-3.1-Pro to generate the corresponding multi-track instructions. The image-generation prompt is:

Image-generation prompt.

```
Generate a photorealistic image of a single person in either
a full-body or half-body composition.  Vary the person's
age, gender, appearance, hairstyle, and clothing across
images.  Use a frontal or three-quarter view, keeping the
face clearly visible, well-lit, and in sharp focus, with
natural facial features and no major occlusions.  Choose a
relaxed, natural pose with anatomically plausible body
proportions and clearly rendered hands when visible.  Place
the person in a visually rich, coherent environment, such as
a furnished living room, café, office, garden, or urban
street.  Include detailed textures, realistic lighting, and
sufficient contrast to distinguish the person from the
background.  Keep the composition focused on the main
subject, without additional people, text, logos, or
watermarks.
```

The prompt used to generate the multi-track instructions is provided below:

Multi-track instruction-generation prompt.

```
SYSTEM_PROMPT = """You create controlled English prompts for
an image-conditioned talking-head video benchmark.
```

```
Inspect the supplied reference image carefully before
choosing any action.  Keep the visible person's identity,
age, clothing, hairstyle, framing, camera, and background
stable.  If several people are visible, preserve all of them
and choose one clearly described lead speaker.  Do not
identify celebrities or infer private facts.

First determine internally how much of the lead person is
actually visible:  face only, head and shoulders, upper
body, or full body; whether the left hand and right hand are
visible; and whether either hand is occupied by an object.
Do not output this analysis.  Use it to avoid impossible
controls.

Return exactly one JSON object and no Markdown.  Required
schema:
{
"source_description":  "one concise identity/scene anchor
grounded only in the image",
"segments":  [
{
"facial_expression":  "one strong observable expression",
"body_action":  "one feasible visible pose, head, hand or
upper-body action, or null",
"background_music":  "an audio-bed description",
"character_speech":  "one natural English line, or null when
a large action should be silent"
},
{"facial_expression":  "...", "body_action":  "...",
"background_music":  "...", "character_speech":  "..."},
{"facial_expression":  "...", "body_action":  "...",
"background_music":  "...", "character_speech":  "..."}
]
}

Rules:
- Output exactly three segments and all text must be
English.
- Make the three facial expressions visibly different,
exaggerated, and easy to judge.  Good choices include
dramatically raised eyebrows, a broad laugh, a wide smile
with visible teeth, strongly lifted cheeks, narrowed joyful
eyes, surprise, or a deep frown.
- Facial-expression text describes an expression, not speech
articulation.  Never request continuous mouth movement,
moving lips, lip sync, talking-mouth motion, repeated jaw
motion, or keeping the mouth moving while speech plays.
- Every body action must be possible within the original
crop.  Never invent an off-screen hand, arm, torso, leg, or
object, and never assume the camera will zoom out or
reframe.
- If both hands and enough upper body are clearly visible
and unoccupied, use one or two large, easy-to-score actions
such as broad two-hand gestures, raising the left hand,
raising the right hand, extending one visible hand, or
pronounced nodding.
- If only one hand is visible or the other hand holds an
object, act only with the visible free hand.  Do not make
the person drop, replace, or switch an object.
- If the image is face-only or a tight head-and-shoulders
crop, set body_action to null in at least two segments.  At
most one segment may use a small visible head action such as
a clear head turn or pronounced nod.
```

```
- Use at most two large body-action segments even when the
full body is visible.  Keep at least one segment free of a
large body action.
- In a segment containing a large hand, arm, or two-hand
action, set character_speech to null so the large action is
explicitly tested during a speech gap.  Do not write a
Silence clause or phrases such as "no speech"; use JSON
null.
- In every segment without a large body action, provide a
natural speech line.  The whole plan must contain at least
one non-null speech line.
- Use short training-style phrases, not cinematography
prose, for facial_expression/body_action/background_music.
- Each non-null speech line must be coherent with the other
lines, 7--14 words, easy to pronounce, and contain no stage
directions.
- Music is non-vocal only.  Prefer quiet room ambience in
one segment, soft instrumental music in one, and no music in
one.  Never request singing, lyrics, chanting, or a vocal
soundtrack.
- Do not include timestamps; the caller assigns
independently staggered times.
"""
```

C VLM Evaluation Protocol and Prompts

Setup and sampling. We use `gpt-4o` to extract atomic event/state targets, cached and shared across methods, and `gemini-3-flash-preview` for visual judgment. Both use temperature 0 and a single user message without an additional system prompt. Each target interval is divided into non-overlapping 0.5-second windows; a final partial window is retained if it lasts at least 0.2 seconds. We sample the nearest frames at 20%, 50%, and 80% of each window and assemble a chronological JPEG contact sheet with timestamps. BODY uses full frames; FACE uses Haar-detected face crops with temporal tracking and box propagation/interpolation. No audio is provided.

Body Event. A BODY event succeeds if visible evidence shows that it occurs at least once within the target interval. Body Event is the number of successful evaluable event targets divided by the total number of evaluable event targets, pooled across cases. Controls without an event target are excluded. State judgments are separate diagnostics: a state window is positive when at least two of its three frames satisfy the target. For BODY states marked as final 40%, only windows with midpoints in the final 40% of the interval are judged.

Face Success. Keyword matching over the instruction and parsed targets selects occurrence or duration mode. Smile/grin variants take precedence and use duration mode; frown/furrow, eyebrow/brow, surprise/astonishment, wide-eyed, blink, and wink variants use occurrence mode; other targets default to duration mode. Occurrence mode requires one clear appearance of the principal action or expression. Duration mode labels each window positive when the expression is established in at least two of its three frames; one positive valid window suffices for control-level success. FACE searches the entire target interval regardless of state-scope metadata. Face Success is the fraction of successful evaluable controls pooled across cases.

FACE filtering. Cases with a detected source-face height below 15% of the image height are excluded; failed source detection defers filtering to generated frames. Generated-face visibility is good for a direct-detection rate of at least two thirds and median face size of at least 48 pixels, fair for at least one third and 24 pixels, and poor otherwise. Face size is the smaller box dimension in the original frame, including propagated boxes; detection rates count only direct detections. Visibility is computed per window in duration mode and across sampled frames in occurrence mode, then combined with VLM visibility by taking the worse rating. Confidence is capped at 1.0/0.7/0.3 for good/fair/poor. Valid judgments require confidence of at least 0.4 and non-poor visibility. An

occurrence control requires a valid event judgment; a duration control requires at least one valid state window. Confidence concerns the correctness of either Boolean label, not target presence.

Scope. Unevaluable entries are excluded rather than counted as failures; empty denominators yield undefined scores. These metrics assess occurrence within sampled target intervals, without checking out-of-interval activity or onset/offset accuracy. Face Success does not require full-interval persistence. Speech Accuracy is evaluated separately using ASR and word timestamps.

Prompts. The five templates below retain the original wording. Double-braced placeholders supply control text, interval endpoints, parsed targets, and window descriptions. Extraction uses three-decimal timestamps; visual judgment uses two decimals. FACE window descriptions also include detector visibility, direct-detection counts, and original face size. Visual requests include the corresponding contact sheet.

BODY target extraction.

```
Decompose this BODY instruction into atomic visual targets
for evaluation.
Instruction: {{CONTROL_TEXT}}
Allowed interval: [{{START_SECONDS_3DP}}, {{END_SECONDS_3DP}}]
seconds
Definitions:
- event: a visible motion or transition that only needs to
occur at least once, such as raising, lowering, turning,
nodding, or shifting posture.
- state: only an explicitly sustained/resulting posture
required by wording such as hold, keep, remain, rest,
maintain, or return to.
- Do not invent a sustained state merely by rewriting a
transient event's pose. For example, "raises both hands,
palms forward" is one event with a manner/pose detail and
has state=null unless the instruction explicitly says the
pose is held or maintained.
- state_scope: "full" only when the instruction requires an
ongoing state throughout the interval without first
transitioning into it.
- state_scope: "final_40_percent" when the state is reached
as the result of a transition, including patterns such as
"brings ... to rest", "raises ... then holds", "lowers ...
returning to", or "ending with".
- Judge only visible body pose/motion. Do not include
facial expression, emotion, speech, inferred intention, or
camera motion.
Return only JSON in this schema:
{"event": "atomic event description" or null, "state":
"atomic state description" or null, "state_scope": "full"
or "final_40_percent" or null}
```

BODY visual judgment.

```
Evaluate visible BODY motion and posture in this
chronological contact sheet.
Each W window has three ordered frames, labeled a, b, then
c. The sheet contains only the target interval.
Ignore facial expression, emotion, speech, identity, image
quality, and camera motion. Judge only visible
body/head/arm/hand motion and posture. Do not infer hidden
motion.
Original BODY instruction: {{CONTROL_TEXT}}
```

```
Target interval: [{{START_SECONDS_2DP}}, {{END_SECONDS_2DP}}]
seconds
Atomic event target: {{ATOMIC_EVENT_OR_NONE}}
Atomic state target: {{ATOMIC_STATE_OR_NONE}}
State windows to judge: {{STATE_WINDOW_INDEX_LIST}}

Windows:
{{WINDOW_DESCRIPTION_LINES}}

Rules:
- event_present is true only if the chronological frames
provide visible evidence that the event occurs at least
once.
- For every requested state window, present is true only if
the target state is visibly established in at least two of
that window's three frames.
- Return one label for every requested state window and no
others.

Return only JSON:
{"event_present": true or false or null, "state_windows":
[{"index": 0, "present": true}], "notes": "at most 15
words"}
```

FACE target extraction.

```
Decompose this FACE instruction into atomic visible targets
for evaluation.

Instruction: {{CONTROL_TEXT}}
Allowed interval: [{{START_SECONDS_3DP}}, {{END_SECONDS_3DP}}]
seconds

Definitions:
- state: a facial expression or visible facial
configuration expected to persist, such as smiling,
frowning, wide eyes, furrowed brows, or pressed lips.
- event: a temporally localized facial action or transition
that only needs to occur once, such as blinking, winking,
opening/closing the mouth, or changing into an expression.
- If wording describes a transition into an expression,
return both the transition event and the resulting
expression state. Transition verbs include "forms",
"adopts", "shifts to", "changes to", "breaks into",
"becomes", "raises", and "lowers".
- For example, "forms a deep frown" means event="transitions
into a deep frown", state="deep frown", state_scope="final_
40_percent".
- Descriptions such as "displays", "shows", "has", or
"maintains" normally specify an expression already present,
so they have no transition event and use state_scope="full".
- state_scope is "full" when the expression itself is the
requested condition over the interval.
- state_scope is "final_40_percent" when wording describes
entering, forming, or adopting the resulting expression,
even if it does not explicitly say "hold" afterward.
- Judge visible face evidence only. Exclude body movement,
speech content, identity, intention, and camera motion.

Return only JSON:
{"event": "atomic event description" or null, "state":
"atomic state description" or null, "state_scope": "full"
or "final_40_percent" or null}
```

FACE occurrence judgment.

```
Evaluate the target person's visible face in this
chronological face-crop contact sheet.
Each W window has three ordered frames labeled a, b, and c.
The sheet contains only the target interval; do not evaluate
anything outside it.
Judge only the target person's visible face.  Ignore body
motion, speech content, identity, background, and camera
motion.  Do not infer details hidden by blur, occlusion, or
framing.
Labels marked [D] use a directly detected face box, [T] use
a temporally propagated/interpolated box, and [X] mean no
face box was available.
Original FACE instruction:  {{CONTROL_TEXT}}
Target interval:  [{{START_SECONDS_2DP}}, {{END_SECONDS_2DP}}]
seconds
Windows:
{{WINDOW_DESCRIPTION_LINES_WITH_DETECTOR_VISIBILITY}}
Task:  OCCURRENCE detection only.
Target facial action/expression:  {{OCCURRENCE_TARGET}}
Rules:
- Inspect every frame in every W window and return event_
present=true if the core target facial action or expression
appears clearly at least once anywhere in the interval.
- Compare later frames with earlier frames when the target
is a change such as raising eyebrows, widening eyes, or
forming a frown.
- For a compound description, a clear occurrence of its
principal named action is sufficient.  Do not require every
secondary cue to be present simultaneously.
- Intensity modifiers such as "dramatically", "deep",
"wide", "clear", or "strongly" are not hard thresholds.  A
visible but weaker occurrence still counts.
- Raised eyebrows count when the brows visibly move higher
or the eye/brow area becomes clearly more open than earlier
frames.
- Use a low threshold for frown/furrow occurrence.  Count it
as present if ANY ONE of these cues is clearly visible in
any frame:  brows move inward toward each other, brows press
downward, vertical or concentrated wrinkles/creases appear
between the eyebrows or at the center of the forehead, or an
unmistakable frowning brow configuration appears.
- For frown/furrow occurrence, do not require tightly
pressed lips, downturned mouth corners, narrowed eyes, a
strong concerned emotion, or multiple cues at the same time.
A brief or mild but visually clear brow/forehead furrow is
sufficient.
- Ignore transient speech mouth shapes unless the mouth
itself is the principal target.
- confidence measures confidence in the true/false judgment,
not confidence that the target is present.
- event_visibility must be "good", "fair", or "poor" based
on whether the relevant face region can be inspected.
Return only JSON:
{"event_present":  true or false, "event_confidence":  0.0 to
1.0, "event_visibility":  "good" or "fair" or "poor", "state_
windows":  [], "notes":  "at most 15 words"}
```

FACE duration judgment.

```
Evaluate the target person's visible face in this
chronological face-crop contact sheet.
Each W window has three ordered frames labeled a, b, and c.
The sheet contains only the target interval; do not evaluate
anything outside it.
Judge only the target person's visible face.  Ignore body
motion, speech content, identity, background, and camera
motion.  Do not infer details hidden by blur, occlusion, or
framing.
Labels marked [D] use a directly detected face box, [T] use
a temporally propagated/interpolated box, and [X] mean no
face box was available.
Original FACE instruction:  {{CONTROL_TEXT}}
Target interval:  [{{START_SECONDS_2DP}}, {{END_SECONDS_2DP}}]
seconds
Windows:
{{WINDOW_DESCRIPTION_LINES_WITH_DETECTOR_VISIBILITY}}
Task:  DURATION measurement only.
Target facial state:  {{FACIAL_STATE_TARGET}}
State windows to judge:  {{STATE_WINDOW_INDEX_LIST}}
Rules:
- For every requested state window, judge whether the
underlying target expression is established during the
majority of its three frames.
- Judge every W window independently.  The expression may
begin partway through the target interval; never copy one
global true/false label to every window merely because
earlier windows differ from later ones.
- This is a talking-avatar video:  ignore transient mouth
opening, closing, and shape changes caused only by speech
articulation.  Do not require the mouth to remain closed,
open, or unchanged while the person is speaking.
- Infer the underlying expression from stable evidence
across eyebrows, eyes, eyelids, cheeks, mouth-corner
tendency, and overall affect.  A speech mouth shape alone is
neither evidence for nor against the expression.
- Treat intensity words such as "deep", "broad", or
"dramatic" and descriptive details as supporting cues.
Label the semantic expression present when its defining
stable cues are clear, even if the generated intensity is
weaker or not every modifier is simultaneously visible.
- For a frown/concern target, brows visibly drawn inward or
downward with a serious/concerned overall expression are
sufficient even while the mouth is articulating speech.  For
a smile target, consistently lifted mouth corners or cheeks
are sufficient even when speech temporarily changes the
mouth shape.
- Do not accept an expression when its defining stable cues
are visibly contradicted.
- Mouth evidence still matters when it persists across
frames and agrees with other facial cues, such as raised
brows and wide eyes for surprise or lifted cheeks and mouth
corners for a smile.
- confidence measures confidence in the true/false judgment,
not confidence that the target is present.
- face_visibility/event_visibility must be "good", "fair", or
"poor" based on face size, blur, occlusion, and whether the
requested facial details can actually be seen.
```

```
- The detector visibility shown beside each window is based
on face size in the original frame and must be treated as an
upper bound on visibility.
- If the face is unclear, keep the best true/false judgment
but lower confidence and use "poor" visibility.  Do not
claim high confidence from body pose or context.
- Return one label for every requested state window and no
others.

Return only JSON:
{"event_present":  null, "event_confidence":  null, "event_
visibility":  null, "state_windows":  [{"index":  0,
"present":  true, "confidence":  0.0 to 1.0, "face_
visibility":  "good" or "fair" or "poor"}], "notes":  "at
most 15 words"}
```

D Additional Qualitative Results

Figures 4 and 5 provide additional comparisons with baseline methods. TempoTrack better preserves character appearance and produces more natural expressions and body movements. The accompanying videos further illustrate improved audio–visual synchronization. Figure 6 presents qualitative comparisons across training stages. The final Stage 3 model exhibits clearer facial features and more detailed hand structures, illustrating the visual benefits of self-forcing refinement. Figures 7 and 8 showcase additional results from TempoTrack across different characters and scenes.

E Inference Throughput Protocol

We measure end-to-end (E2E) latency from request submission to completion of the locally encoded MP4 file. The measured interval includes prompt processing, conditioning, generator setup, recurrent DiT denoising over all 16 blocks, parallel video and audio decoding, and MP4 encoding. The final copy of the completed file to the experiment directory is excluded. All measurements use batch size one, 480×480 resolution, 361 frames at 24 FPS, four-step DMD, and the same input case and random seed. The reported TempoTrack throughput is computed from two formal warm requests after runtime reuse and prompt-cache warm-up; their pooled E2E latency is 23.492 seconds, corresponding to 15.367 FPS. As a runtime optimization, the encoded text context is cached and reused when the active clause tuple and its associated conditioning settings are unchanged. This cache affects only prompt-processing latency and does not change the model, denoising schedule, or generated outputs.

The four-GPU resident configuration places prompt processing on GPU 0, the 22B Transformer and recurrent DiT on GPU 1, video decoding on GPU 2, and audio decoding and the vocoder on GPU 3. The Transformer and decoders remain resident between warm requests, while request-specific KV caches and temporary tensors are released after each request. Video and audio decoding are concurrent, so their wall time is measured as a parallel stage rather than as the sum of the two decoder times. Since baseline methods use different hardware and runtime configurations, their reported FPS values are provided as configuration-specific throughput measurements.



Figure 4: Additional qualitative comparisons with baseline methods, illustrating character consistency, facial expressions, and body movements.

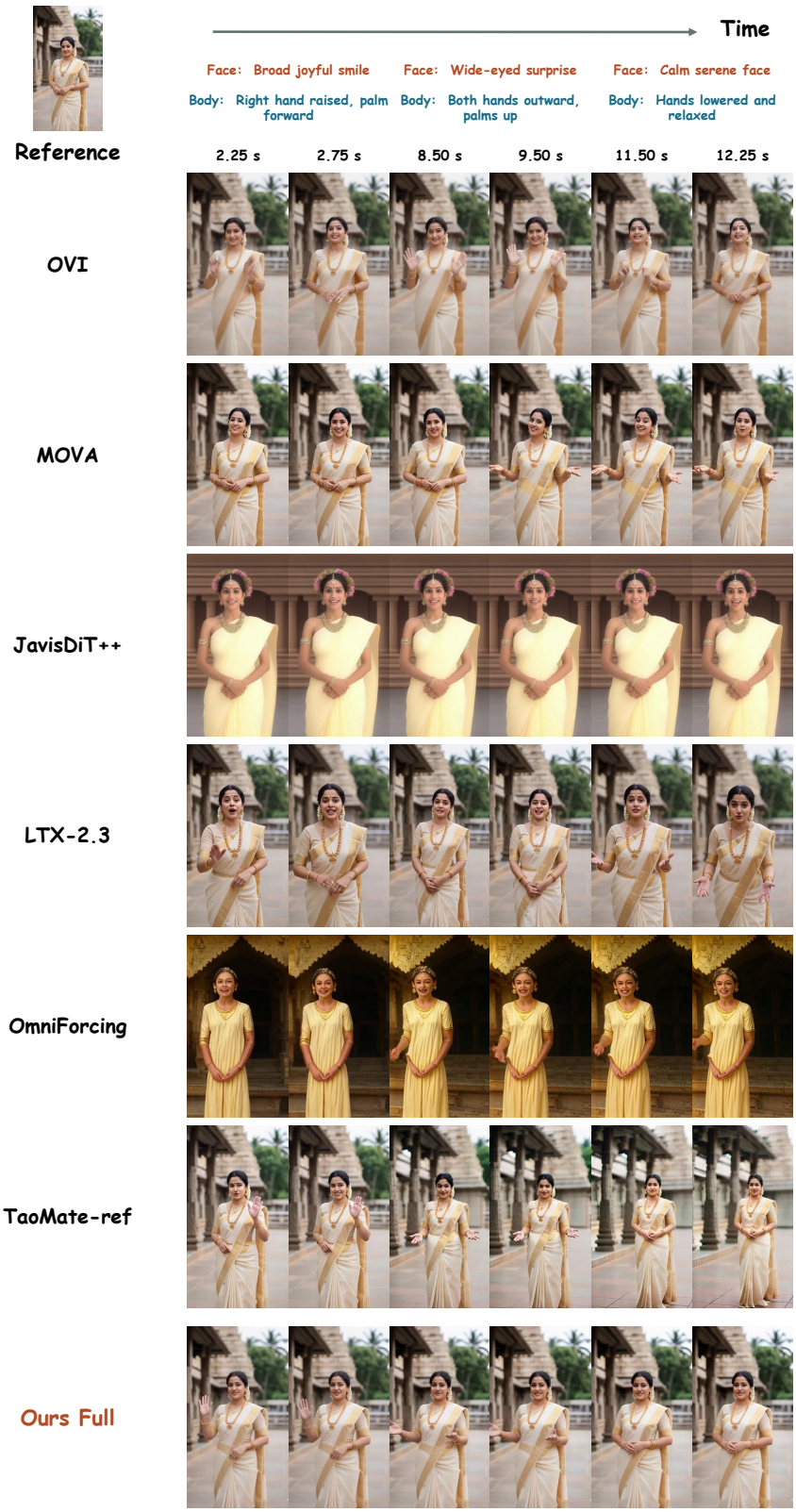

Figure 5: Additional qualitative comparisons with baseline methods, illustrating character consistency, facial expressions, and body movements

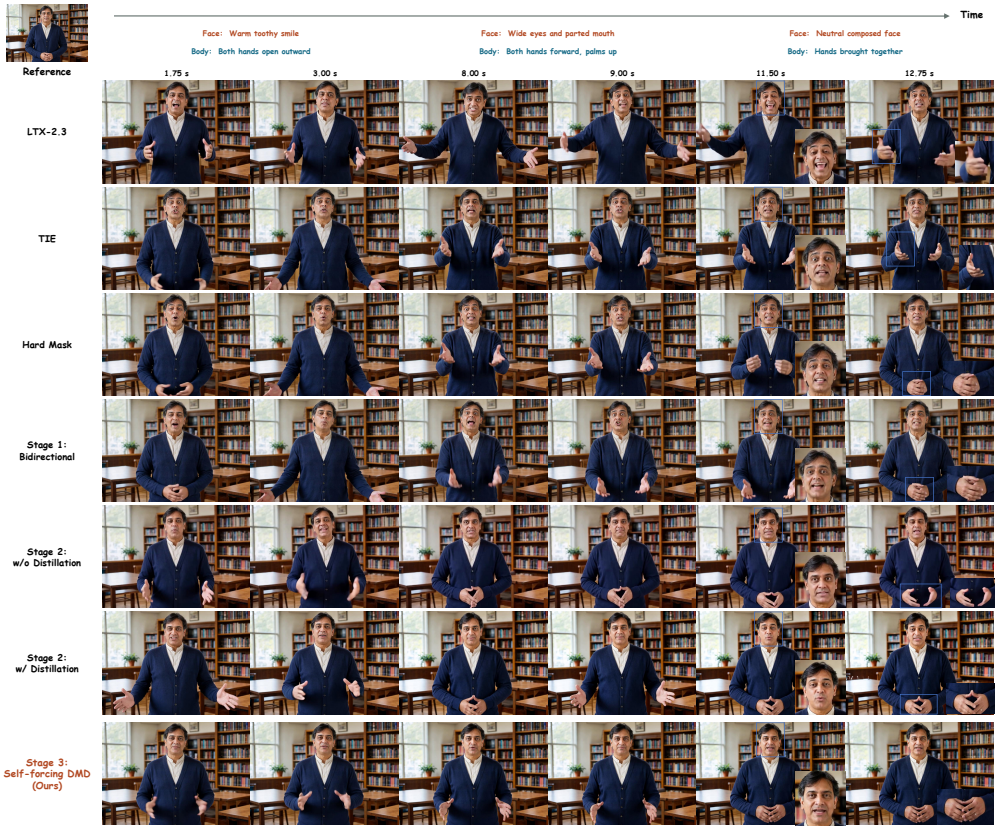

Figure 6: Qualitative ablation across training stages. The final Stage 3 model produces clearer facial features and more detailed hand structures.

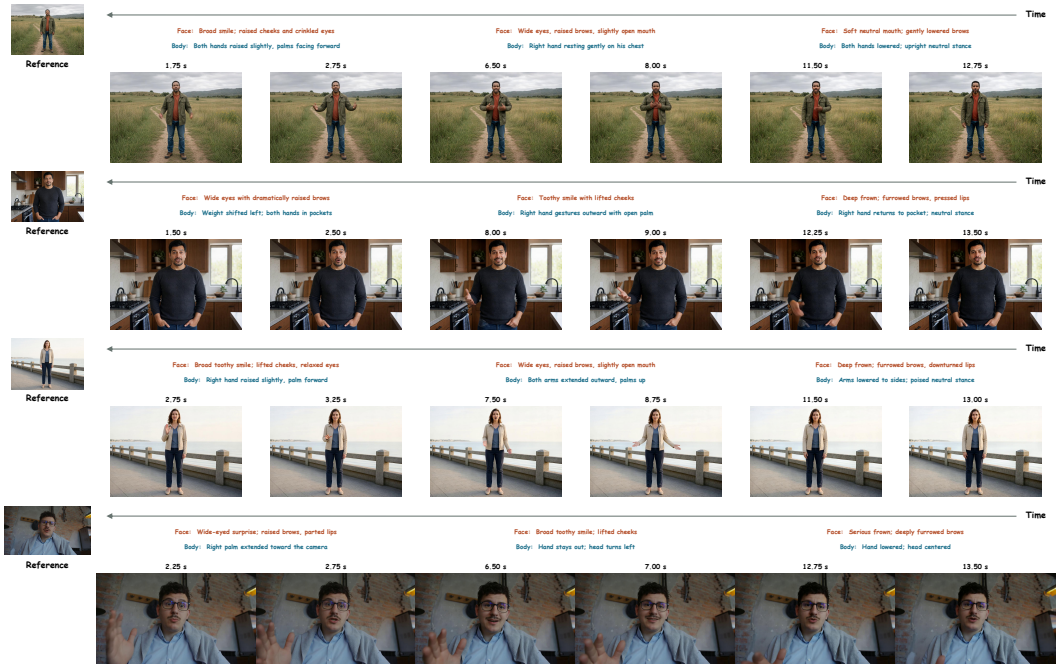

Figure 7: Additional qualitative results generated by TempoTrack across different characters and scenes.

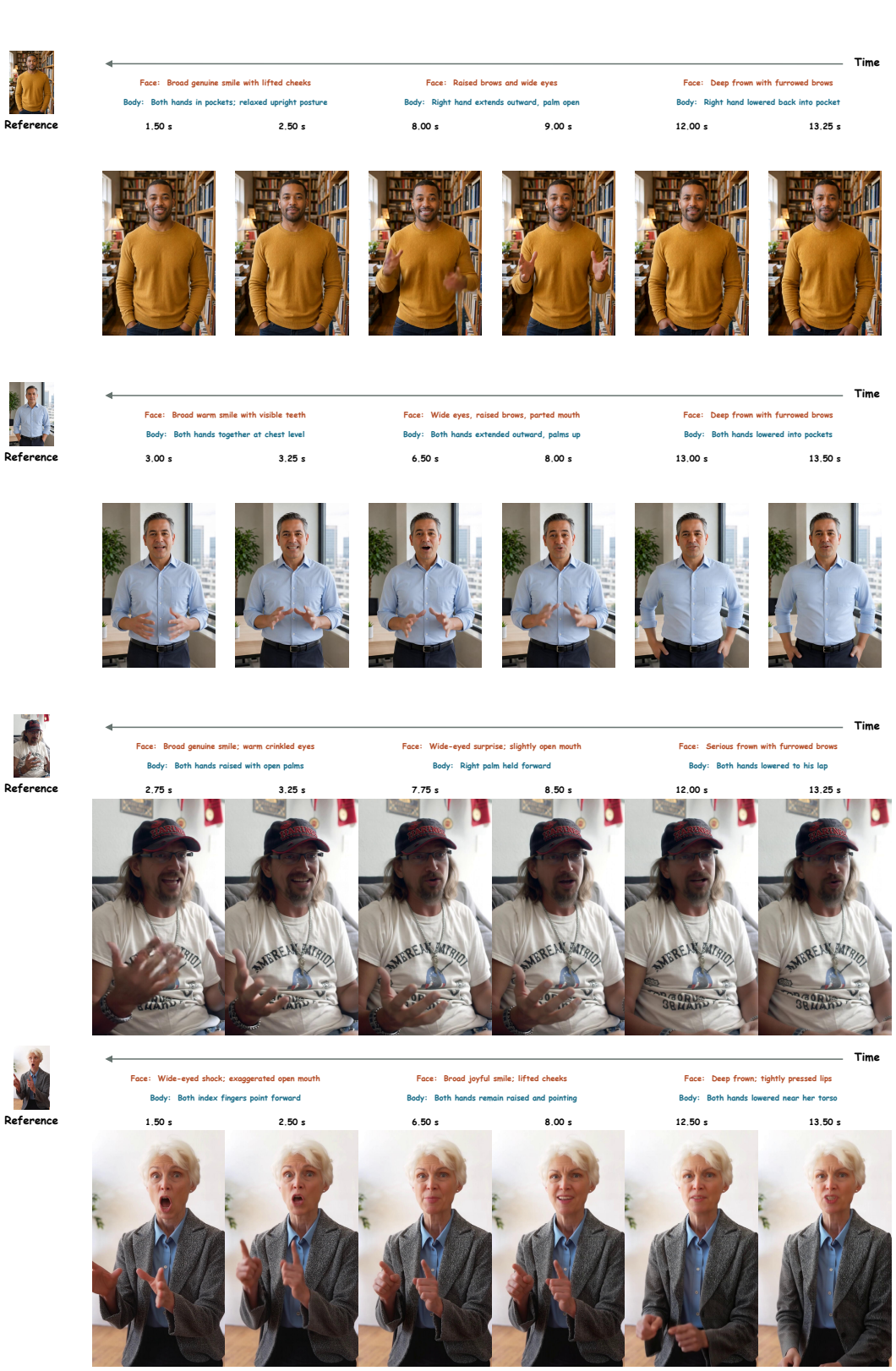

Figure 8: Further examples of TempoTrack's generated audiovisual performances.