JINGDONG

JoyAI-Echo: Pushing the Frontier of Long Audio-Visual Generation

Haoran Li[1,*], **Fredreic Li**[2,*], **Shichen Ma**[1,*], **Jie Huang**[1], **Yijun Liu**[3], **Jiaqi Shi**[4], **Yanwen Ma**[5], **Yaofeng Su**[6], **Xin Lu**[4], **Haoyu Wang**[3], **Xiaoxiao Ma**[4], **Guohui Zhang**[4], **Yaowei Li**[2], **Mingchen Zhong**[4], **Junhao Zhuang**[1], **Songchun Zhang**[7], **Weiyang Jin**[8], **Yuxuan Bian**[9], **Shiyi Zhang**[3], **Haojun Xu**[5], **Shuai Lu**[1], **Xin Han**[1], **Wei Tang**[1], **Tong He**[1], **Jiaqi Wang**[1], **Ping Luo**[8], **Haoyang Huang**[1], **Zeyue Xue**[1,8,*,+], **Nan Duan**[1]

[1]Joy Future Academy, JD, [2]Peking University, [3]Tsinghua University, [4]The University of Science and Technology of China, [5]Beihang University, [6]Fudan University, [7]The Hong Kong University of Science and Technology, [8]The University of Hong Kong, [9]The Chinese University of Hong Kong,

* denotes equal contribution. [+] denotes project lead.

Abstract

Long video generation continues to be plagued by error accumulation, weak temporal coherence, and prohibitive latency, hindering its deployment in interactive settings. We present **JoyAI-Echo**, a framework that circumvents these challenges through four pivotal innovations. At its core, a cross-modal audio-visual memory bank reliably preserves character appearance and vocal timbre across five-minute videos, while a post-training pipeline integrates memory-based reinforcement learning with distribution matching distillation, achieving a 7.5× speedup and substantially enhancing visual quality and alignment. Leveraging these components, JoyAI-Echo markedly outperforms *Happy Oyster* (Directing mode) on long-form generation and surpasses the short-video specialist *Wan 2.6* on human-centric tasks. Beyond generation fidelity, an interactive agent enables real-time editing via conversational instructions, and a lightweight super-resolution module preserves high definition under streaming latency, collectively delivering an instantly editable, conversation-speed video creation experience. For the first time, JoyAI-Echo concurrently achieves long-range cross-modal consistency, real-time inference for minute-scale videos, conversational interactivity, and high-resolution output—without compromise—heralding a new paradigm of interactive video generation. Code and model weights will be publicly released.

Date: June 26, 2026
Code: https://github.com/jd-opensource/JoyAI-Echo
Project Page: https://echo-team-joy-future-academy-jd.github.io/Echo-LongVideo-Page/

1 Introduction

Recent advances in video generation have achieved remarkable success in producing short, high-fidelity clips, yet scaling these methods to long-form, coherent videos remains a fundamental challenge. Prior works such as Pixverse-R1 and Happy Oyster make initial attempts at long video generation, but they often suffer from error accumulation, insufficient long-range temporal modeling, or prohibitive computational costs as the video length grows. These limitations effectively prevent the deployment of video generation in interactive

applications such as virtual storytelling, digital human assistants, and real-time content creation, where consistency, speed, and responsive user control are indispensable.

In this paper, we present **JoyAI-Echo**, a novel framework that advances the frontier of long video generation through four core innovations, organized in two complementary tiers. At its foundation, a cross-modal audio-visual memory bank and a systematic memory-based post-training pipeline directly overcome the central bottlenecks of long-form consistency and inference speed, and together they are responsible for the model's decisive performance advantages. The *audio-visual memory bank* explicitly stores and retrieves identity-related visual features and speaker voice embeddings throughout the generation process, breaking the context-forgetting bottleneck that has plagued multi-step generation and ensuring that character appearance and voice timbre remain highly consistent across videos spanning five minutes. The *memory-based post-training pipeline* incorporates a memory-based Supervised Fine-Tuning (SFT) stage and a cross-modal Reinforcement Learning with Human Feedback (RLHF) stage to optimize visual quality and audio quality, which is followed by a memory-based Distribution Matching Distillation (DMD) stage for a 7.5 $\times$ speedup , thereby removing the hardware barrier and making responsive long audio-visual synthesis practically attainable. Empowered by these two components, the native JoyAI-Echo model decisively outperforms *Happy Oyster* (directing mode) on long-form generation and even surpasses the short-video specialist *Wan 2.6* on human-centric tasks.

Building upon this strong generation backbone, the remaining two innovations further elevate the overall experience and unlock an entirely new mode of creation. An *interactive long audio-visual generation agent* allows users to modify the content on the fly by simply inputting conversational instructions, and the agent instantly iterates on the video based on this real-time feedback. The generation flow thus evolves from a static process into a dynamic dialogue where the output continuously adapts to the user's intent. Complementing this, a *real-time super-resolution module*, designed as a lightweight component that co-operates with the accelerated generation backbone, decouples the computational burden of high-resolution synthesis from the autoregressive generation process. This ensures that visual sharpness keeps pace with the speed of creation under the strict latency constraints of streaming. Together, the interactive agent and the super-resolution module deliver instantly editable, high-definition video content generated at the speed of human conversation.

To the best of our knowledge, JoyAI-Echo is the first framework to simultaneously provide long-term cross-modal consistency, real-time generation for minute-level videos, conversational interactivity, and high-resolution output without compromising any of these properties. Meanwhile, JoyAI-Echo also preserves and further enhances its capability for short-form audio-visual generation. This work thus inaugurates a new era of interactive visual media tools, where coherent, high-resolution, and instantly editable video content is created at the speed of conversation.

2 Data

End-to-end long audio-visual generation requires training data that exposes the model to the same character across temporally separated shots. Unlike short-clip generation, where a sample usually contains a single continuous motion trajectory, minute-scale narrative videos require the model to preserve identity while the background, camera pose, lighting, clothing, facial expression, and speech content may all change. We therefore construct an identity-centric video corpus whose basic unit is not an isolated clip, but a set of visually diverse single-shot clips associated with the same recurring character.

Starting from a million-scale collection of films, TV episodes, and long-form web videos, our pipeline extracts over **one million unique character identities**. Each retained identity is associated with at least a fixed minimum number of scene-disjoint single-shot clips. The clips are temporally bounded, restricted to a small number of co-visible people, and selected to exhibit sufficient scene diversity. This design differs from flat human-video datasets that primarily optimize independent clip quality [4, 12]: our corpus is explicitly organized by identity and is annotated with visual, audio and quality metadata needed for long-horizon identity-consistent generation. Fig. 1 summarizes the construction procedure.

Figure 1 Identity-consistent data construction pipeline. Long-form videos are processed by global identity discovery and scene-level diversity selection first, followed by local tracklet assignment, quality filtering and audio-visual captioning.

2.1 Identity-Consistent Clip Extraction

Given a long-form video, our goal is to identify recurring character identities and extract, for each identity, single-shot clips from visually diverse scenes. However, naive face clustering often over-merges visually similar characters or incorrectly splits the same character under variations in lighting, clothing, and viewpoint. To address this, we perform identity discovery at a global level, while assigning identities to local tracklets only when sufficient within-track evidence is available.

Global identity prototypes. Videos are first routed according to their content domain to select an appropriate recognition backbone and similarity calibration strategy, and are then sampled using a memory-bounded streaming reader. Let $\{\mathbf{e}_i\}_{i=1}^{N}$ denote the unit-normalized face embeddings. We perform global clustering with DBSCAN [6] under cosine distance:

$$D_{ij} = 1 - \mathbf{e}_i^{\top}\mathbf{e}_j. \tag{1}$$

Only clusters with sufficient temporal support are retained. For each valid cluster $\mathcal{C}_k$, we compute a centroid prototype $\bar{\mathbf{e}}_k$ and an adaptive acceptance threshold:

$$t_k = \max(\mu_k - \kappa\sigma_k,\ t_{\text{hard}}), \tag{2}$$

where μ_k and σ_k are intra-cluster similarity statistics, κ is domain-dependent, and t_{hard} is a conservative lower bound. This yields per-identity prototypes robust to varying pose and illumination.

Scene grouping. The video is segmented into shots using PySceneDetect [1]. Adjacent or visually related shots are then merged into scene groups by comparing foreground-masked DINOv2 [17] background embeddings and color histograms, with temporal constraints preventing distant but visually similar scenes from being collapsed into the same group. These scene groups serve as the basis for diversity, as clips within the same group typically share similar environments, wardrobe, lighting conditions, and narrative context.

Local role assignment. Within each scene group, person detections are associated into tracklets by bounding-box overlap and terminated at shot boundaries to reduce identity leakage. For a tracklet $\mathcal{T}$, sharp and temporally distributed face observations are compared with all prototypes. Let $v_k(\mathcal{T})$ be the number of observations assigned to prototype k. The tracklet is assigned to identity:

$$g^{\star} = \arg\max_{k} v_k(\mathcal{T}) \tag{3}$$

only if its consensus ratio:

$$r(\mathcal{T}) = \frac{v_{g^{\star}}(\mathcal{T})}{\sum_k v_k(\mathcal{T})} \tag{4}$$

exceeds a strict confidence requirement. Tracklets with weak occupancy, unstable face evidence, excessive co-visible persons, or ambiguous votes are rejected. Surviving segments are trimmed to single-shot clips and assigned a utility score:

$$Q(\mathcal{T}) = w_v r(\mathcal{T}) + w_s S(\mathcal{T}) + w_m M(\mathcal{T}), \tag{5}$$

where S denotes normalized sharpness, M denotes motion or body displacement, and the weights are calibrated by video domain.

Diversity selection. For each identity, candidates are de-duplicated within scene groups by retaining the highest-quality representative. A greedy diversity pass then ranks by the above utility score Q and suppresses near-duplicates using perceptual hashes, color histograms, and DINOv2 background similarity. Identities whose surviving clips fail to meet the required coverage are discarded, preserving the invariant that every retained identity has multiple high-quality and visually distinct observations.

2.2 Quality Filtering and Metadata Annotation

After extraction, we apply a multi-axis filter combining learned video-quality assessment, frame-level clarity, text-overlay detection, motion statistics, and identity coverage constraints.

First, a video-quality model [15] filters samples with severe compression artifacts, abnormal color, unstable exposure, or low aesthetic quality.

Second, uniformly sampled frames are evaluated by the variance of the Laplacian, an efficient proxy for structural sharpness. This signal is critical for identity learning: motion blur, defocus, and low-resolution faces may pass detection but still provide unreliable supervision. We aggregate frame-level sharpness into clip-level statistics and remove candidates with insufficiently clear identity evidence.

Third, OCR and overlay detectors identify subtitles, logos, watermarks, large text regions, and burned-in graphics; the metadata is used to exclude clips where text dominates the frame or repeatedly occludes the person.

Fourth, optical-flow scores [7] estimate whole-frame dynamics, mainly to avoid a distribution dominated by static talking-head clips while preserving reliable identity evidence. Filtering is also applied at the identity level: if an identity no longer has enough scene-disjoint high-quality clips, all remaining clips for that identity are removed.

2.3 Audio-Visual Captioning

The final stage attaches structured audio-visual captions to the retained clips. The audio stream is extracted and paired with the video under a multimodal captioning prompt, so the annotation can jointly describe visible appearance, motion, speech, and non-speech sound. To keep the signal aligned with long audio-visual generation, the schema is deliberately compact. It records the character identity description and voice timbre, including salient appearance cues and stable vocal characteristics; the main actions, interactions, facial expressions, and body motion in temporal order; the spoken content or dialogue when present; and background audio such as music, ambient sound, crowd noise, or salient effects. We also preserve concise OCR-related descriptions when text is visually relevant.

These captions provide multimodal conditioning that complements identity grouping: across clips, the model observes not only how the same character looks in different scenes, but also how the character speaks, moves, and interacts with the acoustic environment.

3 Long-Term Audio-Visual Memory

Task formulation. Given a story script represented by a sequence of shot-level conditions $\mathcal{C} = \{c_t\}_{t=1}^{T}$, JoyAI-Echo generates a sequence of audio-visual shots $\mathcal{S} = \{S_t\}_{t=1}^{T}$, where each shot contains a video-audio pair $S_t = (V_t, A_t)$. Following memory-based multi-shot generation, JoyAI-Echo decomposes long-form audio-visual generation into iterative shot synthesis conditioned on an evolving memory bank:

$$p_\theta(\mathcal{S} \mid \mathcal{C}) = \prod_{t=1}^{T} p_\theta(S_t \mid c_t, \mathcal{M}_{t-1}), \qquad \mathcal{M}_t = \mathcal{U}(\mathcal{M}_{t-1}, S_t). \tag{6}$$

Here, $\mathcal{M}_{t-1}$ stores compact historical audio-visual cues before generating the t-th shot, and $\mathcal{U}$ denotes the memory update function. Memory tokens are used only as conditional context rather than prediction targets; the model predicts only the current video and audio outputs.

3.1 Memory Mechanism

Slot-paired audio-visual memory. JoyAI-Echo maintains a compact memory bank composed of slot-paired audio-visual events:

$$\mathcal{M}_t = \{m_i\}_{i=1}^{K}, \qquad m_i = (m_i^v, m_i^a), \tag{7}$$

where K is the number of retained memory slots, m_i^v denotes the visual memory of the i-th historical event, and m_i^a denotes its paired audio memory. The slot-paired design explicitly binds appearance and timbre at the event level, which is critical for preserving face-voice correspondence in multi-shot audio-visual storytelling.

Audio memory. Audio memory is organized as an ordered sequence of event-level acoustic segments. For each historical shot, the waveform is converted into a mel-spectrogram, and a bounded temporal window is selected using a maximum-response criterion:

$$\tau_i^* = \arg\max_{\tau} \left\| \mathrm{Mel}(A_i)_{\tau-\frac{\Delta}{2}:\tau+\frac{\Delta}{2}} \right\|_1, \tag{8}$$

where Δ is the temporal length of the selected audio window. The selected segment is then independently encoded by the audio VAE as $m_i^a = E_a(\mathrm{Mel}(A_i)_{\tau_i^*-\frac{\Delta}{2}:\tau_i^*+\frac{\Delta}{2}})$, and the resulting latents are concatenated in slot order to form the audio-memory sequence. This window-bounded design captures the vocal timbre without introducing excessive historical speech content, rhythm, or temporal dynamics.

Visual memory. Visual memory serves as an appearance-oriented conditioning signal aligned with the selected audio-memory event. Specifically, the selected audio window is assigned to the video timeline through $\phi(\tau_i^*)$, and a 9-frame clip around the aligned center frame is encoded by the video VAE. Only the final latent state is retained as visual memory, written inline as $m_i^v = \mathrm{Last}(E_v(V_i[\phi(\tau_i^*) - 4 : \phi(\tau_i^*) + 4]))$. This design keeps the memory compact while preserving identity, appearance, and short-term speaking-related cues, such as mouth state.

Memory update and positional encoding. The memory bank is dynamically updated using a hybrid temporal retention strategy that combines long-range anchor shots with recent contextual shots:

$$\mathcal{M}_t = \mathrm{Encode}\left(\mathrm{First}_3(\mathcal{H}_t) \cup \mathrm{Recent}_4(\mathcal{H}_t)\right), \tag{9}$$

where $\mathcal{H}_t = \{S_i\}_{i=1}^{t}$ denotes the generated history. The first three shots are retained as persistent anchor references, while the most recent four shots provide short-term contextual support. In addition, memory tokens and target tokens use independent positional encoding with separate temporal origins, preventing memory from being interpreted as a temporal prefix of the current shot.

3.2 Memory Interaction

As shown in Fig. 2, the slot-paired memory tokens are injected into the video and audio diffusion branches through layer-specific audio-memory self-attention and slot-aware cross-modal attention masks.

Layer-specific audio-memory interaction. Audio memory is injected into the audio diffusion branch in a stage-wise manner. In the first 70% of audio transformer layers, target audio tokens are isolated from memory tokens through a block attention mask, which allows the model to establish current speech content, local temporal structure, and articulation without historical interference. In the remaining 30% of layers, the mask is removed to enable memory-target interaction and incorporate vocal timbre after target-side audio representations are formed. The attention bias for the l-th audio layer is defined as:

$$B_l^a(q,k) = \begin{cases} -\infty, & l \le 0.7L_a,\ q \in \mathcal{T}^a,\ k \in \mathcal{M}^a, \\ 0, & \text{otherwise}, \end{cases} \tag{10}$$

where $\mathcal{T}^a$ denotes target audio tokens, $\mathcal{M}^a$ denotes audio-memory tokens, and L_a is the number of audio transformer layers.

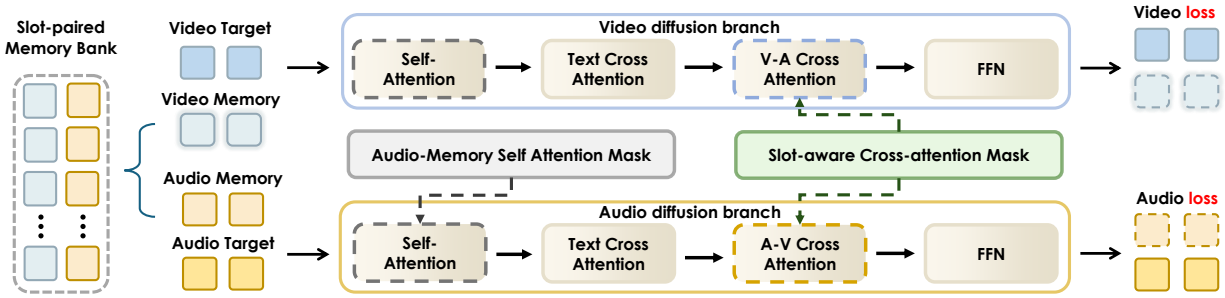

Figure 2 Overview of the slot-paired audio-visual memory interaction mechanism. JoyAI-Echo maintains a slot-paired memory bank in which each historical event contains aligned visual and audio memory tokens. During generation, video and audio target tokens are processed by two diffusion branches, while memory tokens are used only as conditional context and are excluded from loss computation. In the audio branch, an audio-memory self-attention mask controls layer-specific interaction between target audio tokens and audio memory tokens. In the cross-modal modules, a slot-aware cross-attention mask enforces one-to-one interaction between paired visual and audio memory slots, preventing cross-event face-voice mixing. The model therefore preserves long-range visual identity and speaker timbre while predicting only the current video and audio targets.

Slot-aware cross-modal interaction. Cross-modal memory interaction is constrained by a strict slot-aligned masking scheme. In both audio-to-video and video-to-audio cross-attention layers, the i-th visual-memory slot is allowed to interact only with the i-th audio-memory slot:

$$B_{ij}^{av} = \begin{cases} 0, & i = j, \\ -\infty, & i \neq j. \end{cases} \tag{11}$$

Cross-slot interactions across different historical events are explicitly prohibited. Meanwhile, target-video tokens attend to target-audio tokens rather than historical audio-memory tokens, ensuring that the current mouth motion is driven by the current speech. This slot-aware design preserves event-level alignment between visual appearance and acoustic identity, and mitigates incorrect face-voice mixing.

Branch-specific conditioning. The memory bank is used as structured conditioning rather than as an unrestricted context extension. Visual memory mainly contributes identity and appearance cues, audio memory mainly contributes speaker timbre, and cross-modal memory interaction is introduced only through layer-specific and slot-aligned masks. This separates appearance-oriented long-range consistency from synchronization-sensitive current-shot generation.

3.3 Training

Latent-space target prediction. Following the latent diffusion / rectified-flow setting, JoyAI-Echo encodes the target shot into audio-visual latents and trains the model to predict the target velocity conditioned on the prompt and memory. Let $z_0^t = E(S_t)$ denote the clean target audio-visual latent, $z_1 \sim \mathcal{N}(0, I)$ denote Gaussian noise, and $z_\sigma^t = (1-\sigma)z_0^t + \sigma z_1$ denote the interpolated latent. The memory-conditioned velocity objective is:

$$\mathcal{L}_{\mathrm{rf}} = \mathbb{E}_{t,\sigma}\left[\left\| v_\theta(z_\sigma^t, \sigma, c_t, \mathcal{M}_{t-1}) - (z_1 - z_0^t) \right\|_2^2\right]. \tag{12}$$

Memory tokens are treated only as conditions, and losses are computed only on target shot tokens.

Memory-length-aware loss reweighting. During training, the number of memory slots is varied from 0 to 7, matching the inference process where early shots contain little memory while later shots rely on multiple historical references. Since longer memory introduces heavier conditioning and makes speech-driven facial motion more difficult, the video-side objective is reweighted according to memory length:

$$\mathcal{L}_{\mathrm{mem}} = \lambda_v(K)\mathcal{L}_v + \mathcal{L}_a, \qquad \lambda_v(K) = 1 + \alpha\frac{K}{K_{\max}}, \tag{13}$$

where $\mathcal{L}_v$ and $\mathcal{L}_a$ denote video target and audio target losses, respectively. This provides stronger visual supervision under long-context conditioning and mitigates lip synchronization degradation.

Audio-to-video gradient amplification. JoyAI-Echo further amplifies the gradient contribution associated with the audio-to-video cross-modal interaction while keeping the forward computation unchanged, written inline as $\nabla_\theta^{a\to v} \leftarrow \gamma \nabla_\theta^{a\to v}$. This strengthens the coupling between speech and mouth motion without modifying the inference architecture. Training is performed in two stages: the first stage uses a maximum memory-length reweighting factor of 2 and an audio-to-video gradient amplification factor of 2; the second stage increases them to 4 and 6, respectively.

3.4 Inference

Iterative memory-conditioned generation. During inference, JoyAI-Echo generates the story shot by shot. For each shot, the current output is generated according to:

$$S_t \sim p_\theta(S_t \mid c_t, \mathcal{M}_{t-1}), \qquad \mathcal{M}_t = \mathcal{U}(\mathcal{M}_{t-1}, S_t). \tag{14}$$

After generating S_t, representative audio-visual events are extracted and written back into memory using the same slot-paired construction and anchor-recent update strategy as in training. This forms an iterative read-write loop for long-form audio-visual storytelling.

Target-video self-attention scaling. Since memory-conditioned audio tokens may still introduce interference to the video branch, the self-attention of target video tokens is enhanced during inference by scaling its attention temperature, i.e., $\mathrm{Attn}_v = \mathrm{Softmax}(Q_v K_v^\top / (\tau\sqrt{d}))V_v$. This strengthens target-side video token aggregation and alleviates the negative effects introduced by memory-influenced audio representations, while introducing no additional training objective or memory slot.

4 Memory-based Post Training

Although long-term audio-visual memory training equips JoyAI-Echo with cross-shot consistency, the resulting model still has room for improvement in several aspects, including visual quality and resolution, audio-visual fidelity and synchronization, as well as inference speed and efficiency. This motivates a **memory-based post-training** framework that progressively refines the model from complementary perspectives.

The memory-based post-training framework consists of memory-based SFT, cross-modal RLHF, and memory-based DMD. Memory-based SFT introduces high-quality single-shot data, progressively increases the generation resolution from 480p to 720p, thereby enhancing visual quality while preserving memory capability through probabilistic multi-shot training. cross-modal RLHF further aligns the model with human preferences by improving visual quality, audio fidelity, and audio-visual synchronization. Memory-based DMD then accelerates the RLHF-enhanced model by pushing it toward high-probability generation regions, while maintaining overall generation quality and long-form memory-conditioned audio-visual generation ability.

4.1 Memory-based Supervised Fine-Tuning (SFT)

Due to the scarcity of high-quality multi-shot audio-visual videos, we further fine-tune JoyAI-Echo using additional collected high-quality single-shot videos to enhance its generation quality. Since single-shot training can be regarded as a special case of multi-shot training with zero memory, both single-shot and multi-shot data can be naturally integrated within the same training framework. Meanwhile, the multi-shot data from Sec. 3 is sampled with a predefined probability during fine-tuning to preserve the model's multi-shot memory capability and cross-shot consistency.

To better exploit high-quality single-shot data, we reduce the occurrence of zero-memory cases in the multi-shot data, encouraging the model to learn high-quality generation from single-shot samples while retaining memory-conditioned long-form generation ability. For a smoother optimization process, we employ a progressive resolution schedule, first fine-tuning the model on 480p videos and then continuing the training on 720p videos to improve high-resolution generation quality. With this training strategy, JoyAI-Echo improves visual quality not only for long-form video generation, but also for short-form generation.

4.2 Cross-Modal Reinforcement Learning with Human Feedback (RLHF)

After training the base model, JoyAI-Echo aligns the generation policy with human preferences through a cross-modal RLHF [23] stage. Specifically, JoyAI-Echo introduces a modality-aware diffusion reinforcement learning framework termed **OmniNFT** [26] tailored for joint audio-visual generation. This work identifies three fundamental optimization bottlenecks when naively applying vanilla reinforcement learning to multi-modal generation scenarios: first, the reward advantages for video and audio modalities are frequently inconsistent, where high-fidelity video samples do not necessarily correspond to high-quality audio; second, gradients from the video branch tend to leak into the shallow layers of the audio network, disrupting the intra-modal generation process; and third, the uniform credit assignment strategy fails to account for the unequal contributions of critical regions to audio-visual synchronization.

To address these challenges, OmniNFT introduces three coordinated technical designs: 1) Modality-wise advantage routing computes independent advantages for video quality, audio fidelity, and cross-modal synchronization, respectively, and dispatches each advantage signal to its corresponding generation branch; 2) Layer-wise gradient surgery partially detaches video-branch gradients on shallow audio layers while preserving intact gradient flows for deep cross-modal interaction blocks; 3) Region-wise loss reweighting leverages the video-to-audio cross-attention maps as an intrinsic proxy to localize sound-emitting critical regions, and amplifies the optimization intensity on these perceptually sensitive areas.

4.3 Acceleration with Memory-based Distribution Matching Distillation (DMD)

Long-form audio-visual generation requires not only long-range consistency, but also a low-latency generation pipeline. Recent efficient video and audio-visual generators have shown that latent space modeling and streaming distillation can substantially reduce the generation cost [8, 9, 21]. To make JoyAI-Echo practical for interactive scenarios, we introduce a memory-conditioned acceleration framework that converts the original multi-step audio-visual generator into a few-step student while preserving the long-term audio-visual memory interface described in Sec. 3.

Memory-based Bidirectional DMD. Starting from the memory-aware model, memory-based bidirectional DMD is applied to distill the original multi-step bidirectional generator into an 8-step student model. The teacher, student, and distribution-matching critic all receive the same shot condition and committed audio-visual memory. This design ensures that distillation preserves not only short-clip generation quality, but also the memory-conditioned behavior required for long-form video rollout. The memory used in this stage follows the same construction as in Sec. 3, so that the accelerated student remains aligned with the memory distribution learned during multi-shot training.

For the acceleration stages, video and audio losses are balanced as:

$$\mathcal{L}_{\text{stage}} = \mathcal{L}^{v}_{\text{stage}} + \lambda_a \mathcal{L}^{a}_{\text{stage}}, \quad \lambda_a = 0.5. \tag{15}$$

The smaller audio coefficient stabilizes joint audio-visual distillation by balancing the gradient scales of the video and audio branches, while still preserving speech, ambient sound, and synchronization quality. In practice, directly applying standard DMD to the audio branch is unstable and can easily introduce audible noise artifacts after distillation. To improve training stability, EMA smoothing is applied to the optimizer momentum buffers during DMD optimization to reduce abrupt update directions caused by noisy audio gradients. The critic update frequency is set to 1, which keeps the distribution-matching critic sufficiently up-to-date while avoiding an overly slow training process.

The memory-related training strategy is kept consistent with Sec. 3. Specifically, the parameter settings of memory-length-aware loss reweighting and audio-to-video gradient amplification are inherited from the memory-aware training stage. To further reduce the train-test mismatch of memory-conditioned rollout, degradation is applied to the memory inputs during DMD training. This simulates the drift that may accumulate in committed memory as shots are generated sequentially at inference time, encouraging the student to remain robust when conditioned on imperfect self-generated history shots.

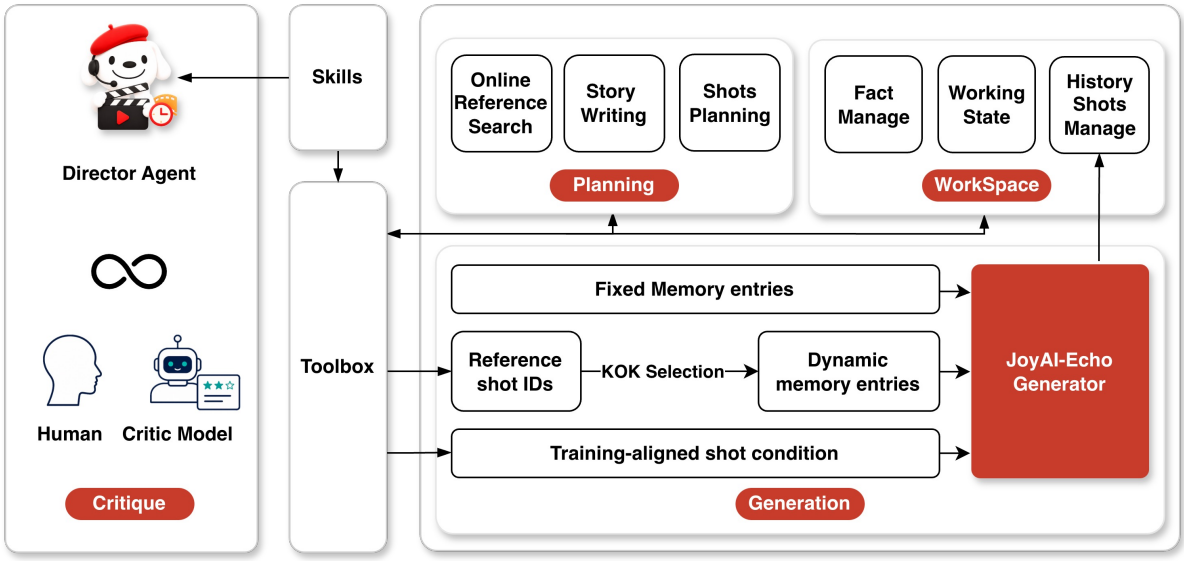

Figure 3 Overview of the Director Agent workflow. The agent organizes long-form video generation into planning, generation, and critique stages. In the planning stage, it expands user intent into structured screenplay, character, scene, and shot-level specifications. In the generation stage, it compiles each shot into training-aligned conditions, retrieves relevant historical shots, constructs fixed and dynamic memory entries with KOK selection, and invokes JoyAI-Echo. In the critique stage, human feedback or an automatic critic model provides shot-level revision signals, which are routed back to the agent for localized regeneration and memory updates.

Extension to Causal Streaming Generation. Beyond the bidirectional DMD student, JoyAI-Echo can be naturally extended to a causal student model for streaming inference. The causal student is first initialized by regressing trajectories from the bidirectional teacher, which bridges full-context denoising and block-wise causal generation, similar to previous slow-to-fast video distillation methods [21, 25]. Memory-causal DMD training is then applied under the same causal contract used at inference time, gradually shifting JoyAI-Echo from ground-truth prefixes and memory states to its own generated histories. This scheduled transition reduces train-test mismatch in autoregressive long video generation and improves robustness to error accumulation [2, 10, 16].

5 Director Agent

Long-form video generation requires not only model-level temporal modeling, but also semantic control across shots. In practical applications, user prompts are often abstract, short, and frequently revised. Although a long-form video generation model may include an internal memory mechanism, its memory retrieval process is usually relatively coarse. Directly using user prompts as model inputs can still cause errors to accumulate across shots, making long-form audio-visual generation challenging for non-expert users.

JoyAI-Echo is trained with explicit and structured shot-level text conditions, while real user inputs are usually much less structured. To bridge this gap, we introduce a Director Agent on top of the generator. The agent converts incomplete or under-specified user inputs into shot conditions aligned with the training distribution, manages long-range references through our agent-level memory mechanism, and supports local revision without regenerating the full video.

5.1 Architecture

As shown in Fig. 3, the Director Agent follows a two-stage planning-and-generation workflow with iterative critique and revision. In the planning stage, the agent expands user intent into a screenplay, characters, scenes, and shot plans. In the generation stage, the agent converts each shot plan into a training-aligned condition, retrieves relevant historical shots, selects dynamic memory entries, invokes the JoyAI-Echo generator, and writes the generated shot back to the history manager. The generated shot can be reviewed by users or automatic critics. If revision is required, the system updates only the affected shot condition and its related memory entries, forming a closed-loop generation process.

At the component level, the workflow is implemented through a persistent Director workspace with a tool-and-skill abstraction. The Director workspace provides a persistent execution context for the proposed screenplay-to-video generation pipeline. It maintains the intermediate and final artifacts produced throughout the workflow, such as the screenplay, shot list, character, scene, and final composition metadata.

To regulate agent-side execution, the system separates state-mutating capabilities from procedural guidance through a tool-and-skill abstraction. For state mutations, tools expose predefined interfaces for modifying persistent workflow artifacts, such as updating project states, recording revisions, and registering generated outputs. By requiring all updates to be performed through explicit interfaces, the system makes state changes traceable and reduces unintended modifications during multi-step agent execution. For procedural guidance, skills encode workflow knowledge that guides the agent through planning, context retrieval, condition preparation, and iterative revision. This separation allows the agent to follow a controlled generation protocol while retaining flexibility in resolving ambiguous user requests, selecting relevant references, and performing localized revisions.

5.2 Planning

The planning stage serves as an intermediate reasoning layer between the user's natural-language request and the structured inputs required by the video generation model. Instead of directly translating the prompt into generation conditions, the agent first organizes the intent into a coherent narrative structure, resolves under-specified details, and establishes constraints that will guide all subsequent shots. This step is crucial for maintaining story coherence, character consistency, and temporal continuity across the generated video.

Given a user prompt, the agent first parses the user's high-level intent. When the prompt is ambiguous, the agent interacts with the user to clarify key requirements, including the story theme, visual style, character settings, scene constraints, video duration, and narrative direction. When necessary, the agent can use external knowledge retrieval tools to supplement background information for the screenplay, characters, or scene details.

The agent then produces a global story description and presents it to the user for confirmation. After confirmation, the global description is decomposed into a shot-level plan that matches the conditioning format preferred by the video generation model. For each shot, the plan specifies visual, narrative, temporal, and continuity information, such as characters, actions, dialogue, and duration.

5.3 Memory Control and Generation

The project state P_t contains the screenplay, character cards, scene cards, shot plan, and revision records. The generated history H_t is maintained by the history manager and contains generated shots, key frames, shot descriptions, metadata, and version information. This conditioning design avoids directly using short user prompts as generator inputs. Instead, the system provides explicit, structured shot-level conditions.

At generation step t, given the current user description u_t, project state P_t, and generated history $H_t = \{S_i\}_{i=1}^{t-1}$, the agent retrieves historical shots related to the current shot and compiles the current shot condition:

$$R_t = \mathrm{TopK}_{S_i \in H_t} \mathrm{Rel}_\phi(q_t, S_i), \qquad c_t = \mathcal{T}_{\mathrm{LLM}}^{\mathrm{skill}}(u_t, P_t, R_t), \tag{16}$$

where q_t denotes the semantic query of the current shot, R_t denotes the retrieved reference shots, Rel_ϕ is a semantic relevance scoring function that measures the relevance between a historical shot S_i and the

current query q_t, and $\mathcal{T}_{\text{LLM}}^{\text{skill}}$ transforms the high-level description into a training-aligned shot-level condition c_t provided to the generator.

The agent uses a unified memory bank with both fixed and dynamic entries. Fixed memory entries follow the internal memory mechanism of JoyAI-Echo in Sec. 3 and remain compatible with the generator's native memory format. These entries are used to preserve low-level consistency, including identity, appearance, voice timbre, and face–voice correspondence. Fixed entries can be constructed from character cards, reference images, reference audio, or initialization shots, and remain relatively stable throughout the project.

Dynamic memory entries are selected by the agent according to semantic relevance. Given the screenplay, current shot plan, and generated history, the agent first identifies historical reference shots that are relevant to character continuity, scene dependency, costumes, props, action flow, or narrative context. For each selected reference shot, the system applies the Key-frame-of-Key-shot (KOK) strategy to obtain the concrete audio–visual memory entries used by the generator. In the current implementation, KOK follows the slot-paired memory construction in Sec. 3 and uses the energy-based audio-window selection in Eq. 8 to obtain synchronized audio-visual memory entries.

The selected synchronized audio–visual pair constitutes the dynamic memory entry for the selected reference shot. This design separates semantic-level memory selection from model-level memory usage. The agent determines which long-range references are needed by the current shot, while the generator still reads memory in its original format. In this way, the memory bank exposes useful historical information to the generator while reducing interference from irrelevant shots.

5.4 Critique and Revision

After each shot is generated, the system enters a critique stage that supports shot-level feedback and local refinement. In the current workflow, this stage is primarily human-in-the-loop: users can inspect each generated shot independently and provide localized revision instructions, such as modifying character appearance, action execution, dialogue, camera motion, scene layout, or audio–visual synchronization. In addition to human feedback, a critic model can be introduced to automatically assess the generated shot for consistency and quality issues. The agent parses the feedback or critic results, localizes the identified issues to the affected shot condition and associated memory entries, rewrites the condition when necessary, and then regenerates only the corresponding shot rather than regenerating the full video.

The revised shot is written back to the history manager together with its updated key frames, metadata, and revision record. If the modification affects temporal continuity or later narrative dependencies, the agent further updates the dynamic memory used by subsequent shots, ensuring that later generations retrieve the revised context. This forms a closed-loop workflow in which generation results are progressively refined through localized feedback.

6 Efficient Super-Resolution

Long-form audio-visual generation at native 720p delivers temporally coherent content but lacks the fine spatial detail expected in production. A joint audio-visual super-resolution (SR) framework is introduced that treats upsampling as *conditional generation*: given a low-resolution (LR) video latent and coarse audio latent produced by the JoyAI-Echo model, the SR model employs a single diffusion step to generate the corresponding high-resolution (HR) video and refined audio in a unified transformer process. The framework supports two resolution tiers—736×1280 → 1152×1920 (1K) and 736×1280 → 1472×2560 (2K)—sharing the same architecture, training, and distillation procedure.

6.1 Conditional Video Super-Resolution and Audio Refinement

Data. Training relies on a curated high-quality audio-visual corpus of ~876K samples (1080p–4K resolution, 5–17s duration). Candidate videos are filtered through a multi-stage quality pipeline: an image quality assessment (IQA) model scores every frame for sharpness, noise level, and compression artifacts, while an audio quality estimator evaluates signal-to-noise ratio, spectral clarity, and the absence of clipping or background

hum. Only samples that pass both visual and auditory quality thresholds are retained, ensuring that the training targets represent genuinely high-fidelity audio-visual content. The data mix deliberately emphasizes hard cases—speech-driven facial motion, on-screen text, small objects, fast motion, dense textures, and shot transitions—to improve robustness on the most challenging content categories.

Formulation. We formulate SR as a text-and-audio-visual-conditioned generation task. Given the text prompt c_{text}, the LR video latent z_v^{lr} and the coarse audio latent z_a^{c}, the SR model generates HR latent predictions $(\hat{z}_v^{\text{hr}}, \hat{z}_a^{\text{ref}})$ via a single transformer forward pass. The video and audio conditions are injected additively into the unified transformer: the video condition enters through a learned cross-resolution projection (details below), while the audio condition is projected via a learned linear layer. The same unified transformer processes video and audio tokens jointly, preserving cross-modal consistency through shared attention.

Cross-resolution condition injection. A key challenge for spatial SR is that the LR latent and HR latent reside on different spatial grids—for example, (23×40) vs. (36×60) at 1K, or (23×40) vs. (46×80) at 2K. We introduce *CondSRPatchifyProj*, a lightweight learned module that maps the raw 5D LR latent $z_v^{\text{lr}} \in \mathbb{R}^{B\times C\times T\times H_{\text{lr}}\times W_{\text{lr}}}$ directly into the HR token space $\mathbb{R}^{B\times(T\cdot H_{\text{hr}}\cdot W_{\text{hr}})\times D}$. It consists of three stages: (1) a per-frame residual convolutional block at LR spatial resolution for local feature refinement; (2) a learned linear spatial mapping $\text{Linear}(H_{\text{lr}}\cdot W_{\text{lr}},\ H_{\text{hr}}\cdot W_{\text{hr}})$ that maps every spatial position from LR grid to HR grid; and (3) a channel projection $\text{Linear}(C, D)$ to the transformer hidden dimension. The spatial projection is initialized as nearest-neighbor interpolation and the channel projection is near-zero, so the condition contribution is negligible at the start of training, allowing stable warm-up.

6.2 Audio-Visual One-Step Distillation

The multi-step SR model described above serves as the teacher for DMD training [24]. The DMD training objective is further combined with reconstruction and perceptual anchors to stabilize the SR training:

$$\mathcal{L} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{dmd}}\mathcal{L}_{\text{dmd}} + \lambda_{\text{lpips}}\mathcal{L}_{\text{lpips}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}. \tag{17}$$

Here, $\mathcal{L}_{\text{rec}}$ is the endpoint reconstruction loss in latent space, computed against the teacher-provided video and audio ODE endpoints, which preserves structure, color, and conditioning alignment. $\mathcal{L}_{\text{lpips}}$ is the LPIPS loss [28] which provides pixel-space perceptual supervision for sharper video local details, while $\mathcal{L}_{\text{dmd}}$ matches the student to the teacher distribution and compensates for the removed intermediate denoising steps. The regularization term $\mathcal{L}_{\text{reg}}$ includes auxiliary constraints such as timestep, noise-prediction, or temporal consistency regularization. Overall, reconstruction and LPIPS terms stabilize the DMD training objective, while DMD improves realism and high-frequency fidelity. This objective compresses the full denoising trajectory into a joint audio-visual single-step generator while preserving the refinement behavior of the multi-step teacher.

Efficient LoRA-based distillation. SR at 1K resolution increases the video token count by ~2.25× compared to the 720p stage, making memory consumption a major bottleneck. We apply Low-Rank Adaptation (LoRA) fine-tuning on frozen base weights and switch adapters between the teacher and student during DMD training, avoiding the need to maintain multiple full model copies in memory. Fully Sharded Data Parallel (FSDP) is used for distributed training, and only the student adapters are optimized while the teacher remains fixed.

Inference. JoyAI-Echo first generates a 720p video latent and its corresponding audio latent. They are then fed into the one-step SR generator, producing high-resolution refined video and refined audio in a single forward pass.

7 Evaluation

7.1 Experimental Setting

Dataset. The evaluation is conducted on a curated long-form audio-visual generation benchmark consisting of 100 stories and 3,000 shots in total. Each story contains 30 sequential shots, with each shot fixed to 241 frames at 25 fps. The benchmark is organized along two orthogonal axes: character source and visual style. Specifically, it includes 50 stories (1,500 shots) with specified IP characters and 50 stories (1,500 shots)

with original characters, as well as 30 stories (900 shots) in animation style and 70 stories (2,100 shots) in live-action style. Shot-level prompts provide detailed multimodal conditions, including character identity, dialogue, actions, emotions, camera motion, background, sound effects, and background music, enabling evaluation of both visual continuity and audio-visual consistency across long-form generation.

Metrics. Our evaluation pipeline covers five complementary dimensions: **(i) Cross-Shot Consistency.** (1) *ViCLIP Similarity:* We compute the pairwise cosine similarity of ViCLIP [22] video embeddings across all shot pairs to measure visual-semantic coherence. (2) *Self-CIDS:* We evaluate character identity consistency on a per-role basis across shots, using GroundingDINO [14] for character detection, Torchreid [29] for body embeddings, and FaceNet [20] for face embeddings. (3) *Voice Consistency:* We measure same-role speaker consistency across shots using cosine similarity between 3DSpeaker [3] speaker-verification embeddings. **(ii) Video Quality.** We adopt the slow-fast evaluation protocol from VBench-Long [11]. Specifically, the concatenated video is split into 2-second clips, on which we report aesthetic quality and imaging quality. **(iii) Text Consistency.** We evaluate the consistency between the generated content and the input story script, and report text-video alignment using the *CLIP score* [18].

(iv) Speech Content Accuracy. We use this metric to assess whether generated speech faithfully reproduces the scripted dialogue. We first transcribe the generated audio using Whisper ASR [19], and then compute the Word Error Rate (*WER*) against the ground-truth dialogue in the script. We report $speech\_accuracy = \max(0, 1 - \mathrm{WER})$, averaged over all dialogue shots. This metric is particularly important for distinguishing joint audio-visual generation models that produce semantically meaningful speech from video-to-audio cascade methods that may generate temporally synchronized but semantically meaningless audio.

Baselines. We compare against three categories of baselines.

(i) Joint Audio-Visual Generation. (1) *JavisDiT++* [13]: A joint audio-visual DiT model with hierarchical spatio-temporal prior synchronization and modality-specific MoE modules. (2) *LTX-2* [9]: A DiT-based audio-visual generation model with Gemma-based prompt enhancement, which generates synchronized audio and video in a single forward pass. (3) *LTX-2.3* [9]: An updated version of LTX-2 with improved audio quality and prompt adherence.

For these joint audio-visual baselines, each shot is generated independently from its corresponding structured text prompt, with a duration of approximately 10 seconds per shot. The generated shots are then concatenated into a single long video for evaluation. For LTX-2 and LTX-2.3, we enable the built-in Gemma prompt-enhancement module to process our detailed structured prompts. In addition, starting from the second shot, each LTX-series generation is conditioned on the last frame of the previous generated shot, so that local visual continuity can be propagated across adjacent shots.

(ii) Cascade Generation: Video Generation followed by Audio Post-processing. (1) *ShotStream* [16] + *MMAudio* [5]: ShotStream first generates silent videos, after which MMAudio performs video-to-audio synthesis for each shot. (2) *StoryMem* [27] + *MMAudio* [5]: StoryMem generates visually consistent silent videos with cross-shot memory, followed by MMAudio for audio generation.

For cascade baselines, we follow the same 30-shot setting. Each method first generates a 10-second silent video for each shot using the corresponding visual prompt. We then apply MMAudio to synthesize audio for each generated shot, using the shot-level audio descriptions when applicable. Finally, all audio-visual shots are concatenated into a single long video for evaluation. This protocol ensures that all baselines are evaluated under the same story script, shot duration, and final long-form video format.

(iii) Native Long Audio-Visual Generation. *Happy Oyster* is a recently released streaming generative world model built upon a native multimodal architecture. Unlike traditional one-shot video generation tools, it features a unified framework that synchronously generates visual and auditory signals. We utilize the *Directing mode* of Happy Oyster.[1] In this mode, the model takes a narrative script as input and continuously outputs a coherent, synchronized audio-visual sequence while maintaining persistent object placement, lighting, and scene causality throughout the generation process. Since Happy Oyster imposes a maximum continuous

[1]Details available at https://www.happyoyster.cn/docs

Table 1 User study based on Good–Similar–Bad (GSB) pairwise comparisons for long-form video and short-form video generation. The numbers denote the percentage of user preferences.

Aspect	Long-form Video			Short-form Video (Human-Centric)		
	JoyAI-Echo	Tie	Happy Oyster (Directing mode)	JoyAI-Echo	Tie	Wan 2.6
Visual aesthetics	**63.6%**	8.8%	27.6%	**58.8%**	14.7%	26.5%
Audio quality	**81.7%**	6.5%	11.8%	32.3%	30.9%	36.8%
Prompt following	**80.6%**	13.5%	5.9%	33.8%	36.8%	29.4%
IP consistency	**59.4%**	12.9%	27.7%	–	–	–

Table 2 Quantitative comparison of baseline methods. **Bold** and underline indicate the best and second-best results, respectively.

	Cross-Shot Consistency			Video Quality		Text Consistency	Speech
Method	ViCLIP	Self-CIDS	Voice	Aesthetic	Imaging	CLIP	Acc
JavisDiT++	0.6955	0.5621	0.7933	0.5019	0.6084	0.2522	0.0073
LTX-2	0.6718	0.5918	0.7339	0.5510	0.5242	0.2537	<u>0.8564</u>
LTX-2.3	0.7047	0.6135	0.6847	0.5436	0.4751	<u>0.2559</u>	0.8466
ShotStream+MMAudio	<u>0.7988</u>	0.7308	<u>0.7945</u>	0.5255	0.6044	0.1998	0.0059
StoryMem+MMAudio	0.7897	<u>0.7491</u>	0.7643	0.5265	0.6320	0.2368	0.0066
Happy Oyster (Directing)	0.7535	0.6940	0.7705	<u>0.5606</u>	<u>0.6701</u>	0.2544	0.0626
JoyAI-Echo	**0.8026**	**0.7793**	**0.8129**	**0.5679**	**0.7058**	**0.2658**	**0.8646**

generation limit of 3 minutes, we adapt our evaluation protocol for this baseline by providing the text scripts corresponding to the first 18 shots of each story.

7.2 Evaluation of Native JoyAI-Echo Model

User preference study. The user study follows a blind pairwise comparison protocol, referred to as Good-Similar-Bad (GSB) comparison. The order of the two videos in each pair is randomized, and each pair is evaluated by multiple annotators. For each evaluation aspect, annotators are asked to choose Good, Similar, or Bad. The final results are reported as percentages averaged over all annotated pairs. Table 1 reports the user study on GSB comparisons for both long-form video and short-form video generation. For long-form video generation, JoyAI-Echo is compared with Happy Oyster in terms of visual aesthetics, audio quality, IP consistency, and prompt following; for short-form video generation, JoyAI-Echo is compared with Wan 2.6 under visual aesthetics, audio quality, and prompt following. As shown in Table 1, JoyAI-Echo obtains higher user preference, demonstrating that the proposed framework improves both perceptual quality and audio-visual faithfulness in human evaluation.

Qualitative comparison. Fig. 4 provides qualitative comparisons between JoyAI-Echo and Happy Oyster across temporally separated shots from the same stories. Across diverse scenarios, including a kitchen scene, a warehouse confrontation, a wizard exploring a castle, and a Batman patrol sequence, JoyAI-Echo better preserves character appearance, scene layout, background details, and action continuity over long rollouts. In contrast, Happy Oyster exhibits more noticeable temporal drift, including changes in character identity, loss or alteration of important objects, and weaker consistency of scene composition across distant shots. For example, in the first story, JoyAI-Echo correctly preserves the female character when she is required to appear beside the male character. In the second story, it better matches the specified character action, such as the female aiming the gun forward. In the third story, it maintains the requested castle-corridor setting, while in the fourth story, it better follows the intended action narrative of Batman chasing a moving van. These examples show that JoyAI-Echo is more reliable in preserving key characters, actions, scenes, and story events specified by the prompt over long-range generation.

Figure 4 Qualitative comparison between **JoyAI-Echo** and **Happy Oyster**. The visualization presents generated frames from several temporally separated shots within the same long-form stories. Red text highlights prompt-critical instructions where the baseline often fails to preserve the requested object, action, spatial relation, or scene detail, while JoyAI-Echo follows these instructions more faithfully.

Table 2 shows that JoyAI-Echo achieves the best performance across all evaluated dimensions. For cross-shot consistency, JoyAI-Echo obtains the highest ViCLIP similarity, Self-CIDS, and voice consistency scores, reaching 0.8026, 0.7793, and 0.8129, respectively. Compared with the strongest baseline in each metric, JoyAI-Echo improves Self-CIDS by 0.0302 and voice consistency by 0.0184, indicating better preservation of character identity and speaker identity across long-horizon multi-shot generation. Although cascade methods such as ShotStream+MMAudio and StoryMem+MMAudio achieve competitive visual consistency, their speech accuracy remains extremely low, since the post-hoc audio generation stage does not reliably reproduce the scripted dialogue.

JoyAI-Echo also achieves the best video-quality scores, with 0.5679 on aesthetic quality and 0.7058 on imaging quality. This suggests that the proposed long-form video generation and refinement pipeline improves local visual fidelity while maintaining global temporal coherence. For text consistency, JoyAI-Echo obtains the highest CLIP score of 0.2658, outperforming LTX-2.3, Happy Oyster, and other baselines. In speech content accuracy, JoyAI-Echo reaches 0.8646, slightly surpassing LTX-2 while substantially outperforming cascade video-to-audio baselines and Happy Oyster. These results show that JoyAI-Echo does not merely generate temporally synchronized audio, but can also preserve the semantic content of scripted dialogue.

7.3 Director Agent and Super-Resolution Extensions

Based on the proposed Director Agent and super-resolution framework, extensive experiments demonstrate improvements in both semantic coherence and generation quality. The Director Agent effectively bridges the gap between ambiguous user prompts and structured generator inputs, reducing error accumulation across shots while enabling flexible local revisions through its memory and critic mechanisms. Meanwhile, the one-step distillation-based super-resolution module achieves efficient joint audio-visual refinement, producing 1.5× higher spatial resolution with competitive fidelity and temporal consistency compared to multi-step baselines. Together, these components enable JoyAI-Echo to generate long-form, high-resolution videos with stable character consistency, natural shot transitions, and real-time inference capability, validating the effectiveness of the hierarchical planning-and-refinement pipeline.

8 Conclusion

We presented JoyAI-Echo, a framework that overcomes the key barriers of long video generation. A cross-modal audio-visual memory bank preserves identity consistency across five-minute videos, while a post-training pipeline consisting of Supervised Fine-Tuning (SFT), Reinforcement Learning with Human Feedback (RLHF), and Distribution Matching Distillation (DMD) improves generation quality and inference efficiency, enabling JoyAI-Echo to outperform Happy Oyster and surpass Wan 2.6 on human-centric tasks. An interactive agent and a lightweight super-resolution module further deliver real-time conversational editing and high-definition streaming output. To our knowledge, JoyAI-Echo is the first to simultaneously achieve long-range cross-modal consistency, real-time minute-level generation, conversational interactivity, and high-resolution output without compromise.

References

[1] Brandon Castellano. PySceneDetect: Video scene cut detection and analysis tool. https://github.com/Breakthrough/PySceneDetect, 2024.

[2] Shuo Chen, Cong Wei, Sun Sun, Ping Nie, Kai Zhou, Ge Zhang, Ming-Hsuan Yang, and Wenhu Chen. Context forcing: Consistent autoregressive video generation with long context, 2026. URL https://arxiv.org/abs/2602.06028.

[3] Yafeng Chen, Siqi Zheng, Hui Wang, Luyao Cheng, et al. 3d-speaker-toolkit: An open source toolkit for multi-modal speaker verification and diarization. 2025.

[4] Zhiqi Chen et al. SpeakerVid-5M: A large-scale speaker-identity video dataset for talking head generation. arXiv preprint, 2024.

[5] Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. Mmaudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 28901–28911, 2025.

[6] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD), pages 226–231, 1996.

[7] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In Scandinavian Conference on Image Analysis (SCIA), pages 363–370. Springer, 2003.

[8] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. Ltx-video: Realtime video latent diffusion, 2025. URL https://arxiv.org/abs/2501.00103.

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

[10] Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion, 2025. URL https://arxiv.org/abs/2506.08009.

[11] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 21807–21818, 2024.

[12] Hui Li, Mingwang Zhang, Yun Bian, Di Zang, et al. OpenHumanVid: A large-scale high-quality dataset for enhancing human-centric video generation. arXiv preprint arXiv:2412.00115, 2024.

[13] 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. In The Fourteenth International Conference on Learning Representations.

[14] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. 2023.

[15] Yiting Lu, Xin Li, Yajing Pei, Kun Yuan, Qizhi Xie, Yunpeng Qu, Ming Sun, Chao Zhou, and Zhibo Chen. Kvq: Kwai video quality assessment for short-form videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024.

[16] Yawen Luo, Xiaoyu Shi, Junhao Zhuang, Yutian Chen, Quande Liu, Xintao Wang, Pengfei Wan, and Tianfan Xue. Shotstream: Streaming multi-shot video generation for interactive storytelling. arXiv preprint arXiv:2603.25746, 2026.

[17] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research, 2024.

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PmLR, 2021.

[19] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL https://arxiv.org/abs/2212.04356.

[20] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 815–823, 2015.

[21] 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, 2026. URL https://arxiv.org/abs/2603.11647.

[22] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In The Twelfth International Conference on Learning Representations.

[23] Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. arXiv preprint arXiv:2505.07818, 2025.

[24] 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, pages 6613–6623, 2024. URL https://arxiv.org/abs/2311.18828.

[25] Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Frédo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 22963–22974. IEEE, june 2025. URL https://arxiv.org/abs/2412.07772.

[26] Guohui Zhang, XiaoXiao Ma, Jie Huang, Hang Xu, Hu Yu, Siming Fu, Yuming Li, Zeyue Xue, Lin Song, Haoyang Huang, Nan Duan, and Feng Zhao. Omninft: Modality-wise omni diffusion reinforcement for joint audio-video generation. arXiv preprint arXiv:2605.12480, 2026.

[27] 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.

[28] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 586–595, 2018.

[29] Kaiyang Zhou and Tao Xiang. Torchreid: A library for deep learning person re-identification in pytorch. arXiv preprint arXiv:1910.10093, 2019.