TAOMATE: ANCHOR-GUIDED MEMORY BRIDGING EVOLVING AND REFERENCE STATES FOR REAL-TIME AUDIO-VIDEO DIGITAL HUMAN GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-time long-form digital-human generation relies on causal models to extend audio-visual content while preserving subject appearance and audio-video synchronization across successive segments. A bounded cache retains local motion and phonetic context but discards older evidence, whereas attending to the complete generated history is computationally expensive and can propagate accumulated errors. We present TaoMate, an anchor-guided persistent-memory framework for few-step joint audio-video generation. The framework preserves an immutable visual anchor, compresses completed video and audio blocks into fixed-capacity dynamic states, and retrieves those states through modality-specific residual attention without extending the active cache. A reference-aware modulation method additionally conditions video features on dynamic and anchor appearance statistics. Anchor-preserving causal-context distillation varies rollout horizon, prefix provenance, and cache-history reliability while keeping the immutable visual anchor unperturbed. By separating persistent memory from stage-local denoising dependencies, TaoMate further admits stage-parallel execution across blocks, accelerating autoregressive inference without pipeline-specific retraining. We evaluate long-form video continuations with appearance, temporal, synchronization, facial, and speech diagnostics. Results show that TaoMate preserves stable appearance across prompt-conditioned segments and strong audio-visual synchronization under autoregressive generation.

1 INTRODUCTION

Joint audio-video diffusion models have substantially advanced digital-human generation by modeling speech, facial motion, and body dynamics within a unified generative process (HaCohen et al., 2026; Liu et al., 2026; Low et al., 2025; SII-OpenMOSS Team, 2026). Recent causal distillation methods further convert short-clip diffusion priors into few-step autoregressive generators (Huang et al., 2025a; Zhu et al., 2026; Huang et al., 2025b; Zhou et al., 2026), making continuous generation increasingly practical. Long-form autoregressive generation, however, introduces a different requirement: an ordered sequence of prompts must drive evolving actions and speech while preserving the subject, scene, and acoustic behavior established earlier in the video.

Despite this progress, two interrelated challenges remain. **i) Long-term audio-visual consistency.** A bounded key-value (KV) cache retains the recent motion and phonetic context needed for local continuity, but earlier subject, scene, and acoustic evidence eventually leaves the cache. Retaining the complete generated history instead causes attention and storage to grow with duration and repeatedly conditions the model on old, imperfect predictions. Under recursive continuation, small errors in facial structure, color, scene layout, or articulation can therefore propagate across prompt transitions. Moreover, causal distillation alone does not eliminate this accumulation because teacher trajectories provide clean bidirectional context, whereas deployment must consume the student's own causal outputs. **ii) Efficient autoregressive inference.** Each temporal block must traverse multiple denoising stages in a large joint audio-video model. Conventional block-wise execution completes every stage for one block before advancing the next, leaving causally ready computation across blocks unexploited.

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.

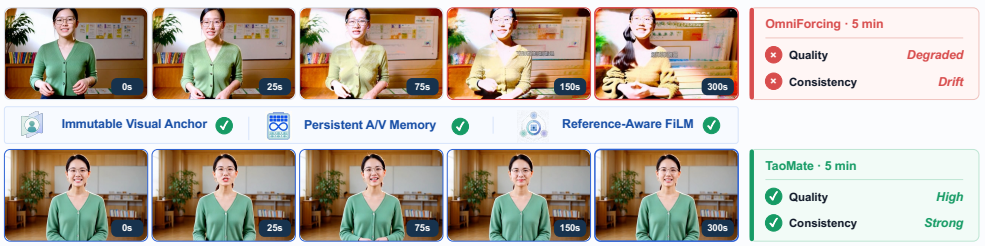


Figure 1: **TaoMate remains stable over 1,000 seconds.** Unlike OmniForcing's visual degradation, identity drift, and color shift, TaoMate preserves consistency through visual anchoring, persistent A/V memory, and reference-aware FiLM.

To address these challenges, we propose TaoMate, an anchor-guided persistent-memory framework for long-form causal audio-video generation. The framework decomposes temporal conditioning into a bounded, position-bearing active context and a fixed-capacity persistent memory. The active context preserves fine-grained local motion and phonetic continuity, whereas persistent memory retains duration-independent visual and acoustic evidence after the corresponding tokens leave the cache. This decomposition separates short-range positional continuity from long-range evidence while keeping the conditioning cost independent of generated duration.

Our framework promotes stable recursive generation through two complementary mechanisms. First, persistent memory separates immutable visual and audio references from adaptive audio-video states. The fixed references preserve the initial subject, scene, and acoustic characteristics, while adaptive states integrate detached representations of completed blocks; anchor agreement further regulates visual updates to suppress unreliable observations. We further separate normalized content representations from channel-level and low-frequency appearance statistics: modality-specific memory attention retrieves structural and acoustic history, while reference-aware FiLM (Perez et al., 2018) reconciles dynamic appearance with the anchor. Second, rollout-aligned causal distillation exposes the student to deployment-like causal contexts. It combines mixed rollout horizons, student-generated prefixes, and selective perturbation of non-anchor history to approximate the context distribution encountered during recursive inference. At inference, the separation between persistent memory and stage-local KV dependencies further enables stage-parallel execution. Successive denoising stages are assigned to different devices, allowing causally ready blocks to advance concurrently after pipeline fill. Only completed clean blocks update persistent memory, preserving the learned causal transition without pipeline-specific retraining. Experiments on long-form Mandarin and English continuations show that TaoMate maintains stable appearance and strong audio-video synchronization across prompt transitions. On three GPUs, the stage-parallel implementation reaches 35 output FPS, enabling real-time long-form generation.

Our main contributions are as follows:

- We introduce anchor-guided persistent multimodal memory that couples an immutable visual anchor with adaptive audio-video states beyond the active KV cache.
- We formulate rollout-aligned causal distillation using mixed rollout horizons and perturbed student histories to reduce exposure bias in recursive generation.
- We enable training-free stage-parallel inference across causally ready blocks, achieving real-time generation.

2 Related Work

Video and audio-video generation. Modern video systems combine diffusion, latent modeling, guidance, and transformer backbones (Ho et al., 2020; Song et al., 2021; Ho & Salimans, 2022; Rombach et al., 2022; Peebles & Xie, 2023; Blattmann et al., 2023b;a; HaCohen et al., 2024). Joint generators model audio and video with asymmetric streams, twin backbones, or experts (HaCohen et al., 2026; Liu et al., 2026; Low et al., 2025; SII-OpenMOSS Team, 2026). Portrait systems instead emphasize lip synchronization, motion, and appearance under specialized controls (Prajwal et al., 2020; Zhou et al., 2020; Wang et al., 2021; Zhang et al., 2023a; Ye et al., 2023; Tian et al., 2024;

Xu et al., 2024; Guo et al., 2024). These methods establish short-horizon generative priors, while long-form causal generation motivates mechanisms for retaining information.

Streaming generation and distillation. In streaming generation, exposure bias arises because training can condition on teacher-derived history whereas inference must consume model-generated history. Self-Forcing and Causal Forcing train causal students on self-generated contexts (Huang et al., 2025a; Zhu et al., 2026). Mutual Forcing instead couples few-step and multi-step modes in a weight-shared autoregressive model, using the few-step mode to construct history for self-distillation (Zhou et al., 2026). Live Avatar combines causal distillation with long-horizon streaming strategies (Huang et al., 2025b). AvatarForcing performs one-step sliding-window denoising with a RoPE-reindexed style anchor, recent clean temporal anchors, and two-stage distribution-matching distillation (Cui et al., 2026). Few-step generation has also been studied through progressive, consistency, latent-consistency, and distribution-matching distillation (Salimans & Ho, 2022; Song et al., 2023; Luo et al., 2023; Yin et al., 2024b;a). PCM decomposes a probability-flow trajectory into adjacent-stage consistency problems (Wang et al., 2024). Several of these methods primarily alter how a causal trajectory or its supervision is constructed. Our training scheme couples self-generated history construction with cache-external memory and keeps the immutable visual anchor unperturbed when perturbing later prefixes.

Position-bearing history retention. Long-context models employ recurrence, compression, retrieval, sinks, heavy-hitter retention, and blockwise attention (Dai et al., 2019; Rae et al., 2020; Wu et al., 2022; Xiao et al., 2024; Zhang et al., 2023b; Liu et al., 2023). For video, TetherCache recalls historical K/V and aligns selected entries to trusted cache statistics, while Pyramid and Sparse Forcing organize retained position-bearing context (Meng et al., 2026; Chen et al., 2026a; Xu et al., 2026). OmniMem sparsely retrieves query-relevant blocks from the full historical K/V cache, whereas FadeMem consolidates older K/V blocks into a distance-aware temporal hierarchy (Zhao et al., 2026; Lu et al., 2026). Anchor Forcing instead stores anchor K/V for prompt-switch re-caching and uses region-specific RoPE origins to control positional drift (Yang et al., 2026). MaineCoon combines agentic prompt planning with inference-time cache management, while AvatarForcing reindexes a style anchor and retains recently emitted clean blocks (Bai et al., 2026; Cui et al., 2026). These approaches preserve, retrieve, consolidate, or reposition elements of position-bearing history.

3 METHOD

3.1 PROBLEM FORMULATION

We consider autoregressive audio-video generation from a temporally ordered sequence of text prompts. The output unfolds causally as a variable-length sequence of synchronized video-audio latent blocks, where each prompt conditions a contiguous temporal segment. Let $\mathbf{z}_t = (\mathbf{z}_t^v, \mathbf{z}_t^a)$ denote the t-th block and let e_t denote its associated text condition. A block is the atomic unit of generation, while consecutive blocks sharing the same condition form a prompt-level output segment. This setting imposes two distinct requirements on temporal conditioning. Fine-grained motion, pose, and phonetic continuity depend on recent latent tokens with explicit temporal positions, whereas subject, scene, and acoustic evidence can remain relevant after those tokens leave a finite attention cache. Retaining the complete generated history would make attention and storage grow with duration and repeatedly expose the generator to old prediction errors; retaining only recent tokens would instead discard long-range evidence. We therefore decompose history into a bounded active context $\mathcal{C}_{t-1}$ and a fixed-capacity persistent memory $\mathcal{M}_{t-1}$. The former preserves position-bearing recent tokens for local continuity, while the latter compresses detached historical evidence for duration-independent long-range conditioning. The resulting causal factorization is

$$\begin{aligned} \mathbf{z}_t &\sim p_\theta(\mathbf{z}_t \mid e_t, \mathcal{C}_{t-1}, \mathcal{M}_{t-1}), \\ \mathcal{M}_t &= U(\mathcal{M}_{t-1}, \mathrm{sg}(\mathbf{z}_t)). \end{aligned} \tag{1}$$

Here sg denotes stop-gradient, which avoids backpropagation through an unbounded chain of memory updates. The update is applied only after $\mathbf{z}_t$ has been generated, so the condition for block t contains information only from preceding blocks. Because both $\mathcal{C}_{t-1}$ and $\mathcal{M}_{t-1}$ have bounded capacity, their storage and conditioning cost do not grow with the duration already generated.

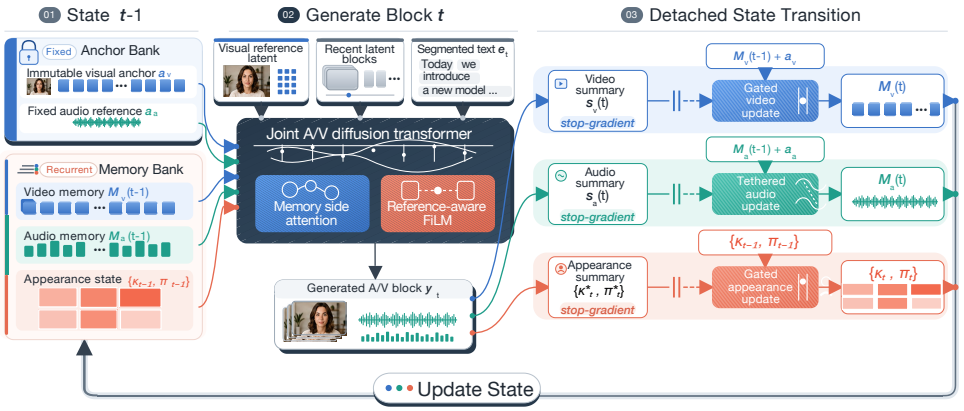


Figure 2: **TaoMate Pipeline.** The pipeline generates synchronized long-form video and audio causally with anchor-guided persistent memory. A bounded context provides local motion and phonetic cues, while a fixed visual anchor and adaptive modality states supply long-range conditioning through side attention and reference-aware FiLM (Perez et al., 2018).

3.2 Anchor-Guided Persistent Memory

Long-horizon generation requires memory to preserve information with different temporal semantics. Subject appearance and acoustic characteristics should remain stable under recursive generation, whereas motion, scene evolution, and phonetic context must adapt to newly generated content. Encoding these roles in a single recurrent representation couples retention and adaptation: rapid updates can overwrite identity-defining evidence, while conservative updates limit responsiveness to evolving context. We therefore factor persistent memory into fixed references and adaptive modality-specific states:

$$\mathcal{M}_t = (\mathbf{a}^v, \mathbf{a}^a, \mathbf{m}_t^v, \mathbf{m}_t^a, \boldsymbol{\kappa}_t, \boldsymbol{\pi}_t). \tag{2}$$

The visual anchor $\mathbf{a}^v$ is fixed from the initial video block, while the audio reference $\mathbf{a}^a$ is initialized from the first available audio observation. The content memories $\mathbf{m}_t^v$ and $\mathbf{m}_t^a$ aggregate evolving spatial and acoustic structure, while $\boldsymbol{\kappa}_t$ and $\boldsymbol{\pi}_t$ preserve absolute channel and low-frequency appearance statistics that are separated from normalized visual content. This factorization allows each information type to follow an update rule and conditioning pathway matched to its temporal role.

All block-derived memory states are detached and have fixed cardinality, which prevents an unbounded optimization graph and keeps memory cost independent of generated duration. The compressed visual anchor is distinct from the uncompressed, position-bearing initial visual latent retained in the active context $\mathcal{C}_t$: the former provides compact long-range reference evidence, whereas the latter remains an explicit token-level condition for causal attention. Figure 2 summarizes the state evolution.

Factorized block representations. The fixed references are instantiated once from the initial block, whereas the adaptive memory components must incorporate evidence from each completed block. To update them without conflating relative content structure with drift-prone absolute appearance, we decompose each block into normalized modality content representations and unnormalized visual appearance representations:

$$\begin{aligned} \mathbf{s}_t^v &= P_v[\mathrm{Norm}_{HW}(\bar{\mathbf{z}}_t^v)], \quad \bar{\mathbf{z}}_t^v = F_v^{-1}\textstyle\sum_f \mathbf{z}_{t,f}^v, \\ \mathbf{s}_t^a &= P_a[\mathrm{Norm}_{F_a}(\mathbf{z}_t^a)], \quad \boldsymbol{\kappa}_t^* = K(\mathbf{z}_t^v), \quad \boldsymbol{\pi}_t^* = P_c(\bar{\mathbf{z}}_t^v). \end{aligned} \tag{3}$$

Here $\mathbf{s}_t^v$ and $\mathbf{s}_t^a$ denote the fixed-cardinality video and audio content representations extracted from block t; they serve as observations for updating $\mathbf{m}_t^v$ and $\mathbf{m}_t^a$, rather than as persistent states themselves. F_v and F_a are the temporal lengths of the video and audio latents, respectively, and f indexes video latent frames. Norm_{HW} standardizes each video channel over its spatial dimensions, whereas Norm_{F_a} standardizes each audio channel over the temporal dimension. P_v, P_a, and P_c are fixed-cardinality projections, and K extracts channel-wise mean and scale. The superscript $*$ denotes a current-block candidate before it is incorporated into persistent memory. The normalized content representations retain relative spatial and acoustic structure, whereas the appearance observations $\boldsymbol{\kappa}_t^*$ and $\boldsymbol{\pi}_t^*$ preserve color, contrast, and coarse spatial statistics for long-range calibration.

Anchor-constrained memory evolution. Block-wise representations extracted from generated history are imperfect observations. Incorporating them at a fixed rate allows transient prediction errors to accumulate in future conditions, whereas freezing the memory prevents legitimate motion and speech evolution. Memory evolution is formulated as an anchor-regularized online estimate:

$$\mathcal{E}(\mathbf{m}, \mathbf{a}, \mathbf{s}; \beta, \lambda) = (1-\lambda)[(1-\beta)\mathbf{m} + \beta\mathbf{s}] + \lambda\mathbf{a}. \tag{4}$$

The coefficients of the previous estimate, current observation, and fixed reference are $(1-\lambda)(1-\beta)$, $(1-\lambda)\beta$, and λ, respectively. They are nonnegative and sum to one, making $\mathcal{E}$ their weighted Euclidean barycenter. Thus β controls evidence assimilation, while λ imposes a persistent reference prior. For constant rates, the coefficient of an observation from k updates earlier is $(1-\lambda)\beta[(1-\lambda)(1-\beta)]^k$; historical evidence and isolated errors therefore decay geometrically, while the anchor is reintroduced at every update.

The reliability of a video observation depends on whether it preserves the structural evidence established by the reference. Because $\mathbf{s}_t^v$ is spatially normalized, its agreement with the visual anchor is insensitive to first-order channel shifts. We convert this agreement into a smooth observation-confidence gate:

$$\begin{aligned} \rho_t &= B^{-1} \sum_b \cos(\mathrm{vec}(\mathbf{s}_{t,b}^v), \mathrm{vec}(\mathbf{a}_b^v)), \\ g_t &= g_{\min} + (1-g_{\min})\sigma((\rho_t - \tau)/T_g), \\ \mathbf{m}_t^v &= \mathcal{E}(\mathbf{m}_{t-1}^v, \mathbf{a}^v, \mathbf{s}_t^v; \beta_v g_t, \lambda_v), \\ \mathbf{m}_t^a &= \mathcal{E}(\mathbf{m}_{t-1}^a, \mathbf{a}^a, \mathbf{s}_t^a; \beta_a, \lambda_a). \end{aligned} \tag{5}$$

Here B denotes the mini-batch size and b indexes samples within the mini-batch. vec flattens the token and channel dimensions, $\sigma(\cdot)$ is the sigmoid function, τ is the agreement threshold, and T_g is the gate temperature. Low agreement reduces the effective video update rate, limiting the influence of structurally inconsistent generations. The nonzero gate floor retains an adaptation path as pose or scene content changes. Audio follows the anchor-tethered update without visual-agreement gating, allowing rapidly changing phonetic content to enter memory while retaining its acoustic reference.

Channel-level errors can propagate as visible color or contrast drift. We therefore apply the same soft confidence gate to $\boldsymbol{\kappa}_t$ and $\boldsymbol{\pi}_t$, together with a hard rejection rule based on their normalized deviation from anchor statistics. The gate scales gradual appearance updates, while rejection prevents an outlying observation from shifting the reference appearance. The resulting memory retains long-range evidence at fixed capacity.

3.3 Memory Conditioning

Persistent memory influences generation without being converted back into additional position-bearing history. Directly concatenating all memory components with the active sequence would blur the distinction between temporally ordered local context and position-free long-range evidence. It would also force token-valued content and global appearance statistics through the same retrieval mechanism, despite their different roles in generation. We therefore use structure-matched conditioning: content memory supports query-dependent retrieval, whereas appearance statistics act coherently across video tokens through feature modulation. For modality $m \in \{v, a\}$, let $\mathbf{r}_{t-1}^v = [\mathbf{a}^v; \mathbf{m}_{t-1}^v]$, $\mathbf{r}_{t-1}^a = \mathbf{m}_{t-1}^a$, and $\mathbf{u}_{t-1}^m = E_m(\mathbf{r}_{t-1}^m)$. Selected transformer layers then apply

$$\begin{aligned} \widetilde{h}_\ell^m &= h_\ell^m + \mathrm{MemAttn}_\ell^m(\mathrm{RMSNorm}(h_\ell^m), \mathbf{u}_{t-1}^m), \\ \widehat{h}_\ell^v &= \widetilde{h}_\ell^v + \boldsymbol{\gamma}_\ell(\mathbf{c}_{t-1}) \odot \mathrm{RMSNorm}(\widetilde{h}_\ell^v) + \mathbf{b}_\ell(\mathbf{c}_{t-1}), \end{aligned} \tag{6}$$

where $\mathbf{c}_{t-1}$ concatenates dynamic and anchor channel statistics and the second line is a reference-aware FiLM residual (Perez et al., 2018). The low-frequency appearance state $\boldsymbol{\pi}_{t-1}$ follows a complementary parameter-free readout that calibrates each completed video latent before it is committed as future context and used to update memory. Memory K/V remain outside the active cache, so fixed token counts keep the additional cost independent of generated duration. Both pathways are residual and initialized as identity mappings; architectural details are deferred to the supplementary material.

3.4 Causal-Context Distillation and Generation

Autoregressive distillation introduces a context-distribution gap: teacher trajectories provide clean bidirectional context, whereas the deployed student must continue from its own imperfect causal

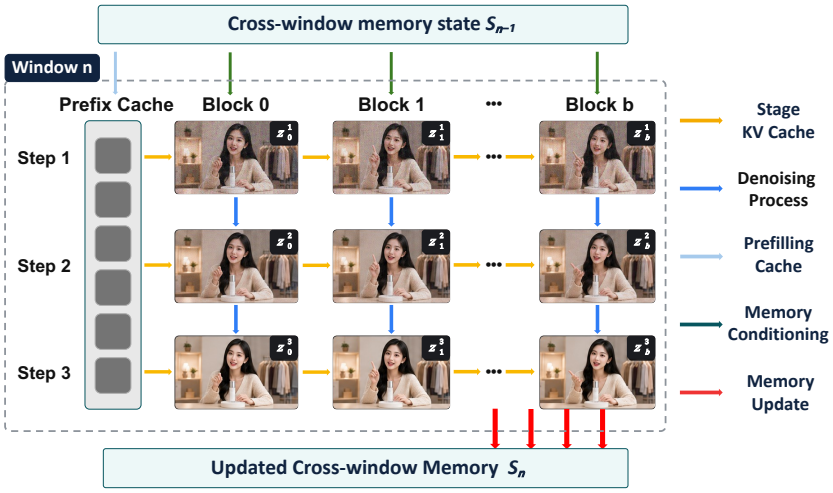


Figure 3: Memory-aware stage-parallel inference. Each stage pre-fills its KV cache from the bounded clean prefix $\mathcal{P}_{n-1}$ and reads the persistent memory $\mathcal{M}_{n-1}$. Blocks advance along an anti-diagonal pipeline, and only final clean outputs update cross-window conditioning.

history. We reduce this gap by distilling on causal self-rollouts that span multiple temporal horizons and draw prefixes from both clean trajectories and detached student generations. Before intermediate non-anchor blocks are reused as causal context, their latent representations are perturbed along the forward flow-matching path, while the persistent visual anchor remains clean. This asymmetric treatment discourages the student from treating recent self-generated context as perfectly reliable and preserves the anchor as a stable source of appearance evidence. Together, these variations expose the student to accumulated generation errors without weakening the persistent identity signal.

On these causal contexts, distribution-matching distillation transfers the guided teacher distribution over the sampled rollout (Yin et al., 2024b;a). We complement this global objective with video PCM (Wang et al., 2024), which promotes consistency between neighboring denoising transitions in latent and appearance-sensitive feature spaces. Persistent-memory updates are detached from the optimization graph, allowing extended causal rollouts without retaining gradients through the complete history.

3.5 Memory-Aware Stage-Parallel Inference

Generating a long video in a single pass would cause the attention context and its position-bearing KV cache to grow continuously. We therefore divide the video into a sequence of bounded windows and propagate the cross-window conditioning pair $\mathcal{S}_{n-1} = (\mathcal{P}_{n-1}, \mathcal{M}_{n-1})$. The clean prefix $\mathcal{P}_{n-1}$ provides explicit latents for short-range continuity, whereas the fixed-capacity persistent memory $\mathcal{M}_{n-1}$ provides learned representations for long-range identity and appearance consistency. In our configuration, the prefix consists of the initial reference sink and up to recent blocks.

Within each window, every pipeline rank is assigned one denoising stage and maintains an independent video–audio KV cache. Before denoising starts, each stage initializes its cache by pre-filling the same explicit clean latents from $\mathcal{P}_{n-1}$, while the persistent memory $\mathcal{M}_{n-1}$ conditions both the prefix prefill and all subsequent denoising forwards. This cross-window prefix is prefilled only once per window. At the window boundary, only final-stage clean blocks construct $\mathcal{P}_n$ and update $\mathcal{M}_n$ for window $n+1$. Each stage discards its previous window cache and pre-fills a new cache from initial reference sink and the most recent clean blocks using positions local to the new window.

During window generation, a block completed at stage k is committed only to the cache of stage k, as indicated by the yellow arrows in Fig. 3. Consequently, when stage k processes block b, its cache contains the bounded clean prefix and all preceding blocks $z_0^k, \ldots, z_{b-1}^k$ committed at the same denoising stage. Cleaner outputs from downstream stages are never written back into an upstream cache. This stage-local cache organization bounds the position-bearing context while preserving causal access to all earlier same-stage blocks. A task (b, k) depends on the same block at the preceding stage, $(b, k-1)$, and on the cache state produced by $(b-1, k)$, which already contains all

Table 1: Comparison with representative speech-to-video (S2V) methods and other baselines on the long-form benchmark. Best and second-best results are shown in bold and underlined, respectively.

Method	Params	LipSync↑	Desync↓	Align.↑	Express.↑	Long Cons.↑	Color Δ ↓	DiT FPS↑
OmniAvatar	1.3B	4.794	–	–	–	0.6467	0.01566	0.948
StableAvatar	1.3B	3.246	–	–	–	0.9242	0.00895	0.352
Wan2.2-S2V-14B	14B	4.898	–	–	–	0.8413	0.00814	0.079
LiveAvatar	14B	<u>5.396</u>	–	–	–	<u>0.9619</u>	<u>0.00541</u>	4.529
Hallo3	5B	4.855	–	–	–	0.7947	0.01383	0.150
LongLive-2.0	5B	–	–	–	–	0.7221	0.01792	<u>10.47</u>
LTX-2.3	22B	2.349	0.050	4.35	3.60	0.3745	0.04387	0.262
JavisDiT++	2.1B	0.904	1.025	**5.00**	3.90	0.3545	0.07159	0.437
MOVA	32B/18B	3.860	0.060	**5.00**	3.75	0.7129	0.02825	0.128
OVI	11B	3.493	0.135	<u>4.60</u>	3.70	0.8252	0.01797	0.231
OmniForcing	22B	1.590	<u>0.045</u>	**5.00**	<u>3.95</u>	0.3404	0.02060	8.14
TaoMate	22.1B	**5.918**	**0.000**	**5.00**	**4.00**	**0.9755**	**0.00260**	**16.32**

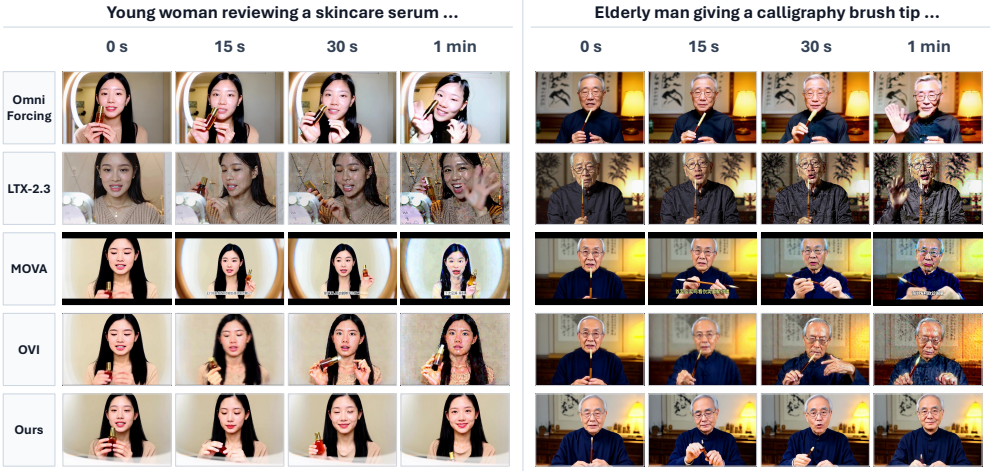


Figure 4: **Qualitative comparison on one-minute video generation.** For two representative prompts, we show frames sampled at 0, 15, 30, and 60 seconds (columns) from videos generated by different methods (rows).

earlier same-stage blocks. Tasks satisfying $b+k=d$ therefore form an anti-diagonal wavefront. After pipeline generation, different blocks are processed simultaneously at different denoising stages, and the final stage continuously retires clean blocks. This substantially reduces inference time over serial block-wise denoising, enabling real-time autoregressive generation without pipeline-specific retraining.

4 Experiments

4.1 Implementation Details

Long-form benchmark. We evaluate single-presenter scenarios in both Mandarin and English. Each scenario is represented by a temporally ordered prompt sequence in which every prompt specifies the action, transcript, subject appearance, scene, camera, and voice for one short segment. The subject, scene, camera, and voice descriptions remain consistent across the sequence, while the action and spoken content evolve. Every internal output contains 60 seconds of video at 768×512 and 24 fps.

Training. The causal generator is initialized from a distilled LTX-2.3 checkpoint and optimized against a frozen bidirectional teacher using three denoising steps per block. The teacher uses classifier-free guidance scales of 3.0 for video and 5.0 for audio. The generator and critic models use learning rates of 10^{-5} and 2×10^{-6}, respectively, and the generator EMA decay is 0.99. Residual memory attention is inserted every four transformer layers, giving 12 memory-enabled

layers in the 48-layer backbone. The video branch receives 48 tokens formed by the visual anchor and dynamic video state, while the audio branch receives 64 dynamic-state tokens; together with the modality encoders and video FiLM, these branches add approximately 97.8M trainable parameters.

Training samples rollout horizons of five and ten blocks with probabilities 0.8 and 0.2 and substitutes detached student-generated prefixes with probability 0.1. Intermediate generated blocks reused as cache history are re-noised with $\sigma \sim \mathcal{U}(0, 0.15)$, while the persistent visual anchor is initialized from the unperturbed reference latent. The adaptive formulation uses $(\beta_v, \lambda_v) = (0.15, 0.20)$ and $(\beta_a, \lambda_a) = (0.10, 0.10)$ for video and audio state updates, with gate parameters $(\tau, T_g, g_{\min}) = (0.05, 0.10, 0.10)$. Appearance states use $(\beta_c, \lambda_c) = (0.03, 0.60)$. The PCM regularizer has weight 0.03, begins at optimization step 400, and is evaluated every two optimization steps.

We train on 6,139 filtered synthetic trajectories generated by the bidirectional LTX-2.3 teacher (HaCohen et al., 2026). Each trajectory contains paired video and audio ODE states, with up to 241 video frames at 768×512 and 24 fps. The audio latents are aligned to the duration of the corresponding video trajectory.

Inference. The EMA student generator produces each one-minute sample by following its ordered prompt sequence autoregressively. At a prompt transition, the active context retains the initial visual reference and at most two recent non-anchor blocks. Only completed blocks update the persistent memory, and local RoPE coordinates restart for each prompt-conditioned segment. The end-to-end single-device implementation requires 130.3 ± 1.3 seconds per 1,441-frame output on an NVIDIA RTX PRO 5000 72GB GPU, corresponding to 11.1 generated frames per second.

4.2 Comparison with State-of-the-Art Methods

Evaluation metrics. We adopt LipSync, Desync, Alignment, and Expressiveness from the official VABench evaluation protocol (Hua et al., 2026). For audio-conditioned video-only methods, LipSync is evaluated against the conditioning audio.

To capture long-form behavior beyond these audio-visual criteria, we use two complementary measures over fixed temporal intervals. *Color* Δ measures local appearance discontinuity between neighboring intervals and is therefore sensitive to visible boundary-level color shifts. *Long Consistency* summarizes appearance stability relative to the opening interval through color drift, brightness variation, and local color discontinuity. Lower Color Δ and higher Long Consistency indicate better temporal stability; their exact definitions are provided in the supplementary material. For the ablation study, *Face Consistency* compares normalized MediaPipe facial geometry between early and late intervals (Kartynnik et al., 2019); it measures face-shape stability rather than biometric identity. *DiT FPS* is the number of generated frames per denoising second and excludes data loading, VAE decoding, and post-processing. All methods are profiled on the same NVIDIA RTX PRO 5000 72GB GPU.

Baselines and evaluation setup. We compare with the joint audio-video generators LTX-2.3, JavisDiT++, OVI, and MOVA (HaCohen et al., 2026; Liu et al., 2026; Low et al., 2025; SII-OpenMOSS Team, 2026), together with an OmniForcing-aligned causal baseline. We additionally include the audio-driven video generators OmniAvatar, StableAvatar, Wan2.2-S2V, LiveAvatar, and Hallo3, and the LongLive-2.0 text-to-video plus image-to-video pipeline (Gan et al., 2025; Tu et al., 2025; Gao et al., 2025; Huang et al., 2025b; Cui et al., 2025; Chen et al., 2026b). LTX-2.3 follows unified chunked continuation, while JavisDiT++ and OVI generate clips assembled according to the benchmark prompts. All methods are evaluated with the same scripts and 768×512 output resolution on the same GPU; their long-form construction follows the respective model implementations.

Quantitative results. Table 1 reports quantitative comparisons with representative S2V systems and joint audio-video baselines. TaoMate ranks first in LipSync, Desync, and Expressiveness and ties for the best Alignment, demonstrating accurate speech-motion coupling with expressive audio-video dynamics. It also achieves the highest Long Consistency and lowest Color Δ, indicating that repeated continuation preserves global appearance while suppressing local boundary shifts. Relative to the strongest external result on each measure, these scores improve LipSync by 9.7%, raise Long Consistency by 0.0136, and reduce Color Δ by 51.9%. At 16.32 DiT FPS, TaoMate is $1.56\times$ faster than the next-highest measured throughput, showing that these stability gains remain compatible with efficient denoising.

Table 2: Multi-GPU end-to-end throughput. Size denotes denoising-generator parameters.

Method	Size	GPUs	FPS ↑
LiveAvatar	14B	5	23.0
LongLive-2.0	5B	4	16.0
TaoMate	22.1B	3	35.0

Table 3: Persistent-memory ablation. Higher is better except Color Δ; bold and underlined values indicate best and second-best results.

Mem.	FiLM	Gate	Long Cons.↑	Color Δ ↓	Face Cons.↑
–	–	–	0.9614	0.00380	0.9955
✓	–	–	<u>0.9653</u>	<u>0.00316</u>	<u>0.9956</u>
✓	✓	–	0.9620	0.00347	0.9947
✓	✓	✓	**0.9676**	**0.00276**	**0.9962**

Qualitative results. Figure 4 presents visual results from one-minute generations. Across both sequences, TaoMate retains the subject's facial outline, clothing appearance, and scene layout while the prompted gestures evolve. At 60 seconds, LTX-2.3 shows visible texture degradation in the face and background; MOVA and OVI exhibit larger changes in facial appearance and scene composition. These frame sequences illustrate the long-horizon appearance differences summarized by Color Δ and Long Consistency in Table 1.

4.3 Inference Throughput

Table 2 evaluates whether the compared systems sustain real-time generation under their reported multi-GPU deployments. TaoMate reaches 35.0 output FPS on three GPUs, exceeding its 24-fps playback rate by 45.8% and the reported throughput of LiveAvatar and LongLive-2.0 by $1.52\times$ and $2.19\times$, respectively. At this throughput, the complete 1,441-frame generation pipeline finishes in approximately 41.2 seconds. By restricting persistent-memory updates to fully denoised outputs, our stage-parallel schedule allows causally ready blocks to progress concurrently through different denoising stages, achieving real-time throughput.

4.4 Ablation Study

All variants use the same data, rollout construction, optimization settings, and serial inference protocol; only the memory pathways listed in Table 3 differ. Relative to the memory-free baseline, fixed-capacity memory raises Long Consistency by 0.0039 and lowers Color Δ by 16.8%, while Face Consistency changes by only 0.0001. This profile indicates that cross-window evidence mainly stabilizes global appearance and scene continuity, with facial geometry already near saturation. Adding ungated FiLM, however, erodes these gains: Long and Face Consistency decrease by 0.0033 and 0.0009, respectively, while Color Δ increases by 9.8%. Gating recurrent updates by anchor agreement instead raises the two consistency scores by 0.0056 and 0.0015 and reduces Color Δ by 20.5%, producing the best result on every metric. This contrast shows that appearance modulation is effective when unreliable recurrent observations are suppressed before conditioning future blocks.

5 Conclusion

We introduced TaoMate, a framework for real-time long-form audio-video generation that separates a bounded, position-bearing active context from fixed-capacity persistent memory. The memory combines immutable references with adaptive content and appearance states for stable long-range conditioning. Rollout-aligned causal distillation further uses mixed rollout horizons, student-generated prefixes, and selective perturbation of non-anchor history to reduce the context gap between teacher-guided training and recursive inference. Experiments and ablations demonstrate strong audio-video synchronization, improved cross-window appearance stability, and the importance of anchor-regulated memory updates. This separation further enables stage-parallel inference, reaching 35 output FPS on three GPUs.

AI USE STATEMENT

Generative AI tools assisted with language editing, figure design, and LaTeX preparation. The authors reviewed the resulting manuscript and take responsibility for its scientific content, citations, and results.

REFERENCES

Lichen Bai, Tianhao Zhang, Shitong Shao, Dingwei Tan, Qiyu Zhong, et al. Mainecoon: Pursuing a real-time audio-visual social world model. *arXiv preprint arXiv:2606.17800*, 2026. URL https://arxiv.org/abs/2606.17800.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023a. URL https://arxiv.org/abs/2311.15127.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22563–22575, 2023b. URL https://openaccess.thecvf.com/content/CVPR2023/html/Blattmann_Align_Your_Latents_High-Resolution_Video_Synthesis_With_Latent_Diffusion_Models_CVPR_2023_paper.html.

Jiayu Chen, Junbei Tang, Wenbiao Zhao, Maoliang Li, Jiayi Luo, Zihao Zheng, et al. Pyramid forcing: Head-aware pyramid kv cache policy for high-quality long video generation. *arXiv preprint arXiv:2605.13111*, 2026a. URL https://arxiv.org/abs/2605.13111.

Yukang Chen, Luozhou Wang, Wei Huang, Shuai Yang, Bohan Zhang, Yicheng Xiao, Ruihang Chu, Weian Mao, Qixin Hu, Shaoteng Liu, Yuyang Zhao, Huizi Mao, Ying-Cong Chen, Enze Xie, Xiaojuan Qi, and Song Han. Longlive-2.0: An nvfp4 parallel infrastructure for long video generation. *arXiv preprint arXiv:2605.18739*, 2026b. URL https://arxiv.org/abs/2605.18739.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21086–21095, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Cui_Hallo3_Highly_Dynamic_and_Realistic_Portrait_Image_Animation_with_Video_CVPR_2025_paper.html.

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. *arXiv preprint arXiv:2603.14331*, 2026. URL https://arxiv.org/abs/2603.14331.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2978–2988, 2019. doi: 10.18653/v1/P19-1285. URL https://aclanthology.org/P19-1285/.

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

Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Dechao Meng, Jinwei Qi, Penchong Qiao, Zhen Shen, Yafei Song, et al. Wan-s2v: Audio-driven cinematic video generation. *arXiv preprint arXiv:2508.18621*, 2025. URL https://arxiv.org/abs/2508.18621.

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

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024. URL https://arxiv.org/abs/2501.00103.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026. URL https://arxiv.org/abs/2601.03233.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. URL https://arxiv.org/abs/2207.12598.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pp. 6840–6851, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Daili Hua, Xizhi Wang, Bohan Zeng, Xinyi Huang, Hao Liang, Junbo Niu, Xinlong Chen, Quanqing Xu, and Wentao Zhang. VABench: A comprehensive benchmark for audio-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23345–23355, 2026.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems*, volume 38, 2025a. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/f4823f831af67a3ef15e41a85434422a-Abstract-Conference.html.

Yubo Huang, Hailong Guo, Fangtai Wu, Weiqiang Wang, Shifeng Zhang, et al. Live avatar: Streaming real-time audio-driven avatar generation with infinite length. *arXiv preprint arXiv:2512.04677*, 2025b. URL https://arxiv.org/abs/2512.04677.

Yury Kartynnik, Artsiom Ablavatski, Ivan Grishchenko, and Matthias Grundmann. Real-time facial surface geometry from monocular video on mobile gpus. *arXiv preprint arXiv:1907.06724*, 2019. URL https://arxiv.org/abs/1907.06724.

Hao Liu, Matei Zaharia, and Pieter Abbeel. Ring attention with blockwise transformers for near-infinite context. *arXiv preprint arXiv:2310.01889*, 2023. URL https://arxiv.org/abs/2310.01889.

Kai Liu, Yanhao Zheng, Kai Wang, Shengqiong Wu, Rongjunchen Zhang, Jiebo Luo, Dimitrios Hatzinakos, Ziwei Liu, Hao Fei, and Tat-Seng Chua. Javisdit++: Unified modeling and optimization for joint audio-video generation. *arXiv preprint arXiv:2602.19163*, 2026. URL https://arxiv.org/abs/2602.19163.

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

Yu Lu, Junjie Yang, Piotr Koniusz, Yuxin Song, and Yi Yang. Fademem: Distance-aware memory consolidation for autoregressive video diffusion. *arXiv preprint arXiv:2606.10671*, 2026. URL https://arxiv.org/abs/2606.10671.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023. URL https://arxiv.org/abs/2310.04378.

Yu Meng, Xiangyang Luo, Letian Li, Wenyuan Jiang, Chen Gao, Xinlei Chen, Yong Li, and Xiao-Ping Zhang. Tethercache: Stabilizing autoregressive long-form video generation with gated recall and trusted alignment. *arXiv preprint arXiv:2606.13035*, 2026. URL https://arxiv.org/abs/2606.13035.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Peebles_Scalable_Diffusion_Models_with_Transformers_ICCV_2023_paper.html.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. URL https://ojs.aaai.org/index.php/AAAI/article/view/11671.

K. R. Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C. V. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 484–492, 2020. doi: 10.1145/3394171.3413532. URL https://doi.org/10.1145/3394171.3413532.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. Compressive transformers for long-range sequence modelling. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1911.05507.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. URL https://iclr.cc/virtual/2022/spotlight/6538.

SII-OpenMOSS Team. Mova: Towards scalable and synchronized video-audio generation. *arXiv preprint arXiv:2602.08794*, 2026. URL https://arxiv.org/abs/2602.08794.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32211–32252. PMLR, 2023. URL https://proceedings.mlr.press/v202/song23a.html.

Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive – generating expressive portrait videos with audio2video diffusion model under weak conditions. In *European Conference on Computer Vision*, pp. 244–260, 2024. doi: 10.1007/978-3-031-73010-8_15. URL https://link.springer.com/chapter/10.1007/978-3-031-73010-8_15.

Shuyuan Tu, Yueming Pan, Yinming Huang, Xintong Han, Zhen Xing, Qi Dai, Chong Luo, Zuxuan Wu, and Yu-Gang Jiang. Stableavatar: Infinite-length audio-driven avatar video generation. *arXiv preprint arXiv:2508.08248*, 2025. URL https://arxiv.org/abs/2508.08248.

Fu-Yun Wang, Zhaoyang Huang, Alexander William Bergman, Dazhong Shen, Peng Gao, Michael Lingelbach, Keqiang Sun, Weikang Bian, Guanglu Song, Yu Liu, Xiaogang Wang, and Hongsheng Li. Phased consistency models. In *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/98a29475083c502c34949f9baa1aa2ef-Abstract-Conference.html.

Suzhen Wang, Lincheng Li, Yu Ding, Chen Fan, and Xin Yu. Audio2head: Audio-driven one-shot talking-head generation with natural head motion. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 1098–1105, 2021. doi: 10.24963/ijcai.2021/152. URL https://www.ijcai.org/proceedings/2021/152.

Yuhuai Wu, Markus N. Rabe, DeLesley Hutchins, and Christian Szegedy. Memorizing transformers. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=TrjbxzRcnf-.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5e5fd18f863cbe6d8ae392a93fd271c9-Abstract-Conference.html.

Boxun Xu, Yuming Du, Zichang Liu, Siyu Yang, Ziyang Jiang, et al. Sparse forcing: Native trainable sparse attention for real-time autoregressive diffusion video generation. *arXiv preprint arXiv:2604.21221*, 2026. URL https://arxiv.org/abs/2604.21221.

Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Yao Yao, and Siyu Zhu. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv preprint arXiv:2406.08801*, 2024. URL https://arxiv.org/abs/2406.08801.

Yang Yang, Tianyi Zhang, Wei Huang, Jinwei Chen, Boxi Wu, Xiaofei He, Deng Cai, Bo Li, and Peng-Tao Jiang. Anchor forcing: Anchor memory and tri-region RoPE for interactive streaming video diffusion. *arXiv preprint arXiv:2603.13405*, 2026.

Zhenhui Ye, Ziyue Jiang, Yi Ren, Jinglin Liu, Jinzheng He, and Zhou Zhao. Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=YfwMIDhPccD.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. doi: 10.52202/079017-1505. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/54dcf25318f9de5a7a01f0a4125c541e-Abstract-Conference.html.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6613–6623, 2024b. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yin_One-step_Diffusion_with_Distribution_Matching_Distillation_CVPR_2024_paper.html.

Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8652–8661, 2023a. doi: 10.1109/CVPR52729.2023.00836. URL https://openaccess.thecvf.com/content/CVPR2023/html/Zhang_SadTalker_Learning_Realistic_3D_Motion_Coefficients_for_Stylized_Audio-Driven_Single_CVPR_2023_paper.html.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Re, Clark Barrett, Zhangyang Wang, and Beidi Chen. H2o: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023b. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6ceefa7b15572587b78ecfcebb2827f8-Abstract-Conference.html.

Lin Zhao, Yushu Wu, Yifan Gong, Yanzhi Wang, and Pu Zhao. Omnimem: Scalable and adaptive memory retrieval for long video generation. *arXiv preprint arXiv:2605.30519*, 2026. URL https://arxiv.org/abs/2605.30519.

Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makeittalk: Speaker-aware talking-head animation. *ACM Transactions on Graphics*, 39(6), 2020. doi: 10.1145/3414685.3417774. URL https://doi.org/10.1145/3414685.3417774.

Yupeng Zhou, Lianghua Huang, Zhifan Wu, Jiabao Wang, Yupeng Shi, et al. Mutual forcing: Dual-mode self-evolution for fast autoregressive audio-video character generation. *arXiv preprint arXiv:2604.25819*, 2026. URL https://arxiv.org/abs/2604.25819.

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