ANAMORA: REFERENCE-GROUNDED STREAMING AUDIO-VIDEO GENERATION WITH ENTITY-STATE MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Streaming joint audio-video generation is an emerging paradigm for continuous audiovisual synthesis, opening new possibilities for immersive, real-time interaction. Stable subject identities and coherent interactions are essential to this process, yet current streaming approaches struggle to maintain consistent subject and scene states when shots change or multiple subjects interact. Stable subject integration and sustained multi-subject interaction are therefore central to coherent, personalized streaming experiences. We present Anamora, a unified framework for reference-grounded multi-shot streaming audio-video generation with entity-state memory. Anamora keeps shot-specific visual references accessible to maintain subject identities while allowing reference updates during generation. A state-aware entity memory selects history compatible with the current requirements and shares a fixed capacity across subjects, supporting state continuity across scene transitions and reappearances. Extensive experiments demonstrate that Anamora achieves high-fidelity subject injection and cross-shot consistency, highlighting the importance of reference grounding and entity-state memory in streaming interactive audio-video generation.

1 INTRODUCTION

Joint audio-video models synthesize coherent motion, speech, and sound (Low et al., 2025; Cheng et al., 2026; HaCohen et al., 2026). Streaming models (Su et al., 2026; Li et al., 2026; Song et al., 2026a) extend this synthesis to continuous, interactive output. Personalized storytelling and immersive experiences require users to introduce subjects and guide their actions across shots, demanding faithful subject integration and preservation of their evolving states.

Subject control remains challenging: text incompletely specifies appearance, while first-frame conditioning ties identity to one scene configuration. StreamChar (Tian et al., 2026) and OmniMate (Song et al., 2026a) introduce visual references, but continuity further requires identity cues that survive shot changes and generated states that persist beyond recent context. Otherwise, continuous output may still exhibit subject drift or forget earlier interactions.

These limitations give rise to two closely related challenges. (1) *Stable subject injection*: a user-specified subject must be faithfully introduced and consistently re-established in each shot despite changes in viewpoint, scene composition, and participating entities. Without persistent, shot-specific grounding, reliance on generated context can progressively weaken agreement with the supplied reference. (2) *Long-horizon subject-state consistency*: preserving identity is insufficient when interactions change a subject's appearance or physical state. If a glass breaks during generation, it should retain that particular damage after leaving view and reappearing, rather than revert to its intact reference. The first challenge concerns grounding each shot in the intended subjects; the second concerns preserving what has happened to those subjects.

We present Anamora, a unified framework for reference-grounded streaming audio-video generation with entity-state memory (Figure 1). Active subject references remain accessible throughout each shot, grounding joint motion and audio synthesis beyond the preceding frames. References can be updated as interaction proceeds, without fixing later shots to the initial composition.

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.

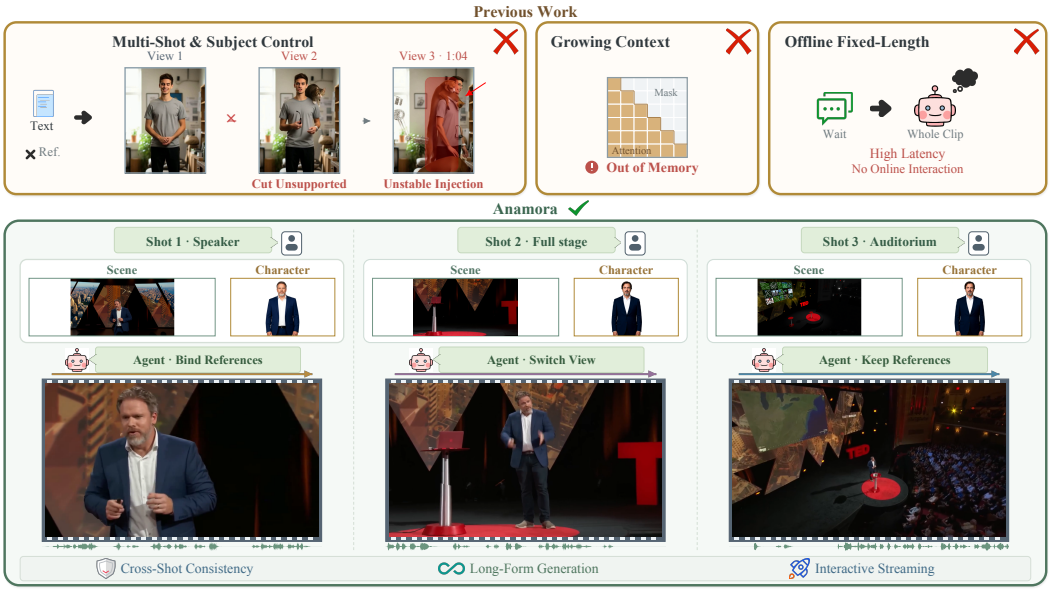

Figure 1: **Reference-grounded audiovisual streaming. Top:** Three limitations motivate our design: unstable subject control across shots, growing memory use during long streams, and high latency with limited interaction in fixed-length generation. **Bottom:** Scene and character references accompany a speaker across a medium shot, a full-stage view, and an auditorium view. All three shots come from one continuous Anamora video.

References anchor identity but do not record subsequently created states. Building on entity-centered history modeling (Dou et al., 2026; Vandersanden et al., 2026), *State-Aware Entity Memory* selects completed observations compatible with current state requirements, excluding obsolete appearances. It shares a fixed capacity across entities, reserving feasible shares for small or temporarily absent subjects. Thus reference identity and generated experience remain distinct without retaining the entire stream.

An external planning agent maintains a global table of reference bindings, desired states, and shot participation. It translates user instructions into condition updates while leaving unmentioned entries unchanged. The table specifies requested changes; visual memory separately records accepted observations.

We evaluate Anamora in terms of reference fidelity, cross-shot consistency, audiovisual quality, and streaming efficiency. Component ablations examine reference features stored in memory, generated-history retention, and state-aware entity selection under a fixed memory budget.

Our main contributions are summarized as follows:

- We present Anamora, a reference-grounded streaming framework for joint audio-video generation that injects subject references into every shot, supporting high-fidelity subject integration and reference updates during continuous synthesis.
- We propose State-Aware Entity Memory to sustain long-horizon subject-state continuity. State-compatible history selection and fixed-budget entity allocation retain relevant generated evidence across shot changes and entity reappearances.
- We integrate an interaction agent that interprets user inputs and maintains a global entity table, coordinating reference bindings, state requirements, and shot participation for coherent multi-entity audiovisual interaction.

2 Related Work

2.1 Streaming Audio-Video Generation

Progress in diffusion-based sequence modeling (Chen et al., 2024; Yin et al., 2025; Huang et al., 2025a) has enabled incremental video synthesis, paving the way for audio-driven streaming avatars (Huang et al., 2025b; Song et al., 2026b). Moving beyond supplied driving audio, joint audiovisual models (Low et al., 2025; Cheng et al., 2026; HaCohen et al., 2026) generate sound and video together, with OmniForcing (Su et al., 2026) and Hallo-Live (Li et al., 2026) extending joint synthesis to streaming output. StreamChar (Tian et al., 2026) and OmniMate (Song et al., 2026a) further combine reference conditioning with high-level generation control to support persistent characters. However, consistent reference appearance does not by itself preserve states created during interaction, particularly when subjects disappear and return in later shots. Anamora combines shot-specific reference grounding with persistent entity history, allowing later shots to draw on both supplied identity cues and the visual consequences of earlier interactions.

2.2 Memory for Long-Horizon Generation

Sliding-window methods (Yang et al., 2025; Chen et al., 2026) use persistent anchors to stabilize long rollouts, but provide limited access to events beyond the recent context. Historical retrieval (Li et al., 2025; Zhang et al., 2025; Ding et al., 2026b) extends access to more distant history, while Ripple (Ding et al., 2026a) compresses past audiovisual context into recurrent memory. Entity-centered approaches (Dou et al., 2026; Liu et al., 2026) further organize history around recurring subjects, and EM-Vid (Vandersanden et al., 2026) supports sparse entity memory with selective appearance editing. Beyond retrieving the right entity, however, incremental interaction requires deciding which of its past observations remain compatible with the current state requirements. Anamora selects state-compatible visual evidence and allocates a fixed total memory budget across entities.

3 Method

Anamora generates synchronized audiovisual chunks from prompts, visual references, desired states, and participation conditions that can change across or within shots (Figure 2). Built on the shared text–reference–audio–video transformer of MiniMax-H3 Ref2AV (MiniMax, 2026), Reference-Grounded Multi-Shot Streaming maintains active identity cues, while State-Aware Entity Memory retains history-specific evidence under fixed capacity. An external agent updates conditions through an entity table, outside the generator and memory policy.

3.1 Preliminaries

Flow Matching. Flow matching (Lipman et al., 2023) learns a transformation from noise to data. Let Z_t^m be a clean latent chunk, where t indexes chunks and $m \in \{v, a\}$ denotes video or audio, and let $\epsilon_t^m \sim \mathcal{N}(0, I)$ be the corresponding noise. Noise time τ is distinct from t. With a modality-specific schedule σ_m, we form $Z_{t,\tau}^m = (1 - \sigma_m(\tau))Z_t^m + \sigma_m(\tau)\epsilon_t^m$. Following the backbone's prediction convention, v_θ^m estimates the clean-ward target $u_t^m = Z_t^m - \epsilon_t^m$, giving the objective

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}\left[\sum_{m\in\{v,a\}} \lambda_m \,\mathrm{MSE}_m(v_\theta^m, u_t^m)\right], \tag{1}$$

where λ_m balances the modalities and each error averages over valid target coordinates. The corresponding clean estimate is $\widehat{Z}_t^m = Z_{t,\tau}^m + \sigma_m(\tau)v_\theta^m$.

Self-Forced Distillation. Self Forcing (Huang et al., 2025a) trains with self-generated rather than ground-truth history to reduce the autoregressive train–test mismatch. Distribution Matching Distillation (DMD) (Yin et al., 2024) uses a frozen teacher and a learned generated-sample score model for few-step training. We apply these foundations after adapting reference and memory reading.

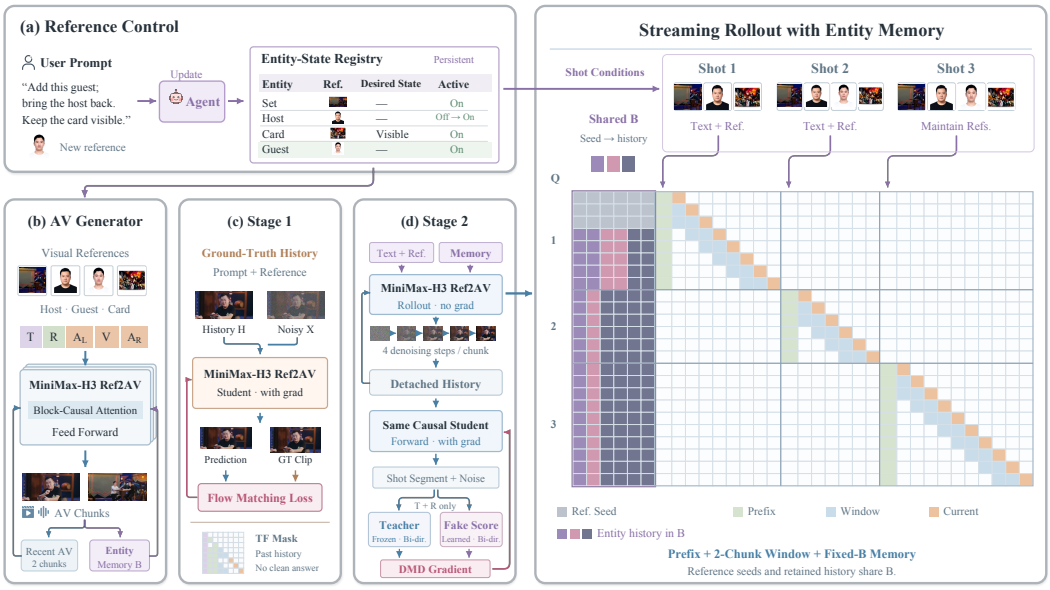

Figure 2: **Overview of Anamora.** **(a)** An agent maintains an entity-state table and organizes prompts and visual references to guide generation. **(b)** The generator produces audio and video chunks from these conditions, recent context, and selected history. **(c)** The first stage uses teacher forcing to train a causal student. **(d)** The second stage distills few-step generation using the student's own history. **Right:** Each shot uses its own prompt and references, while a fixed-capacity memory carries relevant entity history across shots.

3.2 Reference-Grounded Multi-Shot Streaming

Rather than condition only the first frame, active references form a prefix visible to every chunk. Causal attention combines these identity cues with recent audiovisual context and selected older visual evidence.

Chunkwise Joint Generation. Conditions $\mathbf{c}_t$ comprise prompt p_t, active references $\mathcal{R}_t$, state requirements $\mathcal{S}_t$, and shot-start flag b_t; participation determines active entities and references. C_{t-1} holds the two preceding completed chunks of the current shot, and $\widetilde{M}_t^{\text{read}}$ is prepared before chunk t (Section 3.3). Given audiovisual noise ϵ_t:

$$(\widehat{Z}_t^v, \widehat{Z}_t^a) = G_\theta(\epsilon_t; \mathbf{c}_t, C_{t-1}, \widetilde{M}_t^{\text{read}}), \qquad \mathbf{c}_t = (p_t, \mathcal{R}_t, \mathcal{S}_t, b_t), \tag{2}$$

Conditions and pre-block memory remain fixed throughout denoising. Each shot has a text/reference prefix $P_s = [T_s \mid R_s]$; joint synthesis requires no external driving audio.

Causal Audiovisual Attention. Let T encode the prompt and state requirements and R_e contain entity e's reference tokens, with N groups. Each block B_t aligns video tokens V_t and left/right audio tokens A_t^L, A_t^R to the same physical interval:

$$[\,T \mid R_1 \mid \cdots \mid R_N \mid B_1 \mid B_2 \mid \cdots], \qquad B_t = [\,A_t^L \mid V_t \mid A_t^R\,], \tag{3}$$

Positional information and modality tags follow the reordering, allowing within-block audiovisual interaction and between-block causality. To prevent indirect future leakage through the prefix, let H_t, X_t denote history-source and noisy-target copies under fixed conditions, and $R = \bigcup_e R_e$. Visible key sets are

$$\mathcal{N}(T) = T, \qquad \mathcal{N}(R_e) = T \cup R_e, \tag{4}$$

$$\mathcal{N}(H_t) = T \cup R \cup C_{t-1} \cup \widetilde{M}_t^{\text{read}} \cup H_t,$$

$$\mathcal{N}(X_t) = T \cup R \cup C_{t-1} \cup \widetilde{M}_t^{\text{read}} \cup X_t. \tag{5}$$

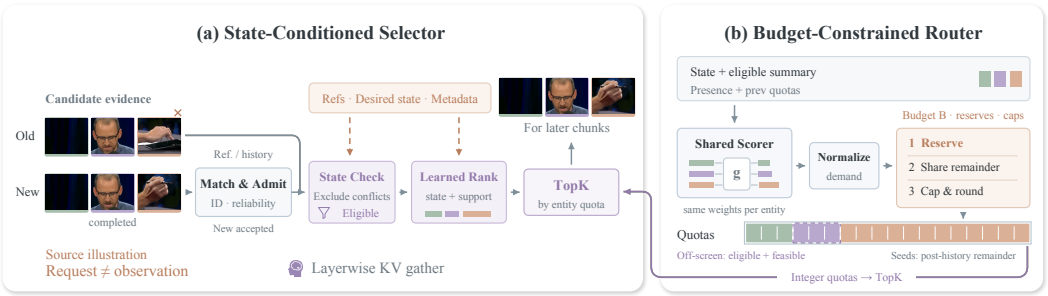

Figure 3: **Entity memory selection and allocation.** **(a)** Accepted observations join each entity's history. Evidence that conflicts with current requirements is excluded, and the remaining candidates are ranked for later generation. **(b)** A fixed budget B is shared according to each entity's estimated needs, with minimum shares for small or temporarily absent entities when capacity permits.

Text reads itself; each reference group reads text and itself, never history or future prompts. Targets cannot read their clean copies or future outputs; KV beyond the recent window enter only through $\widetilde{M}_t^{\text{read}}$. Visibility respects the current shot and condition version.

History-source chunks are encoded at the same small noise level σ_c in teacher forcing, self-forced rollout, and inference. Once a completed chunk is committed, its perturbation is reused. A separate history-encoding forward supplies the KV for subsequent chunks and is outside the four sampling steps. Deep reference KV are reused only when their text, position, and modulation dependencies are unchanged.

Shot Transitions and Reference Updates. At a cut, the generator rebinds the shot prefix and clears prior-shot local KV, retaining valid entity history without an extra bootstrap segment or protected opening chunk. Reference updates change addressed bindings rather than necessarily changing entities; text edits can change desired attributes without replacing references. Updates apply at the next unstarted chunk, remain fixed during denoising, and need not trigger a cut. Affected condition features are refreshed; retained history keeps its original source and condition version, without re-encoding under the new request.

Interaction Interface. An external agent updates reference bindings, desired states, and shot participation in a global entity table, supplying the conditions in Equation 2 while preserving unmentioned entries. This conditioning interface is separate from the generator and memory policy; requests are not observations accepted into memory.

3.3 State-Aware Entity Memory

Entity memory retains history-specific states beyond reference identity, separating state-conditioned selection (*what to retain*) from budget-constrained routing (*how much to retain*; Figure 3). References and desired states guide both; only completed, accepted observations supply new history.

State-Conditioned Entity Selection. We associate individual spatiotemporal positions with entities using reference appearance, descriptions, retained history, and within-chunk correspondence in a shared matching space. State cues accommodate appearance changes; uncertain matches are rejected, while identifiable small entities remain selectable.

A requested state counts as achieved only when supported by a completed observation. Reliability and state checks admit new evidence $\mathcal{C}_{e,t}^{+}$ only for an accepted initial establishment or reliable revision; otherwise it is empty, and idle or clearer frames do not overwrite valid history. Eligibility tests new evidence and $M_{e,t-1}$ against $\mathbf{c}_{e,t}$ (reference binding, desired state, edit scope, and locked attributes). Given features d_j and provenance/observation metadata μ_j, the learned scorer f_ϑ ranks

eligible candidates, and hard TopK retains $k_{e,t}$ under the router's quota:

$$\mathcal{U}_{e,t} = \mathrm{Eligible}\big(M_{e,t-1} \cup \mathcal{C}^{+}_{e,t}; \mathbf{c}_{e,t}\big)\,, \tag{6}$$

$$u_{e,j} = f_\vartheta(d_j, \mathbf{c}_{e,t}, \mu_j), \qquad M_{e,t} = \mathrm{TopK}_{k_{e,t}}(\mathcal{U}_{e,t}; u_{e,\cdot}), \tag{7}$$

Here $\mathcal{U}_{e,t}$ denotes eligible evidence and $u_{e,j}$ its score. Ranking weighs unchanged-attribute support and relevance to accepted changes. Accepted revisions replace scope-conflicting sources, favoring the latest reliable state. Separable unchanged evidence may remain; inseparable conflicting sources are excluded. Intact sources may support transitions but cannot evidence damage, and source-level filtering does not imply perfect semantic disentanglement. Without an accepted revision, valid history persists subject to eligibility and quota reduction. $M_{e,t}$ is available only from block $t+1$, never during the denoising that produced it.

Budget-Constrained Entity Routing. To prevent large or frequent entities from dominating, a shared demand network predicts normalized demand from state conditions, eligible-evidence summaries, shot participation, and previous quotas. Allocation covers all retained entities $\mathcal{E}^{\mathrm{keep}}_t$, including off-screen ones. For sink capacity B, reserve $k_{\min}$, and cap $k_{\max}$, available evidence bounds the quotas:

$$\overline{k}_{e,t} = \min(k_{\max}, |\mathcal{U}_{e,t}|), \qquad \underline{k}_{e,t} = \min(k_{\min}, \overline{k}_{e,t}), \tag{8}$$

$$\underline{k}_{e,t} \le k_{e,t} \le \overline{k}_{e,t}, \qquad \sum_{e \in \mathcal{E}^{\mathrm{keep}}_t} k_{e,t} \le B. \tag{9}$$

The allocator assigns feasible reserves, then shares residual capacity by relative demand through capped integer allocation. Infeasible reserves trigger capacity failure, not budget expansion or silent removal of inactive entities. Routing controls retention capacity, not observation acceptance: revisions need not change quotas, and valid off-screen evidence retains feasible reserves.

Memory Readout. Before chunk t, a fixed read view combines completed history M_{t-1} and reference-tagged seeds F^{ref}_t. Only valid history for active entities $\mathcal{E}^{\mathrm{act}}_t \subseteq \mathcal{E}^{\mathrm{keep}}_t$ enters M^{read}_{t-1}; all retained history and seeds share B:

$$\widetilde{M}^{\mathrm{read}}_t = F^{\mathrm{ref}}_t \cup M^{\mathrm{read}}_{t-1}, \qquad |F^{\mathrm{ref}}_t| + |M_{t-1}| \le B. \tag{10}$$

Initially $M_0 = \varnothing$. Off-screen history counts against B despite being unread; valid seeds occupy only residual capacity and remain reference conditions, never observed-state evidence. Shared position indices gather KV at every backbone layer. Audiovisual queries read them alongside the prefix, two-block recent window, and current block; text/reference queries retain Equation 4. Prefix/seed and recent/memory duplicates, identified by source, condition version, and coordinates, are read once. References and history remain distinct, without averaging or relabelling. After completion, admitted observations and routed quotas update M_t for the next chunk. B bounds sink storage, not total GPU memory.

3.4 Training Strategy

Two generator-training stages bracket memory-policy fitting: adapt the reader on real sequences, fit selection/routing with the generator fixed, then distill on self-generated history with the policy frozen.

Teacher-Forced Reader Adaptation. We adapt the causal generator on real multi-shot sequences using Equation 1. Past entity/state and observation labels construct feasible banks independently of learned policies; valid reference seeds fill residual capacity. Later real continuations supervise cross-shot reading but never supply past candidate features.

Memory-Policy Fitting. With the generator frozen, labels supervise state-conditioned selection. Routing compares feasible allocations by the reader's flow-matching error on later real targets, holding past evidence, conditions, noise, and reader fixed. Future targets provide supervision, never router inputs. Supervision fits hard TopK and integer-allocation decisions without differentiating through either.

Self-Forced Distribution Matching. With the memory policy frozen, Self Forcing and DMD train the few-step generator. Bank admission, selection, and routing follow inference rules without assuming identical feature distributions. A no-gradient rollout supplies detached history for a gradient-enabled student forward producing an audiovisual segment $\bar{Z}$.

After re-noising $\bar{Z}$ as in Section 3.1, the frozen H3 Ref2AV teacher and learnable fake-score model receive the same segment and only its (T_s, R_s), without external history or sink. Bidirectional attention stays within one shot, one condition version, and the supported teacher duration. Their clean estimates $\bar{Z}_\star$ and $\bar{Z}_\phi$ define the DMD update:

$$\nabla_\theta \mathcal{L}_{\mathrm{DMD}} = \mathbb{E}_{\tau,\epsilon}\left[\sum_{m\in\{v,a\}} \lambda_m w_m \left(\frac{\partial \bar{Z}^m}{\partial \theta}\right)^\top \mathrm{sg}(\bar{Z}^m_\phi - \bar{Z}^m_\star)\right]. \tag{11}$$

Positive w_m includes valid-coordinate normalization. Weights and score estimates are detached: DMD gradients traverse only the student path producing $\bar{Z}$, not stored history or either score model. Alternating fake-model updates use Equation 1, replacing clean data by $\mathrm{sg}(\bar{Z})$ in both interpolation and clean-ward targets. DMD uses no real continuation target; it aims to improve four-step quality and reduce exposure bias, rather than directly supervise history-specific recall.

4 Experiments

4.1 Experimental Setup

Our framework is built upon MiniMax-H3 Ref2AV (MiniMax, 2026). Training and inference use adaptive-aspect-ratio 480p video, except for self-forced DMD at 384p. Training uses 64 NVIDIA H20 GPUs, with 5K steps of teacher-forced reader adaptation, 8K steps of memory-policy fitting, and 5K steps of self-forced DMD, following the order in Section 3.4. Under the H3 temporal packing, each full chunk contains five video latent frames, advancing 17 output frames, and 28 or 29 audio latents per channel on a shared 40-Hz clock. Outputs use 24-fps video and 32-kHz stereo audio. We use four consistency-sampling steps with video/audio timestep shifts of 12/3, BF16 DiT computation, sharded parameters, sequence parallelism, and KV caching. A separate history-encoding forward is outside the four denoising steps. Unless otherwise stated, the recent window holds two completed chunks; reference seeds and generated history share a $B = 512$-token sink with per-entity bounds $k_{\min} = 32$ and $k_{\max} = 256$.

4.2 Training Data

We use OpenHumanVid (Li et al., 2025) and Koala (Wang et al., 2025b) as training sources. Quality and audiovisual synchronization filtering precede shot segmentation and speech alignment. Neighboring shots are grouped into multi-shot sequences while preserving their cut boundaries. Shot-level audiovisual captions associate backgrounds, characters, and objects with persistent entity IDs and separate state IDs. Verified entity crops and complementary views provide visual references. This organization pairs supplied appearance cues with later observations of the same entities, supporting reference grounding and cross-shot memory training.

4.3 Evaluation

For evaluation, we sample a total of 200 single-shot and multi-shot videos, each paired with corresponding subject reference images.

Evaluation Metrics. We use six metric groups to assess single-shot audiovisual quality and speed, as well as multi-shot entity continuity. *Speed* comprises FPS (generated video frames divided by generation time) and TTFF (latency to complete latent block 0, before decoding). Both exclude VAE time and are measured after warm-up to avoid first-use CUDA kernel compilation overhead (FlexAttention/FlashAttention). Anamora uses a larger backbone, and its inference implementation is still being optimized. For *audio*, Audiobox Aesthetics (Tjandra et al., 2025) provides Production Quality (PQ) and Content Usefulness (CU), while Whisper-based WER (Radford et al., 2023) measures spoken-content accuracy. For *video*, the LAION aesthetic score (VQ) (Schuhmann, 2022; Huang

Table 1: **Quantitative comparison of joint audio-video generation.** Single-shot cases assess speed and audiovisual quality; multi-shot cases assess cross-shot and reference consistency. Input conditions differ across models. "–" denotes inapplicable metrics; TTFF is in seconds and WER is a ratio. Best and second-best speed and audiovisual scores are **bold** and underlined, including ties.

	Speed		Audio			Video		Lip-Sync		Cross-Shot		Reference	
Model	FPS ↑	TTFF ↓	PQ ↑	CU ↑	WER ↓	VQ ↑	TV ↑	LSE-C ↑	LSE-D ↓	C-S ↑	C-B ↑	R-S ↑	R-B ↑
Offline joint audio-video generation													
Ovi-1.1 (Low et al., 2025)	–	–	5.62	5.36	0.43	3.91	25.23	4.12	7.49	–	–	–	–
MOVA-360P (SII-OpenMOSS Team et al., 2026)	–	–	6.46	**6.24**	**0.29**	3.86	**25.36**	3.59	8.06	–	–	–	–
UniVerse-1 (Wang et al., 2025a)	–	–	4.85	4.60	0.52	3.62	25.19	2.55	9.14	–	–	–	–
DaVinci-Base (SII-GAIR et al., 2026)	–	–	5.95	5.42	0.47	3.72	25.23	3.17	7.58	–	–	–	–
LTX-2.3 (HaCohen et al., 2026)	–	–	6.39	6.12	0.31	3.66	25.34	3.81	**6.44**	–	–	–	–
Streaming joint audio-video generation													
OmniForcing (Su et al., 2026)	8.05	3.14	6.58	6.14	0.56	3.78	24.30	1.68	14.36	–	–	–	–
Hallo-Live (Li et al., 2026)	**12.99**	**2.16**	5.53	5.50	0.38	3.78	25.32	0.93	13.37	–	–	–	–
Anamora (Ours)	6.22	4.32	**6.86**	6.23	0.32	**3.92**	25.34	**5.843**	7.11	**0.82**	**0.75**	**0.88**	**0.83**

et al., 2024) measures visual appeal, and ViCLIP similarity (TV) (Wang et al., 2024) measures video–text alignment. *Lip-sync* uses SyncNet confidence (LSE-C) and audio–lip embedding distance (LSE-D) (Chung & Zisserman, 2016; Prajwal et al., 2020). Following the subject/background distinction in MultiShotMaster (Wang et al., 2026), *cross-shot consistency* uses DINOv2 (Oquab et al., 2023) similarity between matched subject (C-S) and same-scene background (C-B) regions. *Reference consistency* compares those regions with their effective input references (R-S/R-B). We mask out the background for subject matching and foreground entities for background matching before computing DINOv2 feature similarity. FPS/TTFF do not characterize streaming performance for fixed-length offline models; the evaluated baselines' single-shot interfaces lack independent-reference conditioning, precluding cross-shot/reference evaluation.

4.4 Comparison Results

We compare Anamora with offline joint audiovisual models Ovi-1.1 (Low et al., 2025), MOVA-360P (SII-OpenMOSS Team et al., 2026), UniVerse-1 (Wang et al., 2025a), DaVinci-Base (SII-GAIR et al., 2026), and LTX-2.3 (HaCohen et al., 2026; Lightricks, n.d.), as well as streaming models OmniForcing (Su et al., 2026) and Hallo-Live (Li et al., 2026). Execution mode does not imply matched conditioning: text-only, first-frame, and independent-reference inputs are not interchangeable. In particular, OmniForcing is a text-only streaming control, not a reference-matched competitor.

Table 1 shows that Anamora achieves the highest PQ (6.86), VQ (3.92), and LSE-C (5.843) among the listed methods, with second-best CU (6.23) and LSE-D (7.11), tied second-best TV (25.34), and WER of 0.32. Its cross-shot C-S/C-B scores are 0.82/0.75 and reference R-S/R-B scores are 0.88/0.83. Its generation throughput is lower and first-block latency higher than those of the two streaming baselines.

Qualitative Comparisons. Figure 4 compares two paired examples with scene and character references. Each Anamora sequence contains three camera shots from one continuous video, while the corresponding saved baseline clips remain in a single shot. OmniForcing also introduces an extra speaker in example (a), and the baselines differ from the supplied character and scene references. The baseline clips last about 5 seconds, their input interfaces differ, and prompts are shortened in some runs; these examples therefore show the observed outputs rather than isolate intrinsic shot-switching capability.

4.5 Ablation Studies

Reference-Grounded Streaming. Table 2 compares cross-shot continuity while fixing the generator, recent window, and total memory budget. The full model achieves C-S/C-B of 0.82/0.75. Removing reference-derived sink features lowers these scores to 0.51/0.43, while removing historical entity memory gives 0.51/0.46. The latter retains R-S/R-B of 0.87/0.78, compared with 0.88/0.83 for the full model. These results indicate that agreement with supplied references alone does not en-

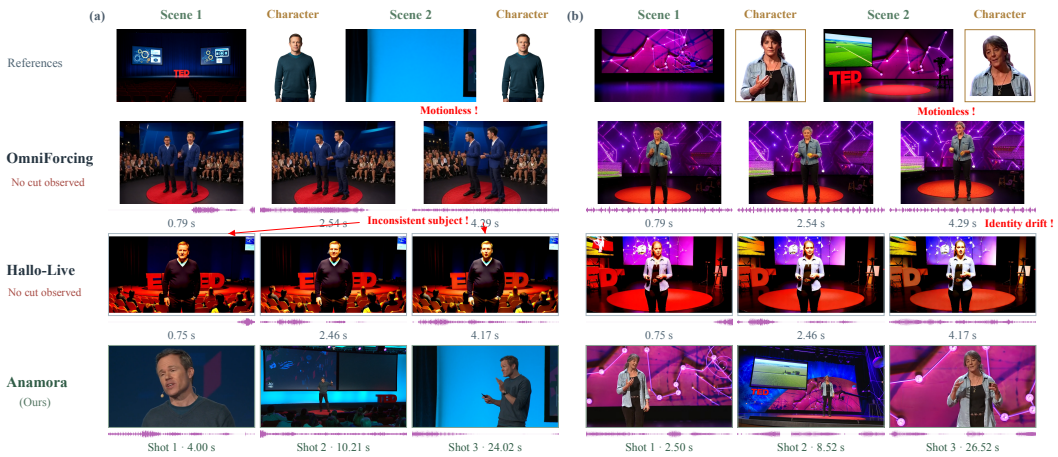

Figure 4: **Qualitative comparison with streaming baselines.** Scene and character references are shown above two paired examples. Anamora preserves the speaker across three camera shots, whereas the saved baseline clips remain single-shot and exhibit appearance or composition differences. Each row samples one video chronologically and uses its own audio; timestamps and output durations differ. Input interfaces and prompt truncation limit causal attribution.

Table 2: **Ablation of reference features and entity memory.** With the generator, recent window, and memory budget fixed, variants test reference features stored in memory, generated history, and state-based selection. Scores assess audiovisual quality, cross-shot continuity, and reference fidelity. "–" denotes unreported results. Best and second-best scores are **bold** and underlined.

Configuration	PQ ↑	CU ↑	VQ ↑	LSE-C ↑	C-S ↑	C-B ↑	R-S ↑	R-B ↑
Reference access								
w/o reference sink	6.73	6.12	3.82	5.56	0.51	0.43	–	–
Historical memory and selection								
w/o historical entity memory	6.42	6.08	3.25	4.87	0.51	0.46	0.87	0.78
State-agnostic entity selection	6.36	5.98	3.91	3.96	0.758	**0.774**	0.84	0.79
Anamora (Ours)	**6.86**	**6.23**	**3.92**	**5.84**	**0.82**	0.75	**0.88**	**0.83**

sure cross-shot continuity: reference features and generated history provide complementary support. Reference-consistency scores are unreported only for the variant without reference sink features.

State-Aware Memory. We compare state-aware and state-agnostic entity selection at the same memory budget. Without state information in selection, C-S decreases from 0.82 to 0.758, while C-B rises from 0.75 to 0.774. R-S/R-B fall from 0.88/0.83 to 0.84/0.79, and PQ falls from 6.86 to 6.36. State-aware selection therefore improves subject consistency and reference fidelity in this comparison, although the state-agnostic variant has higher background similarity. These aggregate appearance scores do not directly measure state correctness.

5 Conclusion

We presented Anamora for reference-grounded streaming audio-video generation. Shot-specific references, fixed-budget State-Aware Entity Memory, and an external agent support identity grounding, compatible-history selection, and condition updates. Experiments show competitive audiovisual quality, while ablations support the complementary roles of reference features and generated history in cross-shot consistency.

AI USE STATEMENT

Generative AI tools assisted with manuscript drafting and language revision, figure layout and editing, media-processing and figure-editing scripts, and qualitative-example selection and presentation. The use of generative models within the audiovisual generation method and reference-construction pipeline is described in the paper and supplementary material. The authors are responsible for verifying all AI-assisted materials and for the final content of this work, including its text, code, claims, and artifacts.

REFERENCES

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion Forcing: Next-token prediction meets full-sequence diffusion. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2407.01392.

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, 2026. URL https://arxiv.org/abs/2605.18739.

Shihao Cheng, Jiaxu Zhang, Quanyue Song, Shansong Liu, Zhizhi Guo, Xiaolei Zhang, Chi Zhang, Xuelong Li, and Zhigang Tu. Unison: Harmonizing motion, speech, and sound for human-centric audio-video generation. arXiv preprint arXiv:2605.08729, 2026. URL https://arxiv.org/abs/2605.08729.

Joon Son Chung and Andrew Zisserman. Out of time: Automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*, 2016. URL https://www.robots.ox.ac.uk/~vgg/publications/2016/Chung16a/.

Yanbo Ding, Zhizhi Guo, Quanyue Song, Yishan He, Zhixiang He, Yongxiang Li, and Yali Wang. Ripple: Real-time streaming audio-video generation with cross-modal recurrent memory. arXiv preprint arXiv:2607.26818, 2026a. URL https://arxiv.org/abs/2607.26818v1.

Yixuan Ding, Jiahao Kong, Wei Huang, Ruijie Quan, and Yi Yang. LayerRecall: A state-conditioned memory router for long-horizon consistency in video generation. arXiv preprint arXiv:2608.28460, 2026b. URL https://arxiv.org/abs/2608.28460.

Weijia Dou, Hui Li, Jiahao Cui, Lei Zhou, Jingdong Wang, and Siyu Zhu. SlotMemory: Object-centric KV memory for streaming long-video generation. arXiv preprint arXiv:2605.31033, 2026. URL https://arxiv.org/abs/2605.31033.

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. arXiv preprint arXiv:2601.03233, 2026. URL https://arxiv.org/abs/2601.03233.

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*, 2025a. URL https://arxiv.org/abs/2506.08009.

Yubo Huang, Hailong Guo, Fangtai Wu, Weiqiang Wang, 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. arXiv preprint arXiv:2512.04677, 2025b. URL https://arxiv.org/abs/2512.04677.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huang_VBench_Comprehensive_Benchmark_Suite_for_Video_Generative_Models_CVPR_2024_paper.html.

Sihui Ji, Xi Chen, Shuai Yang, Xin Tao, Pengfei Wan, and Hengshuang Zhao. MemFlow: Flowing adaptive memory for consistent and efficient long video narratives. arXiv preprint arXiv:2512.14699, 2025. URL https://arxiv.org/abs/2512.14699.

Chunyu Li, Jiaye Li, Ruiqiao Mei, Haoyuan Xia, Hao Zhu, Jingdong Wang, and Siyu Zhu. HalloLive: Real-time streaming joint audio-video avatar generation with asynchronous dual-stream and human-centric preference distillation. arXiv preprint arXiv:2604.23632, 2026. URL https://arxiv.org/abs/2604.23632.

Hui Li, Mingwang Xu, Yun Zhan, Shan Mu, Jiaye Li, Kaihui Cheng, Yuxuan Chen, Tan Chen, Mao Ye, Jingdong Wang, and Siyu Zhu. OpenHumanVid: A large-scale high-quality dataset for enhancing human-centric video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7752–7762, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Li_OpenHumanVid_A_Large-Scale_High-Quality_Dataset_for_Enhancing_Human-Centric_Video_Generation_CVPR_2025_paper.html.

Lightricks. LTX-2.3 model card. Hugging Face model documentation, n.d. URL https://huggingface.co/Lightricks/LTX-2.3. Accessed September 24, 2026.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Yilai Liu, Xin Zhang, Shiyuan Zhang, and Hongyang Du. SlotMem: Character-addressable internal memory for narrative long video generation. arXiv preprint arXiv:2607.15772, 2026. URL https://arxiv.org/abs/2607.15772.

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.

MiniMax. Open general intelligence: MiniMax H3 is now open source. Model release and technical documentation, August 2026. URL https://www.minimax.io/news/minimax-h3-open-source.

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 Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193, 2023. URL https://arxiv.org/abs/2304.07193.

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://arxiv.org/abs/2008.10010.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Christoph Schuhmann. LAION-Aesthetics. LAION technical blog, August 2022. URL https://laion.ai/blog/laion-aesthetics/.

SII-GAIR, Sand.ai, Ethan Chern, Hansi Teng, Hanwen Sun, Hao Wang, Hong Pan, Hongyu Jia, Jiadi Su, Jin Li, Junjie Yu, Lijie Liu, Lingzhi Li, Lyumanshan Ye, Min Hu, Qiangang Wang, Quanwei Qi, Steffi Chern, Tao Bu, Taoran Wang, Teren Xu, Tianning Zhang, Tiantian Mi, Weixian Xu, Wenqiang Zhang, Wentai Zhang, Xianping Yi, Xiaojie Cai, Xiaoyang Kang, Yan Ma, Yixiu Liu, Yunbo Zhang, Yunpeng Huang, Yutong Lin, Zewei Tao, Zhaoliang Liu, Zheng Zhang, Zhiyao Cen, Zhixuan Yu, Zhongshu Wang, Zhulin Hu, Zijin Zhou, Zinan Guo, Yue Cao, and Pengfei Liu. Speed by simplicity: A single-stream architecture for fast audio-video generative foundation model. arXiv preprint arXiv:2603.21986, 2026. URL https://arxiv.org/abs/2603.21986.

SII-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. arXiv preprint arXiv:2602.08794, 2026. URL https://arxiv.org/abs/2602.08794.

Quanyue Song, Yishan He, Yanbo Ding, Zhixiang He, Yongxiang Li, Caigui Jiang, and Zhi Zhi Guo. OmniMate: Open-ended real-time streaming audio-visual generation for interactive avatars. arXiv preprint arXiv:2607.23023, 2026a. URL https://arxiv.org/abs/2607.23023v2.

Quanyue Song, Yishan He, Yanfei Zhang, Shihao Cheng, Zhixiang He, Zhizhi Guo, Chi Zhang, Xuelong Li, and Caigui Jiang. InteractiveAvatar: Real-time streaming video generation for consistent and intent-aware avatars. arXiv preprint arXiv:2606.22905, 2026b. URL https://arxiv.org/abs/2606.22905.

Yaofeng Su, Yuming Li, Zeyue Xue, Jie Huang, Siming Fu, Haoran Li, Ying Li, Zezhong Qian, Haoyang Huang, and Nan Duan. OmniForcing: Unleashing real-time joint audio-visual generation. arXiv preprint arXiv:2603.11647, 2026. URL https://arxiv.org/abs/2603.11647.

Linrui Tian, Qi Wang, and Bang Zhang. StreamChar: Long-horizon streaming character audio-video generation with decoupled orchestration. arXiv preprint arXiv:2605.25659, 2026. URL https://arxiv.org/abs/2605.25659.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, and Wei-Ning Hsu. Meta Audiobox Aesthetics: Unified automatic quality assessment for speech, music, and sound. arXiv preprint arXiv:2502.05139, 2025. URL https://arxiv.org/abs/2502.05139.

Jente Vandersanden, Matheus Gadelha, Chun-Hao P. Huang, Hyeonho Jeong, and Yulia Gryaditskaya. EM-Vid: Training-free entity-centric memory for efficient and consistent multi-shot video generation. arXiv preprint arXiv:2605.23610, 2026. URL https://arxiv.org/abs/2605.23610.

Duomin Wang, Wei Zuo, Aojie Li, Ling-Hao Chen, Xinyao Liao, Deyu Zhou, Zixin Yin, Xili Dai, Daxin Jiang, and Gang Yu. UniVerse-1: Unified audio-video generation via stitching of experts. arXiv preprint arXiv:2509.06155, 2025a. URL https://arxiv.org/abs/2509.06155.

Qinghe Wang, Xiaoyu Shi, Baolu Li, Weikang Bian, Quande Liu, Huchuan Lu, Xintao Wang, Pengfei Wan, Kun Gai, and Xu Jia. MultiShotMaster: A controllable multi-shot video generation framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16268–16278, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wang_MultiShotMaster_A_Controllable_Multi-Shot_Video_Generation_Framework_CVPR_2026_paper.html.

Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, Fei Yang, Pengfei Wan, and Di Zhang. Koala-36M: A large-scale video dataset improving consistency between fine-grained conditions and video content. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8428–8437, 2025b. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_Koala-36M_A_Large-scale_Video_Dataset_Improving_Consistency_between_Fine-grained_Conditions_CVPR_2025_paper.html.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. InternVid: A large-scale video-text dataset for multimodal understanding and generation. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=MLBdiWu4Ew.

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. arXiv preprint arXiv:2509.22622, 2025. URL https://arxiv.org/abs/2509.22622.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo 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, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yin_One-step_Diffusion_with_Distribution_Matching_Distillation_CVPR_2024_paper.html.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22963–22974, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yin_From_Slow_Bidirectional_to_Fast_Autoregressive_Video_Diffusion_Models_CVPR_2025_paper.html.

Kaiwen Zhang, Liming Jiang, Angtian Wang, Jacob Zhiyuan Fang, Tiancheng Zhi, Qing Yan, Hao Kang, Xin Lu, and Xingang Pan. StoryMem: Multi-shot long video storytelling with memory. arXiv preprint arXiv:2512.19539, 2025. URL https://arxiv.org/abs/2512.19539.

ANAMORA: SUPPLEMENTARY MATERIAL

Anonymous authors
Paper under double-blind review

A DATA PREPARATION

Figure 5 summarizes the preparation of OpenHumanVid (Li et al., 2025) and Koala (Wang et al., 2025), aligning audiovisual content with entities, states, and references. Prepared inventories and actual training consumption are recorded separately.

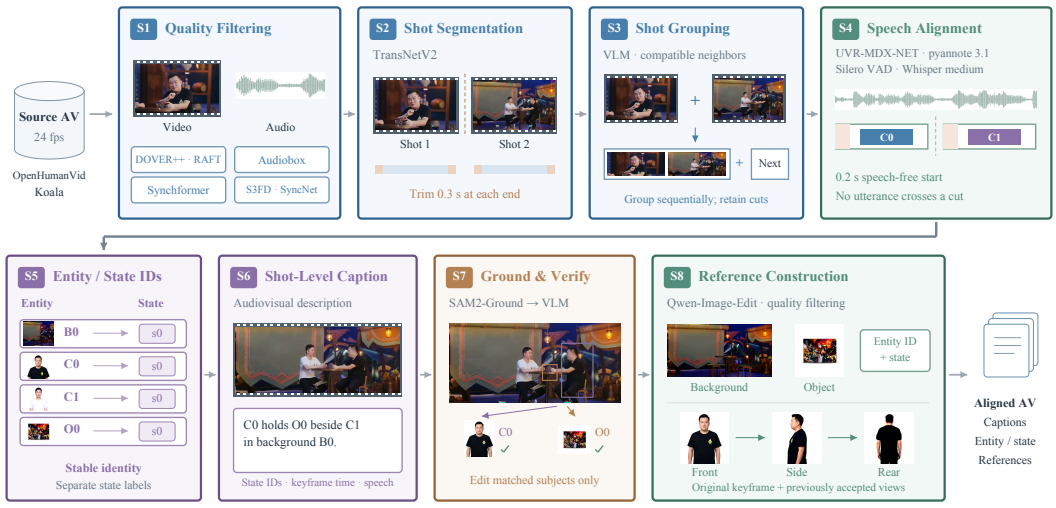

Figure 5: **Data preparation pipeline.** Top: quality filtering, shot segmentation and grouping, and speech alignment. Bottom: entity/state indexing, shot-level captions, verified subject extraction, and reference construction. Character views use the original keyframe and previously accepted views. Paired source frames and references illustrate the workflow; speech intervals and grounding boxes are schematic. Trimming and speech-boundary rules concern data preparation, not inference latency.

S1: Quality Filtering. Videos are converted to 24 fps and screened for resolution, aspect ratio, DOVER++ quality, and RAFT motion estimates. Audio checks cover silence ratio, audio resolution, and Audiobox Aesthetics (Tjandra et al., 2025). Synchformer (Iashin et al., 2024) checks global synchronization, while S3FD face tracks support SyncNet (Chung & Zisserman, 2016) lip alignment for visible speakers.

S2–S4: Shot and Speech Alignment. TransNetV2 identifies shots, which are trimmed by 0.3 seconds at each end. A vision–language model (VLM) groups compatible neighboring shots sequentially, retaining their cut timestamps rather than treating a group as one continuous shot. UVR-MDX-NET, pyannote 3.1, Silero VAD, and Whisper medium (Radford et al., 2023) provide source separation, speaker identities, speech intervals, and transcripts. Validation requires a 0.2-second speech-free start and complete utterances within each retained shot; speech-free does not mean silent, and these constraints do not impose an inference-time waiting period.

S5–S6: Entity-State Annotation. Backgrounds, characters, and objects receive persistent entity IDs, with separate state IDs rather than new identities for every appearance change. Shot-level audiovisual captions bind these IDs to applicable states, keyframe timestamps, and aligned speech, making identity and state separately addressable in subsequent reference construction and memory supervision.

S7–S8: Reference Construction. SAM2-Ground extracts entities from keyframes, and a VLM verifies their correspondence with caption subjects before editing. Qwen-Image-Edit constructs unobstructed backgrounds and complementary character/object views; front, side, and rear character views are generated progressively using the original keyframe and previously accepted views. Quality and correspondence filtering precede binding the resulting references to entity/state IDs; canonical multiview inputs are optional at inference.

B Additional Related Work

Advances in artificial intelligence have enabled progress in visual perception, multimodal understanding, reasoning, and content generation. Research across these areas explores how models can interpret complex inputs, integrate information across modalities, and support richer interactions with users.

References

Joon Son Chung and Andrew Zisserman. Out of time: Automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*, 2016. URL https://www.robots.ox.ac.uk/~vgg/publications/2016/Chung16a/.

Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5325–5329, 2024. URL https://www.robots.ox.ac.uk/~vgg/research/synchformer/.

Hui Li, Mingwang Xu, Yun Zhan, Shan Mu, Jiaye Li, Kaihui Cheng, Yuxuan Chen, Tan Chen, Mao Ye, Jingdong Wang, and Siyu Zhu. OpenHumanVid: A large-scale high-quality dataset for enhancing human-centric video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7752–7762, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Li_OpenHumanVid_A_Large-Scale_High-Quality_Dataset_for_Enhancing_Human-Centric_Video_Generation_CVPR_2025_paper.html.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, and Wei-Ning Hsu. Meta Audiobox Aesthetics: Unified automatic quality assessment for speech, music, and sound. arXiv preprint arXiv:2502.05139, 2025. URL https://arxiv.org/abs/2502.05139.

Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, Fei Yang, Pengfei Wan, and Di Zhang. Koala-36M: A large-scale video dataset improving consistency between fine-grained conditions and video content. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8428–8437, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_Koala-36M_A_Large-scale_Video_Dataset_Improving_Consistency_between_Fine-grained_Conditions_CVPR_2025_paper.html.