



JoyAI-Echo-1.5: Long-Horizon Audio-Visual Generation for Persistent Stories and Interactive Worlds

Echo Team @ Joy Future Academy, JD

Abstract

Video generation is progressing beyond isolated clips toward long-form narratives and interactive worlds, requiring models to preserve identities, follow user controls, and remain stable over extended rollouts. We present **JoyAI-Echo-1.5**, a unified audio-visual generation system with two purpose-built variants. The **long-video** variant introduces composable cross-shot memory that aggregates visual evidence across multiple prior shots and speaker cues derived from speech-filtered full-shot audio, enabling persistent character appearance and voice identity across flexible combinations of text, image, and memory conditioning. The **world model** variant converts heterogeneous navigation inputs into calibrated metric 6-DoF camera trajectories and injects them through a geometry-aware conditioning pathway, enabling controller-agnostic interaction across flexible viewpoints. To support efficient long-horizon generation, we transform a bidirectional audio-visual backbone into a causal few-step generator using progressive teacher forcing and short- and long-horizon Self-Gradient Forcing on self-generated rollouts. Experiments demonstrate strong performance in both settings. JoyAI-Echo-1.5 achieves improvements over existing long-video baselines in cross-shot consistency, visual quality, text alignment, and speech fidelity. Its world-model variant ranks **first** on WBench, with an average score of 81.7, and achieves leading visual quality and long-horizon persistence on SANA-WM-Bench. Together, these results indicate that memory, geometric control, and rollout-aware training provide a practical foundation for generating coherent stories and continuously evolving interactive worlds.

Project Page: <https://echo-team-joy-future-academy-jd.github.io/Echo-1.5-Page/>

Code: <https://github.com/jd-opensource/JoyAI-Echo>

1 Introduction

The next frontier in video generation is not merely improving individual clips, but achieving **continuity and control**. Long-form creation requires characters and voices to remain consistent across shots, while interactive world modeling requires environments to respond precisely to user intent and remain stable during continuous rollouts. Yet existing video models are still largely designed around fixed temporal windows: they forget identities, drift when conditioned on self-generated histories, and struggle to translate heterogeneous action inputs into consistent motion. Overcoming these limitations requires more than scaling model capacity; it requires rethinking how models remember, act, and evolve over time.

We present **JoyAI-Echo-1.5**, a technical system comprising two purpose-built versions. Its **long-video version** targets long-form audio-visual generation and introduces composable cross-shot memory to preserve character appearance and speaker identity across changing scenes. It aggregates visual evidence from multiple prior shots, derives robust voice cues from speech-filtered full-shot audio, and supports flexible combinations of text,

image, and memory conditions. Its **world-model version** targets interactive world modeling by converting heterogeneous navigation inputs into calibrated metric 6-DoF camera trajectories and injecting them through a geometry-aware control pathway. The first extends stories beyond individual shots; the second transforms generated video into an environment that can be explored.

Both versions are designed to operate beyond fixed temporal windows. JoyAI-Echo-1.5 progressively converts a bidirectional audio-visual backbone into a causal few-step generator through audio-visual teacher forcing followed by short- and long-horizon Self-Gradient Forcing on self-generated rollouts. Training on model-generated histories exposes the system to accumulated generation errors, helping it preserve synchronized audio-visual dynamics and sustain long streaming sequences within a bounded computational budget. **The long-video version gives generation memory. The world-model version gives it control. Together, JoyAI-Echo-1.5 pushes video generation from creating isolated moments toward sustaining stories and evolving worlds.**

Main contributions. Our main contributions are summarized as follows:

- **A long-video system that remembers.** We introduce multi-shot audio-visual memory with speech-filtered voice representations, separated positional regions, and composable text, image, and memory conditioning to preserve identity across shots.
- **A world model that responds.** We establish a controller-agnostic action interface based on calibrated metric 6-DoF trajectories, enabling precise camera control across heterogeneous data sources, environments, and viewpoints.
- **A generator built to continue.** We develop a causal few-step training pipeline that combines audio-visual teacher forcing with short- and long-horizon Self-Gradient Forcing, allowing JoyAI-Echo-1.5 to learn from its own rollout distribution and remain stable over extended generation.

2 Data

2.1 Long-Video Training Data

JoyAI-Echo-1.5 adopts a two-stage data curriculum, as illustrated in Fig. 1. **Stage I** reuses the identity-centric corpus of JoyAI-Echo-1.0 [17] and constructs memory-target pairs from scene-disjoint clips of the same recurring character to train audio-visual memory conditioning. **Stage II** uses high-quality videos with broader resolution and aspect-ratio coverage, increases the proportion of subtitle-free and Chinese-language samples, and incorporates captioned single-shot and multi-shot videos, with the latter explicitly describing transitions across cuts. It improves general audio-visual quality, spatial and linguistic coverage, and transition modeling, while preserving text-to-audio-video (T2AV) and image-to-audio-video (I2AV) generation capabilities. Both stages use the same quality-control pipeline, with their sampling distributions and caption formats adapted to their respective objectives.

2.1.1 Stage-I Memory-Construction Corpus

Identity-centric source groups. The memory-construction corpus retains the identity-centric organization introduced in JoyAI-Echo-1.0 [17]. Starting from long-form films, television episodes, and web videos, recurring characters are associated with scene-disjoint single-shot clips. Following the extraction pipeline of v1.0, the data undergo global identity discovery, scene grouping, local tracklet assignment, and quality filtering. Consequently, the basic unit used by v1.5 is not an isolated video clip, but an identity group

$$\mathcal{G}_k = \{S_{k,n}\}_{n=1}^{N_k}, \quad S_{k,n} = (V_{k,n}, A_{k,n}), \quad (1)$$

where all clips in $\mathcal{G}_k$ contain the same recurring character while differing in scene, clothing, viewpoint, pose, motion, illumination, facial expression, dialogue, and acoustic environment. Maintaining this diversity is important because a useful memory representation should capture persistent character-level evidence rather than scene-specific appearance alone.

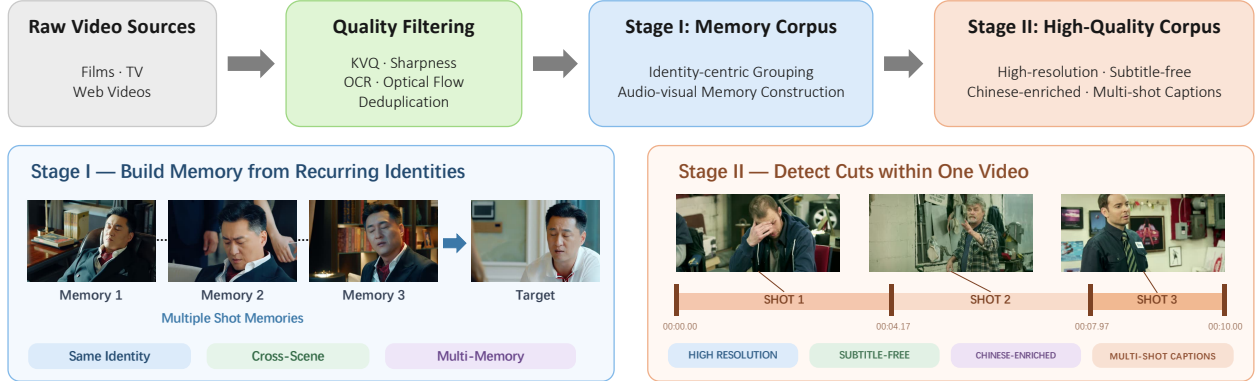


Figure 1 Two-stage data construction pipeline of JoyAI-Echo-1.5. Raw videos are quality-filtered and organized into an identity-centric Memory corpus and a high-quality corpus with broader resolution, language, and multi-shot coverage.

Memory-target sample construction. At each training iteration, one clip $S_{k,t}$ is selected as the prediction target and a variable-size subset of the remaining clips is sampled as memory:

$$\mathcal{M}_{k,t} = \{S_{k,i_1}, S_{k,i_2}, \dots, S_{k,i_K}\}, \quad i_j \neq t. \quad (2)$$

The memory and target clips therefore show the same identity through different observations rather than through temporally adjacent frames from the same trajectory. This separation discourages direct content copying and requires the model to recover stable identity information under changes in background, camera pose, wardrobe, action, and speech content. The number of sampled memory shots K is varied during training so that the model does not rely on a fixed memory capacity and remains effective when only a small amount of character evidence is available.

For every selected memory shot, one video frame is randomly sampled and encoded as visual memory. In contrast to the visual branch, which uses a compact frame-level observation, the complete shot audio is retained. The audio is first processed by the speech-filtering operator and is then encoded as audio memory. Using the complete speech-bearing interval provides a more reliable description of speaker characteristics than selecting a short response window, while speech filtering reduces the leakage of background music and environmental effects into the speaker condition. The visual and audio observations from each memory shot preserve the same slot order, allowing the model to jointly access the appearance and voice evidence associated with that shot.

Conditioning configurations. The availability of memory and the first-frame image condition is independently varied when constructing training samples. This produces four conditioning configurations under a unified interface:

- **T2AV**: text-to-audio-video generation without memory or an input image;
- **I2AV**: audio-video generation conditioned on text and the first frame;
- **MT2AV**: text-to-audio-video generation conditioned on audio-visual memory;
- **MTI2AV**: audio-video generation jointly conditioned on memory, text, and the first frame.

The memory-construction corpus primarily supplies MT2AV and MTI2AV examples, while the same sample representation also supports T2AV and I2AV when the corresponding conditions are absent. Randomizing both memory length and image-condition availability prevents the model from binding a particular generation task to a fixed input layout. It also enables memory and the first-frame condition to be used compositionally, rather than treating memory-conditioned generation and image-conditioned generation as separate models.

2.1.2 Stage-II High-Quality and Multi-Shot Enhancement Corpus

The identity-organized corpus provides the supervision required for cross-shot memory retrieval, but its distribution is constrained by the availability of recurring characters and scene-disjoint observations. Training exclusively on such samples can reduce the diversity and standalone quality of ordinary T2AV and I2AV generation. Stage II therefore uses an enhancement corpus selected independently of the identity-group requirement. It combines high-quality single-shot videos with samples containing multiple shots and explicit transitions. Its purpose is to preserve and strengthen general audio-visual generation, broaden spatial and linguistic coverage, and improve transition modeling after the memory interface has been established in Stage I.

Resolution and aspect-ratio coverage. The high-quality corpus contains videos at 720×720 , 1280×720 , and 720×1280 , corresponding to square (1:1), landscape (16:9), and portrait (9:16) formats. Rather than concentrating training on one canonical spatial shape, we explicitly balance these formats so that the model encounters comparable high-quality examples across common generation settings. This improves robustness to changes in token layout and spatial composition and allows the same model to serve cinematic landscape generation, portrait-oriented content, and square social-media formats.

Subtitle-free visual data. The retained samples emphasize clear visual content and stable audio-visual quality, with particular attention to videos without embedded subtitles. Burned-in subtitles create persistent, high-contrast patterns that are strongly correlated with dialogue. When such patterns are overrepresented, the model may reproduce incomplete glyphs, garbled character fragments, or subtitle-like regions even when no text is requested. Increasing the proportion of subtitle-free video weakens this undesirable correlation and provides cleaner supervision for faces, body motion, lower-frame image regions, and speaking scenes. OCR metadata is still retained for filtering and analysis, but samples with dominant or repeatedly occluding text are excluded from the enhancement corpus.

Chinese-language expansion. We additionally increase the proportion of Chinese-language data, including Chinese captions and dialogue-oriented audio-visual content. The broader language distribution supplies direct supervision for Chinese prompt following, spoken-content generation, and the alignment between visible speaking motion and Chinese dialogue. This adjustment is particularly important for long-form generation, where errors in linguistic conditioning or speaker behavior can accumulate across shots. The language expansion is performed together with visual-quality and subtitle filtering so that improved Chinese coverage does not come at the cost of image cleanliness or acoustic fidelity.

2.1.3 Quality Filtering and Distribution Balancing

Both corpora inherit the multi-axis quality-control pipeline of JoyAI-Echo-1.0, but the retained distribution is recalibrated for high-resolution and multi-shot training.

Learned perceptual quality. We first apply a learned video-quality estimator [23] to assess perceptual quality over both spatial content and temporal behavior. It removes samples with severe compression artifacts, abnormal color, unstable exposure, visually distracting noise, or consistently low aesthetic quality. This learned score complements the following hand-crafted measurements: it captures degradations that are difficult to describe with one low-level statistic, whereas the explicit operators provide interpretable control over identity visibility, text overlays, and motion coverage.

Structural sharpness. Uniformly sampled frames are evaluated using the variance of the Laplacian. For a sampled frame I_t , its sharpness response is

$$s_t^{\text{sharp}} = \text{Var}(\nabla^2 I_t), \quad (3)$$

and the frame responses are robustly aggregated into a clip-level score

$$S_{\text{sharp}}(V) = \text{Agg}\left(\{s_t^{\text{sharp}}\}_{t \in \mathcal{T}}\right), \quad (4)$$

where $\mathcal{T}$ denotes the uniformly sampled timestamps. Unlike face-detection confidence alone, this operator directly measures whether structural details remain visible. It is therefore particularly important for memory training: a face may still be detected under defocus, motion blur, or low spatial resolution, while providing unreliable identity supervision. Clips with persistently weak responses or insufficiently clear character observations are removed.

Text and overlay filtering. OCR and overlay detectors identify subtitles, logos, watermarks, large text regions, and burned-in graphics. We consider not only detected text content but also its occupied area, temporal persistence, and overlap with the main character. Clips in which text dominates the image or repeatedly occludes faces and bodies are rejected. For the high-quality enhancement corpus, this filter is made deliberately stricter to increase the proportion of subtitle-free data. OCR metadata is nevertheless preserved for caption construction: an empty result is explicitly verbalized as no on-screen text or subtitles, whereas visible text is introduced with an OCR marker.

Motion statistics. Dense optical flow [7] is used to estimate whole-frame dynamics. Given the flow field $\mathbf{u}_t(p)$ between adjacent frames, we compute

$$M(V) = \frac{1}{T-1} \sum_{t=1}^{T-1} \frac{1}{|\Omega|} \sum_{p \in \Omega} \|\mathbf{u}_t(p)\|_2, \quad (5)$$

where Ω is the image domain. The motion score is not used to retain only high-motion samples; instead, it supports distribution control. Very low-motion clips are downsampled to prevent the corpus from being dominated by static talking heads, while sharpness and perceptual quality filters prevent large flow caused by blur, camera shake, or corrupted frames from being mistaken for useful motion. The retained data therefore covers static dialogue, subtle facial and body motion, camera movement, and larger character displacement without collapsing toward either motion extreme.

Identity-level coverage and de-duplication. For the memory-construction corpus, all of the above filters are applied at both clip and identity-group levels. If an identity no longer contains enough scene-disjoint high-quality clips after filtering, the complete group is removed because it cannot support a meaningful memory-target split. Within each identity, perceptual hashes, color histograms, and DINOv2 background similarity are used to suppress near-duplicate observations. This ensures that adding memory slots introduces new evidence about appearance, viewpoint, scene, or voice rather than repeated copies of the same shot.

Distribution balancing. Finally, the retained samples are balanced across square, landscape, and portrait formats; Chinese and English content; subtitle presence; and motion ranges. The identity-organized and high-quality corpora are also mixed so that memory-conditioned examples do not overwhelm ordinary T2AV and I2AV training. These controls prevent one source domain or generation mode from dominating optimization and allow improvements in composable memory conditioning to be obtained without sacrificing general visual and acoustic fidelity.

2.1.4 Single- and Multi-Shot Audio-Visual Captioning

JoyAI-Echo-1.5 expands the clip-level audio-visual annotation of v1.0 into bilingual captions covering both single-shot and multi-shot videos. For each structure, we construct a text-to-video caption and derive a corresponding image-to-video caption. The T2V caption must independently specify the visible subjects and their temporal behavior, whereas the I2V caption can rely on the input first frame for appearance information and is therefore rewritten to reduce redundant identity descriptions.

Single-shot T2V captions. For a single-shot clip, structured fields are concatenated in the following order:

$$c_{\text{single}}^{\text{T2V}} = R \oplus A \oplus S \oplus C \oplus B \oplus O \oplus E \oplus G, \quad (6)$$

where R is `Roles_and_Subjects`, A is `Action_and_Dialogue`, S is `Style`, C is `Camera_Movement`, B is `Background`, O is the subtitle/OCR description, E is `Sound_Effects`, and G is `BGM`. The action field describes actions, expressions, body movement, and dialogue in temporal order, while voice timbre and speaking manner are retained in the subject and dialogue descriptions. When no OCR or subtitle is detected, O explicitly states that no on-screen text or subtitles are present; otherwise, the recognized content is introduced by an OCR marker. Repeated empty-text statements are removed during concatenation.

Multi-shot T2V captions. For a video containing N shots, we use the rewritten identity-consistency annotation and serialize the individual shot records in chronological order:

$$c_{\text{multi}}^{\text{T2V}} = H_N \oplus \bigoplus_{s=1}^N [Q_s \oplus R_s \oplus A_s \oplus C_s \oplus O_s \oplus E_s \oplus G_s], \quad (7)$$

where H_N states the number of shots and Q_s is the explicit `Shot` s boundary. Each shot preserves its own subjects, actions and dialogue, camera behavior, OCR state, sound effects, and background music. This format provides direct supervision for shot order, subject changes, dialogue allocation, camera changes, and soundtrack continuity across cuts. Characters recurring across shots retain consistent identifiers in the T2V annotation, while characters appearing for the first time are fully introduced in the shot where they enter.

I2V caption rewriting. Both single-shot and multi-shot T2V captions are converted into I2V captions. Because the first frame already provides the initial appearance, the full identifier-based subject description in the first shot is replaced by a concise semi-referential phrase, such as “the short-haired man” or “the uniformed middle-aged man.” The synthetic identifiers such as `ID_A` are removed. Voice timbre, speaking manner, dialogue, sound effects, and background music are retained because they cannot be recovered reliably from the image. For multi-shot videos, any new character appearing after the first shot receives a sufficiently detailed natural-language description at first appearance, but is likewise described without an identifier token. The complete caption is then lightly shortened to remove repeated visual facts without discarding temporal actions or audio information.

A bilingual I2V instruction prefix is randomly sampled and prepended to the rewritten caption. The same procedure is applied independently to the Chinese and English annotations, and the resulting strings are stored in `video_audio_i2v_caption`. This design reduces conflict between the image and redundant textual appearance descriptions while preserving the information required to generate motion, dialogue, voice, sound, and multi-shot evolution.

The two-stage curriculum separates the acquisition of memory conditioning from the subsequent enhancement of general quality and transition modeling. Stage I teaches the model to retrieve and compose cross-shot identity evidence. Stage II uses high-quality subtitle-free data with increased Chinese coverage to maintain standalone T2AV and I2AV fidelity; within the same stage, captioned single-shot samples preserve fine-grained intra-shot actions and multi-shot samples provide explicit supervision for transitions across cuts. This progression allows JoyAI-Echo-1.5 to extend long-form controllability and composable memory conditioning while maintaining broad spatial, linguistic, visual, and acoustic coverage.

2.2 Action Annotation and Trajectory Data

Action-conditioned world generation requires a consistent motion representation across data collected from different sources. However, these sources provide substantially different levels of supervision: synthetic UE renders expose exact camera poses and control inputs, internal gameplay provides action logs but not camera poses, while human gameplay and web videos provide audio-visual observations but no explicit action logs or camera telemetry. We therefore build a unified processing pipeline that converts all sources into metric 6-DoF camera trajectories and, whenever available, aligns them with source-native control signals.

Trajectory Data from Heterogeneous Sources. We collect motion supervision from three complementary data sources:

Table 1 Representative abbreviated examples of the four caption formats used in JoyAI-Echo-1.5. English renderings are shown for readability, while both Chinese and English captions are retained during training. T2V captions preserve explicit identity definitions, whereas I2V captions use concise natural references because appearance is supplied by the first frame.

Structure	T2V caption	I2V caption
Single shot	ID_A is a young man with short brown hair, wearing a dark gray hoodie, with a clear mid-range voice. He sits slightly off-camera, blinks naturally, and says, “my games often focus on that ...” A static chest-up close-up uses shallow depth of field in an indoor gym. No on-screen text or subtitles. No obvious sound effects. No background music.	[Random I2V prefix] A short-haired man speaks with a clear, stable mid-range voice and a natural conversational tone. He sits slightly off-camera, blinks naturally, and says, “my games often focus on that ...” A static chest-up close-up keeps him in focus against a softly blurred gym. No on-screen text or subtitles. No obvious sound effects. No background music.
Multi-shot	This video contains two shots. Shot 1: ID_A is a middle-aged man in a dark green dress uniform with a deep voice; ID_B is a middle-aged man in digital camouflage with a bright voice. They walk and talk side by side as a medium shot follows them. OCR: military slogans and a name label. Soft piano music plays. Shot 2: ID_C is a silent middle-aged man in a dark blue police uniform. A police formation advances in a static medium shot. No OCR or subtitles; the piano music continues.	[Random I2V prefix] This video contains two shots. Shot 1: A middle-aged man in a dress uniform and a middle-aged man in camouflage walk and talk side by side. The former has a deep, steady voice and the latter a bright, natural voice. A medium shot follows them; soft piano music plays. Shot 2: A serious middle-aged man in a dark blue police uniform appears silently within an advancing formation. A static medium shot follows the group. No OCR or subtitles; the piano music continues.

- *Unreal Engine (UE) Renders.* UE environments provide accurate geometric supervision. Avatars are controlled through local navigation commands under the engine’s native physics and collision system. Each rollout is rendered from one first-person and four character-relative third-person cameras. For every frame, we synchronously record RGB observations, keyboard inputs, camera intrinsics, and ground-truth camera-to-world poses in metric units. During training, one viewpoint is randomly sampled from each sequence, encouraging the model to generalize across camera configurations.
- *Internal Gameplay.* Internal game environments provide diverse interaction scenarios, including open-world navigation, driving, and mounted traversal in both first- and third-person views. Scripted agents execute predefined control sequences, allowing us to record accurately aligned action commands together with RGB video, stereo audio, and scene metadata. Interface elements such as HUDs and crosshairs are removed during rendering. Since these game engines do not expose camera pose matrices, their 6-DoF trajectories are estimated from the recorded videos and temporally aligned with the corresponding action logs.
- *Human Gameplay and Web Video.* These videos complement scripted data with natural navigation behavior, human reaction patterns, and richer audio-visual content. They do not contain explicit action logs or camera telemetry, so we use the recovered camera trajectory itself as motion supervision. We extract temporally continuous, shot-consistent segments of approximately one minute and estimate their metric 6-DoF camera motion using geometry-based reconstruction.

Together, these sources provide a spectrum from precise synthetic supervision to large-scale natural motion, while sharing the same trajectory representation.

Long-Sequence Metric Pose Recovery. For videos without ground-truth camera poses, directly estimating motion on short training clips often leads to scale ambiguity and unstable trajectories. We instead estimate camera motion over longer, continuous windows of approximately one minute and only afterward divide them into training clips.

Specifically, we apply the metric-depth-assisted ViPE pipeline [12] to temporally subsampled frames. The longer temporal context provides a larger geometric baseline and improves the robustness of pose reconstruction. We then recover frame-rate trajectories by interpolating rotations with Spherical Linear Interpolation (SLERP) and translations with linear interpolation. Finally, both estimated and ground-truth trajectories are transformed into a common camera-to-world coordinate convention, metric scale, axis orientation, and temporal sampling

Unified Audio-visual Memory Conditioning

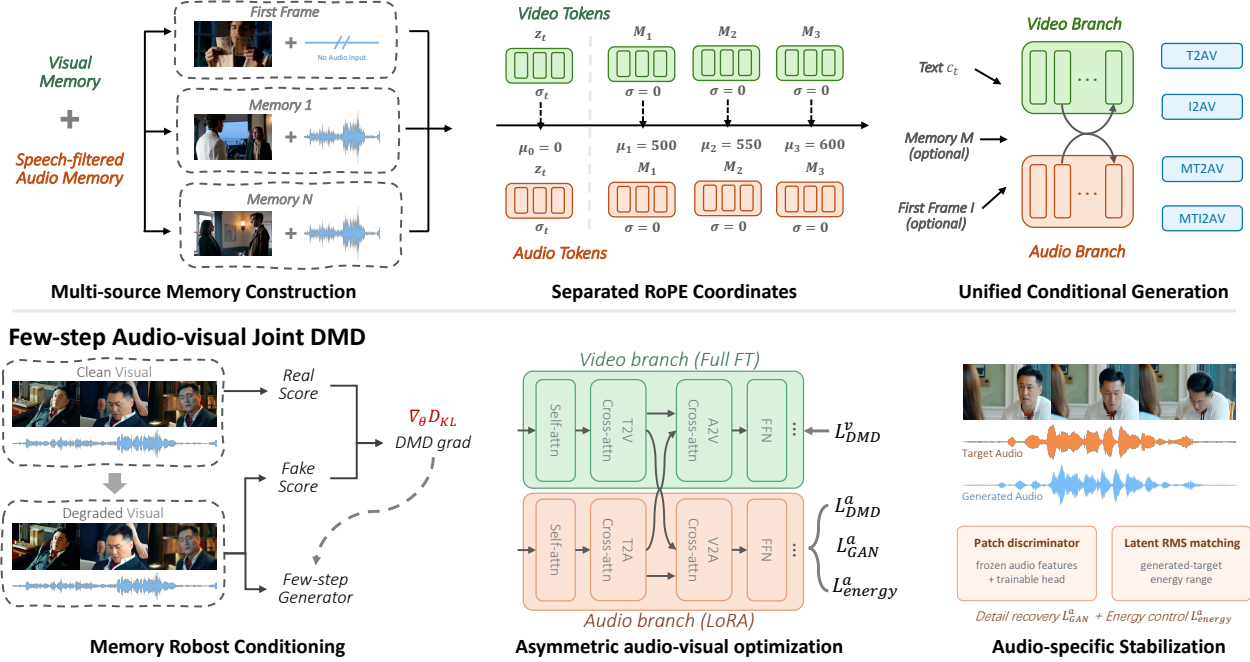


Figure 2 Overview of memory-conditioned long-form generation and few-step acceleration in JoyAI-Echo-1.5. Top: unified conditioning with single-frame visual memories, speech-filtered audio memories, and separated RoPE coordinates supports multiple audio-visual generation tasks. Bottom: memory-robust joint DMD combines asymmetric video-audio optimization with audio-specific adversarial and energy constraints.

rate.

Trajectory Quality Filtering. We apply trajectory-level filtering before training. Clips are removed when pose estimation has low confidence or exhibits implausible motion, including severe frame-to-frame jitter, abrupt rotational changes, negligible displacement, or unstable camera intrinsics. Ground-truth UE trajectories are subjected to the same motion-range checks, while their original metric poses are kept unchanged.

Action-Decoupled Text Annotation. We structure each audio-visual caption into persistent scene attributes, a dynamic-motion narrative, and audio-related descriptions. Persistent attributes include the scene, visual style, viewpoint, and subject identity, while the audio component describes speech and environmental sounds. During Action-SFT, we omit both the dynamic-motion narrative and the audio-related fields, retaining only the persistent static descriptors because this stage uses a visual-only denoising objective. During Joint-FT, we restore the audio-related fields while continuing to exclude the dynamic-motion narrative. In both stages, camera motion is specified explicitly by the trajectory condition rather than redundantly through language. This encourages the model to follow the provided trajectory while preserving the remaining prompt semantics.

3 Long-Horizon Video Generation

JoyAI-Echo-1.0 [17] introduced a long-term audio-visual memory mechanism that reuses compact visual and acoustic cues from previous shots, substantially improving character appearance and speaker-timbre consistency during multi-shot generation. Building on this foundation, JoyAI-Echo-1.5 improves both long-range consistency and generation efficiency. First, it introduces a unified audio-visual memory-conditioning framework that constructs visual memory from multiple shots, extracts audio memory from speech-filtered full-shot audio, separates memory and target tokens through RoPE coordinates, and jointly trains different combinations of memory and image conditions. Second, it applies distribution matching distillation to convert the original multi-step generator into an 8-step student, with optimization strategies designed for the distinct

characteristics of video and audio generation. We first formulate the long-horizon generation task and present the unified memory-conditioning framework, and then describe the proposed distillation method for efficient audio-visual generation.

3.1 Unified Audio-Visual Memory Conditioning

Task formulation. Given a story script represented by shot-level text conditions $C = \{c_t\}_{t=1}^T$, JoyAI-Echo generates a sequence of synchronized audio-visual shots $S = \{S_t\}_{t=1}^T$, where $S_t = (V_t, A_t)$. Let M_0 denote an optional memory bank supplied before generation, and let $I = \{I_t\}_{t=1}^T$ denote optional image conditions for individual shots. Long-form generation is factorized as:

$$p_\theta(S \mid C, M_0, I) = \prod_{t=1}^T p_\theta(S_t \mid c_t, M_{t-1}, I_t), \quad M_t = \mathcal{U}(M_{t-1}, S_t). \quad (8)$$

Before generating shot t , M_{t-1} contains both the initial memory M_0 and the memory extracted from shots $S_1, \dots, S_{t-1}$. After S_t is generated, the update function $\mathcal{U}$ extracts its audio-visual information and updates the memory bank for the next shot. The image condition I_t , when provided, specifies the first frame of the current target shot and does not consume a memory slot.

Audio-visual memory construction. Each memory item is written as $m_i = (m_i^v, m_i^a)$, where m_i^v and m_i^a denote its visual and audio components. From the available history, we select up to K shots and arrange them in chronological order. For the i -th selected shot, one frame is sampled and encoded by the video VAE:

$$m_i^v = E_v(V_i[f_i]), \quad f_i \sim \mathcal{U}\{1, \dots, |V_i|\}. \quad (9)$$

The resulting latents provide compact visual observations from different historical shots instead of restricting the condition to a single adjacent frame.

For audio memory, we use the complete audio content of each selected shot. Since full-shot audio may contain background music, environmental sounds, and other non-speech components, we first apply a speech-filtering operator $\mathcal{F}_{\text{speech}}$ and then encode the filtered waveform using the audio VAE:

$$\tilde{A}_i = \mathcal{F}_{\text{speech}}(A_i), \quad m_i^a = E_a(\tilde{A}_i). \quad (10)$$

Here, $\mathcal{F}_{\text{speech}}$ is a generic speech-separation operator that can be instantiated by different filtering models; our current implementation uses a speech-extraction model through the Music-Source-Separation-Training (MSST) toolkit. In v1.0, a short audio window is selected according to acoustic energy. However, the maximum-response window may be dominated by music or other loud background components, which can be repeatedly reinforced when generated audio is reused during autoregressive rollout. The new construction instead uses the complete filtered speech content of each shot, providing cleaner and more stable speaker information. The visual and audio latents are then arranged in the same slot order, so that each memory item contains one compact visual observation and its corresponding event-level audio segment.

RoPE for memory conditioning. All historical conditions and user-provided memory inputs are represented in a common memory format. Following the discontinuous RoPE offsets introduced in ShotStream [24], we assign separated temporal RoPE regions to the current target, historical memory, and memory supplied before generation. The target video and audio tokens retain their original temporally aligned coordinates in a low-offset region below 500. When an image condition I_t is provided, its tokens remain in the coordinate system of the current target shot and are placed at temporal coordinate zero as its clean first frame.

The visual and audio components extracted from the i -th historical shot share the same temporal RoPE center:

$$\mu_i = 500 + 50(i - 1), \quad (11)$$

such that successive audio-visual memory items are centered at 500, 550, 600, $\dots$. All spatial tokens of the sampled visual memory frame use μ_i as their temporal coordinate while retaining their original spatial RoPE

coordinates. For the corresponding audio latent with native temporal coordinates $\{\tau_{i,\ell}^a\}_{\ell=1}^{L_i^a}$, we assign:

$$r_{i,\ell}^a = \mu_i + \tau_{i,\ell}^a - \bar{\tau}_i^a, \quad \bar{\tau}_i^a = \frac{1}{L_i^a} \sum_{\ell=1}^{L_i^a} \tau_{i,\ell}^a, \quad (12)$$

where ℓ indexes audio tokens. In other words, the temporal RoPE coordinates of the complete audio latent are shifted together so that its midpoint is aligned with the same center μ_i . This operation preserves the relative temporal structure of the audio event and does not modify the audio latent features themselves.

Memory supplied before generation is encoded independently and placed in a high-offset region:

$$M_0 = \{m_j^0\}_{j=1}^J, \quad m_j^0 = (m_j^{0,v}, m_j^{0,a}), \quad (13)$$

where

$$m_j^{0,v} = E_v(X_j^{\text{mem}}), \quad m_j^{0,a} = E_a(\mathcal{F}_{\text{speech}}(A_j^{\text{mem}})), \quad \rho_j \geq 5000. \quad (14)$$

The visual and audio components of the j -th initial memory item are both centered at ρ_j using the same positional assignment: the visual tokens use ρ_j as their temporal coordinate, while the audio RoPE coordinates are shifted together so that their midpoint is aligned with ρ_j . The centers $\{\rho_j\}_{j=1}^J$ are distinct and are assigned with enough spacing to keep the shifted audio-coordinate spans of different memory items non-overlapping. Consequently, the current target, historical memory, and initial memory occupy separated temporal RoPE regions. This separation keeps memory outside the target-shot timeline while allowing memory and image conditions to be provided simultaneously.

Memory interaction. After memory construction and RoPE assignment, memory and target tokens are jointly processed by the model. Unlike v1.0, v1.5 removes the memory-specific audio self-attention mask and the slot-paired cross-modal attention mask. All valid tokens interact through the original self-attention and bidirectional cross-modal attention paths. Their roles remain distinguishable because memory tokens use dedicated RoPE coordinates, remain clean, and are excluded from the prediction loss, whereas target tokens follow the target timeline and are optimized as generation outputs. This unified interaction removes additional hard constraints and naturally supports the joint use of memory and image conditions.

Joint conditional training. Training follows the two-stage data curriculum described in Sec. 2.1. Throughout the curriculum, we use a unified conditional interface that supports text-to-audio-video (T2AV), image-to-audio-video (I2AV), memory-conditioned text-to-audio-video (MT2AV), and memory-conditioned text-and-image-to-audio-video (MTI2AV). Let $b_M, b_I \in \{0, 1\}$ indicate whether audio-visual memory and a first-frame image condition are provided. The model condition for shot t is:

$$\mathcal{C}_t = (c_t, b_M M_{t-1}, b_I I_t), \quad (15)$$

where $(b_M, b_I) = (0, 0), (0, 1), (1, 0), (1, 1)$ correspond to T2AV, I2AV, MT2AV, and MTI2AV, respectively.

Let z_0^v and z_0^a denote the clean target video and audio latents, and let $\epsilon^v, \epsilon^a \sim \mathcal{N}(0, \mathbf{I})$. For each modality $q \in \{v, a\}$, rectified-flow training constructs the interpolation and target velocity for the tokens to be predicted:

$$z_\sigma^q = (1 - \sigma)z_0^q + \sigma\epsilon^q, \quad u^q = \epsilon^q - z_0^q. \quad (16)$$

We write the joint noised audio-visual target as $z_\sigma = (z_\sigma^v, z_\sigma^a)$ and the paired noise as $\epsilon = (\epsilon^v, \epsilon^a)$. Memory tokens are inserted as clean conditions with zero noise level and are excluded from the prediction loss. When $b_I = 1$, the first target video frame is also kept clean and excluded from the video loss. Let W_v and W_a be binary loss masks selecting only noised target tokens. The joint training objective is:

$$\mathcal{L} = \mathbb{E}_{\text{data}, \epsilon, \sigma, b_M, b_I} \left[\|W_v \odot (v_\theta^v(z_\sigma, \mathcal{C}_t) - u^v)\|_2^2 + \lambda_a \|W_a \odot (v_\theta^a(z_\sigma, \mathcal{C}_t) - u^a)\|_2^2 \right], \quad (17)$$

where λ_a balances the audio and video objectives. The resulting training distribution covers the four combinations of (b_M, b_I) , enabling generation without memory, with image conditioning alone, with memory alone, and with both conditions. This unified formulation supports T2AV, I2AV, MT2AV, and MTI2AV without changing the model interface or introducing separate objectives.

3.2 Acceleration with Distribution Matching Distillation

Long-form audio-visual generation requires both long-range consistency and low-latency inference. To accelerate JoyAI-Echo-1.5, we apply distribution matching distillation (DMD) to convert the original multi-step audio-visual generator into an 8-step student generator.

Data curation and training pipeline. JoyAI-Echo-1.5 performs DMD using memory conditions constructed from real long-form videos following Sec. 3.1. From the large-scale training corpus, we curate a compact subset with high-quality audio-visual memory and diverse generation conditions. The resulting data mixture covers T2AV, I2AV, and MT2AV, together with samples annotated by multi-shot captions. During training, we randomly sample these tasks according to predefined mixing ratios, allowing the student generator to preserve both single-shot generation quality and memory-conditioned multi-shot capability.

We find that DMD is particularly sensitive to the quality and distribution of visual memory. Low-level discrepancies in memory appearance, including brightness and saturation shifts, can bias the conditional distribution learned by the student and become progressively amplified when generated shots are written back into memory. Before DMD training, we therefore calibrate the low-level statistics of visual memory to better match the target training distribution. This preprocessing substantially improves distillation stability and reduces quality drift during multi-shot rollout.

The curated task mixture and calibrated memory conditions are used throughout the complete DMD pipeline. We first perform high-shift DMD with a relatively large learning rate to establish video quality and global structure, and then continue from the converged checkpoint with a lower shift and learning rate to refine audio and other fine-grained details. In both stages, the video branch is fully fine-tuned, while the audio and cross-modal modules are adapted through LoRA.

Asymmetric audio-visual optimization. Video and audio exhibit substantially different structures in the latent space. Video latents are highly structured and spatially redundant, with strong correlations across neighboring spatial locations and adjacent frames. Their temporal evolution is also relatively smooth, allowing global appearance, scene layout, and motion patterns to be represented through slowly varying latent structures. Audio latents, in contrast, encode dense one-dimensional temporal signals with rapidly changing local patterns. Fine-grained information such as phonetic content, pitch, transients, and acoustic textures may vary significantly within a short temporal interval [35].

These differences lead to asymmetric optimization requirements during few-step DMD. The video branch must substantially adapt its spatial and temporal representations to recover visual structure and motion along the shortened sampling trajectory. Restricting its trainable capacity results in insufficient adaptation and degraded visual fidelity. The audio branch, however, is more sensitive to aggressive distribution-matching updates. Small deviations in the audio latent can remove perceptually important local details, while excessive parameter updates may overwrite the acoustic prior of the multi-step teacher and introduce muffled, metallic, or electronic-sounding artifacts [37].

We therefore allocate different trainable capacities to the two modalities:

$$\theta_G^{\text{train}} = \{\theta_v^{\text{full}}, \Delta\theta_a^{\text{LoRA}}, \Delta\theta_{\text{cross}}^{\text{LoRA}}\}, \quad (18)$$

where the video branch is fully fine-tuned, while the audio branch and cross-modal attention modules are adapted through low-rank adaptation (LoRA). Full fine-tuning provides sufficient capacity for the video branch to learn the structural and motion changes required by few-step generation. In contrast, LoRA constrains the effective update space of the audio branch, slowing its deviation from the well-trained acoustic prior. Applying LoRA to the cross-modal attention modules preserves the flexibility required to adapt audio-visual correspondence without aggressively modifying the audio representation itself. This asymmetric parameterization therefore balances video adaptability with audio stability during joint DMD training.

Coarse-to-fine DMD noise scheduling. In DMD, the clean prediction produced by the student generator is re-noised to an auxiliary noise level σ , at which the teacher and fake-score model estimate the real and generated

scores [45, 46]. The choice of σ is important because it determines the granularity of the distributional information exposed to score matching. Higher noise levels place greater emphasis on coarse distributional properties and help stabilize global structure, whereas lower noise levels retain more sample-specific information and provide stronger supervision for fine-grained details.

Video and audio benefit differently from these noise regimes. Video generation relies strongly on high-noise DMD updates to establish stable appearance, scene layout, motion, and temporal structure under the shortened sampling trajectory. Audio quality, in contrast, is particularly sensitive to the low-noise regime, where local temporal patterns such as phonetic details, pitch, transients, and acoustic textures are more directly preserved and corrected. A fixed noise regime therefore cannot simultaneously provide the most suitable optimization signal for both modalities.

We address this mismatch with a two-stage DMD schedule. The first stage uses a relatively large timestep shift and learning rate, biasing score matching toward higher noise levels to establish stable video quality and structure. Starting from the converged first-stage checkpoint, the second stage reduces both the timestep shift and learning rate, shifting the optimization toward lower noise levels to refine audio quality and other fine-grained details. This coarse-to-fine schedule accommodates the different noise-level preferences of video and audio during joint few-step distillation.

Memory-robust and task-adaptive DMD. For the DMD stage described here, we consider T2AV/I2AV/MT2AV, where MT2AV generates the current shot conditioned on the audio-visual memory bank M_{t-1} . Although the training memory is constructed from real long-form videos, the memory available during inference is progressively extracted from previously generated shots. Even small generation errors can therefore alter the memory distribution and accumulate over a multi-shot rollout. A student trained only with undegraded memory may become sensitive to this discrepancy, leading to gradual degradation of visual quality and cross-shot consistency.

To improve robustness, we apply stochastic degradation to the visual memory during DMD training. The student generator and fake-score model receive the same probabilistically degraded memory, allowing the fake-score model to track the student distribution under imperfect memory conditions. In contrast, the real-score model receives the original, undegraded memory and provides a stable target corresponding to the desired conditional distribution. The resulting score difference encourages the student to recover a high-quality current shot even when its historical memory contains moderate errors. This stochastic degradation complements the visual-memory distribution calibration described above: calibration corrects systematic low-level biases in the training memory, while degradation simulates the sample-dependent errors encountered during autoregressive rollout.

Moreover, MT2AV and other tasks also exhibit different audio optimization dynamics. In T2AV or I2AV, the student must construct the complete audio distribution from text and noise without any audio reference. In MT2AV, the memory condition already provides informative cues about speaker timbre, acoustic context, and temporal continuity, making its audio component easier to learn. Consequently, using the same audio weight for both tasks can cause the MT2AV audio objective to contribute disproportionately to joint DMD optimization.

We therefore introduce a task-dependent audio weight:

$$\mathcal{L}_{\text{DMD}} = \mathcal{L}_{\text{DMD}}^v + \lambda_a(\kappa)\mathcal{L}_{\text{DMD}}^a, \quad (19)$$

where $\kappa \in \{\text{T2AV}, \text{I2AV}, \text{MT2AV}\}$ denotes the DMD task, with $\lambda_a(\text{T2AV/I2AV}) = 0.25$ and $\lambda_a(\text{MT2AV}) = 0.10$. The reduced MT2AV weight prevents its easier memory-conditioned audio objective from dominating joint distillation, while the larger weight retains sufficient supervision for learning audio generation directly from noise. Together, stochastic memory degradation and task-adaptive weighting improve the robustness of few-step generation across both single-shot and memory-conditioned settings.

Audio distillation with adversarial and energy constraints. Despite the conservative Audio-LoRA parameterization and low-noise refinement stage, applying score-based DMD to audio remains challenging. The

distribution-matching gradient mainly corrects the discrepancy between the teacher and student distributions at noisy latent states. For dense audio signals, this correction can favor dominant acoustic modes while suppressing subtle temporal variations, producing over-smoothed, muffled, or electronic-sounding outputs. We further observe that DMD may cause the overall audio latent level to increase continuously during training, leading to excessive loudness, clipping, and harsh distortion after decoding. We therefore introduce two complementary audio-specific objectives to constrain global signal level and local acoustic realism.

First, we match the root-mean-square (RMS) level of the generated audio latent to that of its ground-truth counterpart:

$$\mathcal{L}_{\text{energy}}^a = \psi(\log(\text{RMS}(\hat{z}_0^a) + \varepsilon_{\text{num}}) - \log(\text{RMS}(z_0^a) + \varepsilon_{\text{num}})), \quad (20)$$

where $\hat{z}_0^a$ and z_0^a denote the predicted and ground-truth clean audio latents, respectively. The robust penalty ψ limits the influence of extreme samples, and $\varepsilon_{\text{num}} > 0$ ensures numerical stability. Computing the discrepancy in the log domain makes the loss sensitive to relative rather than absolute level differences. This constraint suppresses persistent latent-level drift and reduces the risk of clipping after audio decoding.

Second, following AudioX-Turbo [37], we introduce a diffusion-based audio discriminator. We reuse intermediate features from the frozen audio teacher as the discriminator backbone and train only lightweight heads to distinguish generated and real audio latents. By preserving the pretrained audio representation and limiting adversarial training to the added heads, the discriminator provides a stable perceptual signal without modifying the teacher. Its local adversarial feedback complements the distribution-level DMD objective by preserving transients, acoustic textures, and other fine-grained temporal details.

The complete student-generator objective is

$$\mathcal{L}_G = \mathcal{L}_{\text{DMD}}^v + \lambda_a(\kappa)\mathcal{L}_{\text{DMD}}^a + \lambda_{\text{GAN}}\mathcal{L}_{\text{GAN}}^a + \lambda_{\text{energy}}\mathcal{L}_{\text{energy}}^a. \quad (21)$$

The DMD terms align the global audio-visual generation distribution, the adversarial loss restores local acoustic details, and the RMS-level constraint stabilizes the overall audio signal level. Together, these objectives mitigate both perceptual over-smoothing and excessive energy growth during few-step distillation.

4 Action-Conditioned World Modeling

4.1 Overview and Problem Formulation

Given an optional clean media context $\mathcal{M}$, a text condition $\mathcal{C}$, and a user control sequence $U_{1:T}$, our goal is to jointly generate synchronized video $V_{1:T}$ and audio $A_{1:T}$ under user-specified camera intent. The media context $\mathcal{M}$ accommodates an empty condition $\emptyset$, a single reference image I_{ref} , or an audio-visual prefix $(V_{1:\tau}, A_{1:\tau})$. We serialize user navigation inputs into a time-indexed metric camera trajectory event stream $\mathcal{E}_P = \{(t, P_t)\}_{t=1}^T$, where P_t represents a calibrated relative 6-DoF pose. The joint generation process is formulated as:

$$p_{\theta, \phi}(V_{1:T}, A_{1:T} \mid \mathcal{M}, \mathcal{C}, \mathcal{E}_P), \quad (22)$$

where θ denotes the audio-visual backbone, and ϕ represents the parallel relative trajectory-conditioning pathway. Static text fields specify scene semantics, while $\mathcal{E}_P$ dictates the continuous temporal trajectory evolution. When $\mathcal{M}$ contains an observed audio-visual prefix, those prefix samples are clamped and only the unobserved suffix is generated; the full-sequence notation in Eq. (22) is used for compactness.

4.2 Unified Camera-Intent Interface & Scale Calibration

Realizing continuous world modeling across heterogeneous datasets requires overcoming two key obstacles: the incompatibility of controller-specific action spaces and the scale mismatch of relative motion amplitudes. To address these issues, we decouple user-facing controls from model-level conditions through a two-fold strategy: (1) mapping arbitrary navigation inputs into a unified, subject-agnostic 6-DoF trajectory, and (2) calibrating relative translations with a robust global scale.

Relative 6-DoF Representation. We define camera intent using relative 6-DoF trajectories rather than actor-specific control states. Let $T_t \in \text{SE}(3)$ denote the camera-to-world extrinsic matrix at frame t . The relative motion between frames i and j is given by:

$$\Delta T_{ij} = T_i^{-1} T_j, \quad \xi_{ij} = \text{Log}(\Delta T_{ij})^\vee \in \mathbb{R}^6, \quad (23)$$

where ξ_{ij} is the corresponding 6-DoF twist coordinate. For discrete keyboard inputs $\mathbf{u}_t$, a fixed control mapping $\mathcal{F}_{\text{ctrl}}$ converts each input into a step-wise camera increment:

$$\xi_t = \mathcal{F}_{\text{ctrl}}(\mathbf{u}_t), \quad T_t = T_{t-1} \text{Exp}(\widehat{\xi}_t), \quad \Delta T_t = T_0^{-1} T_t, \quad (24)$$

where $\widehat{\xi}_t \in \mathfrak{se}(3)$. Continuous metric trajectories bypass $\mathcal{F}_{\text{ctrl}}$ and directly produce relative poses $\Delta T_{1:T}$. This unifies raw actions and estimated poses under a single geometric interface.

Global Translation Scale Calibration. To ensure consistent velocity semantics without introducing boundary artifacts in multi-turn rollouts, we calibrate relative translations across all data sources using a fixed global dataset-level scale s_{global} . For clip i , let $\Delta \mathbf{t}_{i,k}$ denote the relative translation at frame k with respect to the initial frame. We compute

$$s_{\text{global}} = Q_{0.9} \left(\left\{ \max_k \|\Delta \mathbf{t}_{i,k}\|_2 \right\}_{i \in \mathcal{D}_{\text{train}}} \right). \quad (25)$$

Relative translations are calibrated via $\widehat{\mathbf{t}}_{i,k} = \Delta \mathbf{t}_{i,k} / s_{\text{global}}$, yielding $P_{i,k} = [\Delta \mathbf{R}_{i,k} \mid \widehat{\mathbf{t}}_{i,k}]$ while keeping rotation matrices $\Delta \mathbf{R}_{i,k}$ intact.

4.3 Relative Trajectory Conditioning

Instead of absolute Plücker coordinates which depend on arbitrary world origins, we inject trajectory geometry through Unified Camera Positional Encoding (UCPE) [49]. UCPE makes pairwise camera geometry invariant to global coordinate shifts while preserving calibrated relative translation magnitudes.

For a latent token $i = (t, s)$ at frame t and spatial cell s , we define its local ray transformation $\mathbf{D}_i = (\mathbf{T}_i^{\text{wr}})^{-1}$ derived from calibrated pose P_t and intrinsics K_t . We inject this geometry into parallel attention heads via matrix scaling $\mathbf{G}_i = \mathbf{I}_{d/8} \otimes \mathbf{D}_i$ combined with spatiotemporal RoPE:

$$\begin{aligned} \widetilde{\mathbf{q}}_i^c &= (\mathbf{G}_i^\top \oplus \text{RoPE}_i) \mathbf{q}_i^c, \\ (\widetilde{\mathbf{k}}_i^c, \widetilde{\mathbf{v}}_i^c) &= (\mathbf{G}_i^{-1} \oplus \text{RoPE}_i) (\mathbf{k}_i^c, \mathbf{v}_i^c), \\ \mathbf{o}_i^c &= (\mathbf{G}_i \oplus \text{RoPE}_i^{-1}) \text{Attn}_{\text{cam}}(\widetilde{\mathbf{q}}^c, \widetilde{\mathbf{k}}^c, \widetilde{\mathbf{v}}^c)_i. \end{aligned} \quad (26)$$

Here, $\mathbf{T}_i^{\text{wr}}$ is the UCPE ray-to-world transform constructed from (P_t, K_t, s) , so $\mathbf{D}_i$ maps world coordinates to the ray-local frame; d is the attention-head width, and the superscript c identifies the parallel camera-conditioning stream. The symbol $\oplus$ denotes a block-diagonal direct sum with the spatiotemporal RoPE operator, and Attn_{cam} denotes attention in the camera pathway. The resulting camera-attention outputs pass through zero-initialized linear projections in the trajectory pathway parameterized by ϕ and are added to the main video self-attention blocks. The trajectory condition is applied exclusively to the visual backbone. One UCPE pathway is shared across first- and third-person data. Geometry alone does not determine whether a trajectory is realized from an egocentric or third-person observer-subject configuration, so the static viewpoint field supplies this initial relationship while $P_{1:T}$ supplies camera intent. This division allows the same trajectory representation and parameters to serve both viewpoints while leaving their camera-character realization to the learned world prior.

4.4 Progressive Training Curriculum

Data sources with rich audio-visual semantics and those providing precise, clean control trajectories are fundamentally heterogeneous. To resolve the resulting optimization conflicts, we employ a progressive three-stage curriculum that decouples audio-visual prior acquisition, control alignment, and joint consolidation.

Stage 1: Audio-Visual Continued Pretraining (AV-CPT). This stage adapts the base model to the visual styles, motion statistics, and acoustic dynamics (footsteps, ambient audio, and speech) of interactive game and general environments. We optimize the full backbone θ on the AV-rich data mixture while completely disabling trajectory conditioning. To simultaneously establish text-to-AV generation, image-to-AV generation, and temporal continuation capabilities, we randomly sample the clean media condition $\mathcal{M}$ from an empty set $\emptyset$, a single reference frame I_{ref} , or a synchronized audio-visual prefix $(V_{1:\tau}, A_{1:\tau})$. Unobserved video and audio latent regions are generated jointly, embedding strong acoustic priors and temporal continuity into the backbone.

Stage 2: Action Fine-Tuning (Action-SFT). This stage isolates control responsiveness without degrading the acquired audio-visual generation priors. We freeze the entire backbone θ and train only the lightweight UCPE camera pathway parameters ϕ on control-clean data using a visual-only denoising objective. To eliminate information shortcuts, we omit the dynamic-motion narrative and all audio-related fields, retaining only persistent scene, style, viewpoint, and subject descriptors $\mathcal{C}_{\text{static}}$. This forces the trajectory $\mathcal{E}_{\mathcal{P}}$ to serve as the sole source of motion guidance. Furthermore, we heavily bias condition sampling toward reference-frame conditioning (I_{ref}), fixing initial scene content so the network concentrates parameter updates entirely on camera trajectory following.

Stage 3: Joint Fine-Tuning (Joint-FT). To consolidate control sensitivity with joint audio-visual synthesis, we unfreeze both the backbone θ and trajectory pathway ϕ for end-to-end co-adaptation on the balanced high-quality mixture. We restore environmental sound and speech fields to the text prompt (while keeping dynamic motion descriptions excluded) and resume joint audio-visual loss optimization. Training executes with a reduced learning rate to harmoniously align multimodal quality with trajectory responsiveness without collapsing the pre-trained priors.

5 Autoregressive Audio-Visual Generation

We progressively adapt a pretrained bidirectional, multi-step audio-visual diffusion model into a streaming autoregressive generator with high throughput and low latency. This transition requires two key transformations: converting bidirectional temporal attention into causal autoregressive computation for online generation of synchronized audio-visual chunks, and compressing iterative denoising into a few-step sampler that remains stable under self-generated histories. To achieve this, we first apply audio-visual teacher forcing [2, 42] to initialize a causal streaming generator from the bidirectional backbone. We then introduce short-horizon Self-Gradient Forcing (SGF) [13, 54] to train on self-generated rollout states while distilling few-step generation, and further extend SGF to long-horizon self-rollouts to improve robustness and stability during sustained streaming inference.

5.1 Audio-Visual Teacher Forcing

We first use teacher forcing to initialize a causal generator from the pretrained bidirectional model. Specifically, during training, ground-truth audio-visual chunks are provided as the historical context, such that the generation of the current chunk depends only on preceding clean chunks rather than future observations. This stage provides a stable causal initialization before exposing the model to self-generated histories in subsequent self-forcing training.

Causal initialization. Let $m \in \{v, a\}$ index the modality, with v denoting video and a denoting audio. We divide both latent streams into N temporally aligned macro-chunks. The clean latent of modality m in chunk i is denoted by x_i^m , where $i \in \{1, \dots, N\}$. An aligned video-audio pair spans the same time interval, although the two chunks may contain different numbers of tokens. For each pair, we sample a shared noise level $\sigma_i \in (0, 1)$ and construct:

$$x_{\sigma_i, i}^m = (1 - \sigma_i)x_i^m + \sigma_i\epsilon_i^m, \quad \epsilon_i^m \sim \mathcal{N}(0, \mathbf{I}), \quad (27)$$

where ϵ_i^m is Gaussian noise and $\mathbf{I}$ is the identity covariance matrix. The shared σ_i places the paired audio and video chunks at a consistent diffusion stage.

To enable causal generation without sacrificing training parallelism, we concatenate the clean latents with their noisy counterparts. Let $\mathcal{B}_{i,\text{clean}}$ and $\mathcal{B}_{i,\text{noisy}}$ denote the clean and noisy latent blocks of the i -th paired audio-visual chunk, respectively. We impose the following causal attention pattern:

$$\mathcal{S}(\mathcal{B}_{i,\text{clean}}) = \{\mathcal{B}_{j,\text{clean}} : j \leq i\}, \quad \mathcal{S}(\mathcal{B}_{i,\text{noisy}}) = \{\mathcal{B}_{j,\text{clean}} : j < i\} \cup \{\mathcal{B}_{i,\text{noisy}}\}, \quad (28)$$

where $\mathcal{S}(\cdot)$ denotes the set of latent blocks accessible to the corresponding attention query. Under this mask, each noisy chunk can attend to all preceding clean audio-visual chunks and to the noisy tokens within the current chunk, while the corresponding clean target and all future chunks are masked out. We apply the same causal constraint to video self-attention, audio self-attention, and both directions of audio-visual cross-attention. For efficient causal attention computation, we implement this pattern using PyTorch FlexAttention [5] with a block-level causal mask. The resulting BlockMask exploits the structured sparsity of chunk-wise causal attention, allowing masked blocks to be skipped without materializing a dense attention mask. As illustrated in Fig. 3(a), this causal attention pattern enables each chunk to use the complete preceding clean history while preventing information leakage from the current clean target and future chunks. This converts the original bidirectional temporal attention into chunk-wise causal attention, while retaining parallel prediction of all chunks during training.

Training objective. To preserve the pretrained generative capability while efficiently adapting the model to the new causal attention pattern, we perform this stage using LoRA-based parameter-efficient fine-tuning [11]. The pretrained backbone is kept fixed, while the LoRA parameters are optimized to learn the causal audio-visual generation behavior. Given the causally masked sequence, the model jointly predicts the flow velocities of the audio and video latents in a single forward pass. We retain the standard flow-matching objective of the pretrained diffusion model:

$$\mathcal{L}_{\text{TF}} = \sum_{m \in \{v,a\}} \lambda_m \mathbb{E} \left[\sum_{i=1}^N \|v_{\theta}^m(x_{\sigma_i,i}^m, \sigma_i) - (\epsilon_i^m - x_i^m)\|_2^2 \right], \quad (29)$$

where v_{θ}^m denotes the predicted flow velocity for modality m , λ_m balances the video and audio losses, and θ denotes the trainable LoRA parameters of the generator. Thus, audio-visual teacher forcing adapts the pretrained bidirectional diffusion model to causal, chunk-wise autoregressive generation while preserving its original flow-matching training objective.

5.2 Short-Horizon Audio-Visual Self-Gradient Forcing

Starting from the causal teacher-forcing initialization, we adopt Self-Gradient Forcing (SGF) [54] to adapt the generator to its own autoregressive audio-visual histories. SGF follows the deployment-time self-rollout distribution while restoring gradients through the computation that encodes generated contexts into future-readable causal memory. Unlike standard Self Forcing [13], which treats the historical KV cache as detached rollout state, SGF reconstructs the sampled rollout state in a differentiable forward pass. Losses on later chunks can therefore supervise how preceding generated audio-visual contexts are encoded into causal memory for subsequent generation, without backpropagating through the full sequential sampling trajectory. We further combine SGF with a DMD objective [46] to distill the original multi-step diffusion model into a few-step autoregressive generator, jointly improving autoregressive robustness and sampling efficiency.

Forward 1: Autoregressive self-rollout. The generator first performs a complete chunk-wise rollout with gradient computation disabled, using exactly the few-step sampler deployed at inference. For example, with a four-step sampler, each synchronized audio-visual chunk is generated through four denoising steps, followed by an additional clean-context forward at $t = 0$ that prefills the causal KV cache for subsequent chunks. During the rollout, we sample an exit step t_e and record, for each chunk i , the corresponding noisy audio-visual latents $r_i = (r_i^v, r_i^a)$ together with the final generated clean latents $\hat{x}_i = (\hat{x}_i^v, \hat{x}_i^a)$. Since Forward 1 is executed under `no_grad`, it faithfully reproduces the inference-time autoregressive state distribution without retaining the computation graph of the sequential rollout.

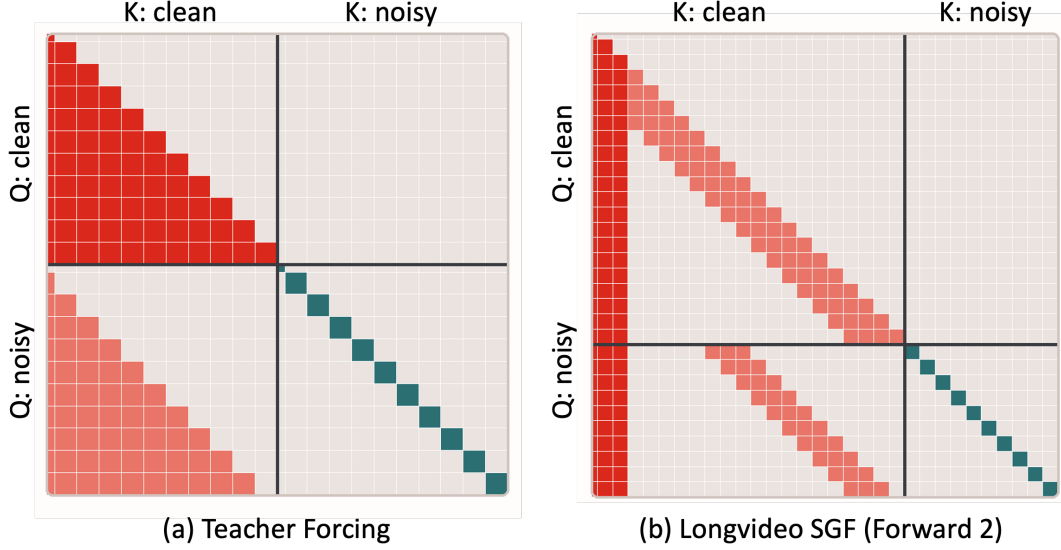


Figure 3 Illustration of the causal attention masks used in streaming post-training. (a) Audio-visual teacher forcing, where each noisy chunk attends to its preceding clean history and the current noisy chunk, while the clean target and future chunks are masked out. (b) Long-horizon SGF Forward 2, where the causal context is further restricted by the sink-plus-FIFO KV cache.

Forward 2: Audio-visual context-gradient reconstruction. We then reconstruct the sampled rollout computation in a single differentiable forward pass using the states recorded in Forward 1. Specifically, the generated clean latents $\{\hat{x}_i\}_{i=1}^N$ are provided as stop-gradient context inputs at the clean context timestep $t = 0$, while the recorded noisy audio-visual latents $\{r_i\}_{i=1}^N$ retain their sampled exit timestep t_e . The clean-context and noisy target tokens are jointly arranged according to the causal construction in Eq. (28): the clean latents of chunk i provide autoregressive context for subsequent chunks, whereas r_i represents the prediction target of chunk i at the sampled exit state. This construction reproduces the context-target relation of the serial rollout while allowing all recorded chunks to be evaluated in parallel.

Our LTX-2.3-based audio-visual backbone contains five attention pathways that transmit rollout-dependent information: video self-attention, audio self-attention, audio-to-video (A2V) attention, video-to-audio (V2A) attention, and the spatial attention adapter associated with Unified Camera Positional Encoding (UCPE). As illustrated in Fig. 3(a), we apply the same chunk-level causal attention pattern to all five pathways. For a query associated with chunk i , the corresponding keys and values are restricted to the current or preceding chunks according to the autoregressive rollout geometry. The unimodal masks prevent video and audio tokens from accessing future content; the A2V and V2A masks preserve the same temporal causality during bidirectional cross-modal interaction; and the UCPE mask prevents future camera-conditioned features from leaking into the current prediction. Text cross-attention remains unchanged because the text prompt is globally available and contains no rollout-dependent future information.

Although both the exit-step states and the generated clean latents are detached from Forward 1, their context representations are recomputed with gradient tracking in Forward 2. In particular, the $t = 0$ clean-context computation reconstructs the K/V projections and context attention operations through which generated audio and video chunks are written into causal memory. Later-chunk losses can therefore propagate through the reconstructed context representations in all five attention pathways. This supervision improves not only how each modality writes its own autoregressive history, but also how audio and video histories are exchanged through A2V and V2A attention and how camera information is incorporated through UCPE. Applying the matched causal attention masks allows all chunks to be reconstructed in parallel, yielding the recovered clean audio-visual predictions $\{\hat{x}_i = (\hat{x}_i^v, \hat{x}_i^a)\}_{i=1}^N$.

DMD optimization. We optimize the generator with a joint audio-visual DMD objective evaluated on the recovered predictions conditioned on self-generated autoregressive histories. Since these predictions are conditioned on the generator’s own audio-visual history rather than ground-truth context, the objective adapts the generator to the state distribution encountered during deployment. Moreover, because the reconstructed predictions are coupled through the causal A2V and V2A attention pathways, the objective supervises not only the quality of each modality but also the consistency of their future-readable representations. It therefore improves few-step sampling efficiency while mitigating unimodal error accumulation and audio-visual synchronization drift caused by self-generated histories.

The generator is initialized from the causal model obtained after the teacher-forcing stage, while both the real-score and fake-score models are initialized from the pretrained bidirectional diffusion model. The real-score model is kept frozen to represent the target audio-visual distribution, whereas the fake-score model is trainable and continuously updated to track the evolving distribution of the few-step generator.

For a sampled diffusion timestep s , we perturb the recovered clean audio-visual latents as

$$\bar{x}_{s,i} = (1 - s)\bar{x}_i + s\eta_i, \quad \eta_i \sim \mathcal{N}(0, \mathbf{I}), \quad (30)$$

where $\bar{x}_i = (\bar{x}_i^v, \bar{x}_i^a)$ denotes the jointly recovered audio-visual prediction. The frozen real-score model s_{real} estimates the score of the pretrained data distribution, while the trainable fake-score model s_{fake} estimates the score of the current generator distribution. The generator is optimized with the DMD objective

$$\mathcal{L}_{\text{SGF-DMD}} = \mathbb{E}_{i,s,\eta_i} [w(s) \langle \text{sg}(s_{\text{fake}}(\bar{x}_{s,i}, s) - s_{\text{real}}(\bar{x}_{s,i}, s)), \bar{x}_i \rangle], \quad (31)$$

where $\text{sg}(\cdot)$ denotes stop-gradient and $w(s)$ is the diffusion-dependent weighting factor. This objective drives the few-step generator toward the distribution represented by the pretrained multi-step model while preserving gradients through the reconstructed causal audio-visual context.

Following each generator update, we perform multiple fake-score updates. The generated samples are detached and used to train the fake-score model with the standard flow-matching objective, allowing it to track the evolving generator distribution. This short-horizon audio-visual SGF stage establishes robust few-step generation under deployment-time autoregressive states before training is extended to substantially longer streaming horizons.

5.3 Long-Horizon Audio-Visual Self-Gradient Forcing

Short-horizon audio-visual SGF reconstructs the causal computation within the original training horizon, but sustained streaming generation encounters histories containing progressively accumulated prediction errors. We therefore extend SGF to substantially longer synchronized audio-visual trajectories. The few-step sampler and DMD formulation remain unchanged; the main difference is that Forward 1 exposes the generator to long-horizon self-generated states, while Forward 2 reconstructs the resulting trajectory under a sparse causal attention pattern. This stage adapts both the denoising computation and the audio-visual memory-writing process to the distribution shift encountered during long-running autoregressive generation.

Long-horizon rollout construction. We construct a long-horizon trajectory by sequentially rolling out multiple temporal segments, each corresponding to the training horizon used during short-horizon SGF. To preserve temporal continuity across segment boundaries, the terminal video frame of the k -th rollout segment is reused as the initial frame of the $(k + 1)$ -th segment. Adjacent rollout segments therefore share a boundary frame instead of restarting from independent initial conditions. Throughout the trajectory, all subsequent audio-visual chunks are generated autoregressively from the model’s preceding outputs. Consequently, later segments are conditioned on histories that already contain prediction errors accumulated during earlier segments.

Bounded multimodal causal context. Maintaining a full causal KV cache is impractical for sustained streaming generation because its size grows with the rollout duration. We instead use a sink-plus-FIFO policy that preserves a small persistent prefix together with a bounded window of recent history. This policy is applied consistently to the video, audio, A2V, V2A, and UCPE attention pathways.

Let S_m denote the number of persistent sink chunks and R_m the number of recent chunks retained for modality m . The historical context directly available when generating chunk i is

$$\mathcal{H}_i^m = \underbrace{(\hat{x}_1^m, \dots, \hat{x}_{\min(S_m, i-1)}^m)}_{\text{attention sink}} \parallel \underbrace{(\hat{x}_{\max(S_m+1, i-R_m)}^m, \dots, \hat{x}_{i-1}^m)}_{\text{recent FIFO history}}, \quad (32)$$

where $\parallel$ denotes ordered concatenation and an indexed range is empty when its lower bound exceeds its upper bound. The attention sink preserves a fixed long-term prefix, whereas the FIFO portion continuously retains the most recent generated chunks. Consequently,

$$|\mathcal{H}_i^m| \leq S_m + R_m, \quad (33)$$

making the number of historical K/V entries directly accessed at each layer independent of the total rollout duration. The same sink-plus-FIFO policy and causal attention geometry are used during training and inference.

As illustrated in Fig. 3(b), Forward 2 reconstructs the long audio-visual trajectory under the corresponding sink-plus-FIFO masks. The reconstruction retains the long trajectory as a parallel sparse computation graph, while each query is allowed to read only the context specified by its sink-plus-FIFO neighborhood. The clean audio-visual chunks are recomputed at $t = 0$, and the noisy target chunks retain the sampled exit timestep t_e , following the same context-target construction as in short-horizon SGF.

Layer-wise long-range context gradients. Importantly, the sink-plus-FIFO bound constrains only the K/V entries that a query can read directly within an individual Transformer layer. It does not restrict the end-to-end receptive field or gradient horizon of the stacked network to the same single-layer window. At an upper Transformer layer, a recent FIFO chunk visible to the current query already contains information aggregated from its own sink-plus-FIFO neighborhood in preceding layers. A loss on the current chunk can therefore first propagate to the directly visible K/V representations and subsequently continue through their lower-layer hidden states to the K/V-writing computations of earlier chunks. Repeating this process across layers creates a multi-hop context-gradient path whose temporal reach can extend substantially beyond the sink and FIFO entries directly visible at any single layer.

For example, suppose that chunk i directly attends to a recent chunk j at layer ℓ . Although an earlier chunk h may lie outside the direct sink-plus-FIFO window of chunk i , chunk j may have attended to h in a preceding layer. The gradient from the loss on chunk i can then follow the path

$$\mathcal{L}_i \longrightarrow \text{KV}_j^{(\ell)} \longrightarrow \text{KV}_h^{(\ell-1)}, \quad (34)$$

and analogous paths can be composed across additional layers and intermediate chunks. Hence, $S_m + R_m$ bounds the direct attention fan-in at each layer, but does not bound the effective multi-layer context-gradient reach.

This layer-wise propagation also operates across modalities. A video loss may reach earlier audio representations through A2V attention, an audio loss may reach earlier video representations through V2A attention, and both may propagate through the camera-conditioned representations introduced by UCPE. Moreover, the shared boundary between adjacent rollout segments is preserved during the parallel reconstruction rather than treated as a gradient boundary. Losses from later segments can consequently supervise more distant audio-visual context-writing operations through intermediate chunks, cross-modal attention, and successive Transformer layers.

This transitive supervision remains fundamentally different from full rollout backpropagation. The generated latents, noise states, sampling decisions, and serial KV-cache trajectory recorded in Forward 1 remain detached. Gradients flow only through the parallel Forward-2 reconstruction, including the clean-context projections, the five causally masked attention pathways, and the target-side denoising computation. Long-horizon audio-visual SGF therefore supervises distant memory-writing operations through a bounded-degree sparse reconstruction graph without retaining the recurrent autograd graph of the sequential rollout.

Temporal scope of DMD score evaluation. During long-horizon training, the autoregressive generator retains its original RoPE configuration. The sink-plus-FIFO policy only controls which historical K/V states are retained and visible, without modifying the generator’s positional encoding. Since the real-score and fake-score networks are initialized from a bidirectional diffusion model pretrained on fixed-duration sequences, their DMD evaluations during training are restricted to temporal spans covered by the pretraining horizon. This keeps score-based supervision within the temporal regime observed by the score models while leaving the generator’s positional representation unchanged.

Long-horizon DMD supervision. We apply the DMD objective to every rollout segment instead of supervising only the final segment of the long trajectory. Let $\mathcal{L}_{\text{DMD}}^{(k)}$ denote the joint audio-visual DMD loss computed from the recovered predictions of the k -th rollout segment. For a trajectory containing K segments, the long-horizon objective is accumulated across all segments:

$$\mathcal{L}_{\text{Long-SGF}} = \sum_{k=1}^K \mathcal{L}_{\text{DMD}}^{(k)}. \quad (35)$$

Supervision is therefore distributed throughout the complete self-generated trajectory. Later-segment losses are evaluated under histories that already contain errors accumulated in earlier segments and, through the layer-wise paths described above, can supervise context-writing computations beyond their directly visible KV windows. Jointly optimizing these segment-wise losses encourages the few-step audio-visual world model to preserve visual identity, scene geometry, audio continuity, and cross-modal synchronization under progressively shifted autoregressive states.

Sparse background temporal regularization. We observe that the four-step distilled LTX generator can occasionally produce transient gray-speckle artifacts in locally stable regions, which become particularly noticeable during long-horizon autoregressive rollouts. To distinguish these artifacts from genuine motion, we compare two consecutive temporal transitions: the historical change from frame $t-2$ to $t-1$, and the current change from frame $t-1$ to t .

Specifically, for each generated latent token $\hat{z}_{t,p}$ at spatial location p , we first find its closest latent token in a local neighborhood of frame $t-1$. We then trace this matched token back to its closest correspondence in frame $t-2$:

$$q_{t,p}^* = \arg \min_{q \in \mathcal{N}(p)} \|\hat{z}_{t,p} - \hat{z}_{t-1,q}\|_2, \quad r_{t,p}^* = \arg \min_{r \in \mathcal{N}(q_{t,p}^*)} \|\hat{z}_{t-1,q_{t,p}^*} - \hat{z}_{t-2,r}\|_2. \quad (36)$$

The corresponding current and historical changes are

$$d_{t,p}^{\text{cur}} = \frac{1}{\sqrt{C}} \|\hat{z}_{t,p} - \hat{z}_{t-1,q_{t,p}^*}\|_2, \quad d_{t,p}^{\text{hist}} = \frac{1}{\sqrt{C}} \|\hat{z}_{t-1,q_{t,p}^*} - \hat{z}_{t-2,r_{t,p}^*}\|_2, \quad (37)$$

where C denotes the latent dimension. If both transitions exhibit comparable changes, the variation is likely caused by persistent object or camera motion. In contrast, a current change that is substantially larger than its historical counterpart is more likely to correspond to a transient artifact.

To account for textured regions and global frame-level variation, we construct an adaptive normal-variation scale $s_{t,p}$ by combining the historical change, the local spatial variation around the matched token, and a stabilizing frame-wise floor. We select a sparse set $\mathcal{A}$ of locations for which $d_{t,p}^{\text{cur}}/s_{t,p}$ is unusually large and penalize only the change exceeding an adaptive tolerance:

$$\mathcal{L}_{\text{bg-temp}} = \frac{1}{|\mathcal{A}|} \sum_{(t,p) \in \mathcal{A}} \rho\left([d_{t,p}^{\text{cur}} - \alpha s_{t,p}]_+\right), \quad (38)$$

where ρ is a robust penalty and α controls the tolerated normal variation. This sparse, motion-adaptive formulation selectively suppresses transient gray-speckle artifacts while preserving persistent scene, object, and camera motion. The complete generator objective is

$$\mathcal{L}_{\text{gen}} = \mathcal{L}_{\text{Long-SGF}} + \lambda_{\text{bg}} \mathcal{L}_{\text{bg-temp}}, \quad (39)$$

where λ_{bg} controls the strength of the auxiliary regularization.

5.4 Autoregressive Inference

At inference, the generator produces synchronized audio-visual chunks autoregressively using the same few-step sampler and sink-plus-FIFO cache introduced during training. For each new chunk, the retained modality-specific histories $\mathcal{H}_i = (\mathcal{H}_i^v, \mathcal{H}_i^a)$ provide the causal context, while outdated FIFO entries are discarded to keep the KV cache bounded. Consistent with long-horizon training, the temporal RoPE indices of the sink, FIFO, and current chunks are rebased into a fixed positional window as generation proceeds, maintaining a consistent temporal positional layout throughout streaming inference.

Let Θ collect the active generator parameters (including the trajectory pathway ϕ for the world-model variant), and let $\mathcal{D}_i$ denote the variant-specific conditions: text together with memory and/or an image for the long-video variant, or media context, text, and trajectory for the world-model variant. The next synchronized chunk is sampled as:

$$(V_i, A_i) \sim p_{\Theta}(\cdot \mid \mathcal{H}_i, \mathcal{D}_i), \quad (40)$$

where V_i and A_i denote the generated video and audio chunks, respectively.

5.5 Towards Causal Memory

Teacher forcing in Eq. (28), long-horizon SGF, and inference already share one causal rule. A noisy chunk may attend only to previously generated clean history. The current clean target and all future chunks remain masked. The sink-plus-FIFO cache in Eq. (32) is the bounded form of this rule. The attention sink preserves a fixed early prefix, the FIFO retains the most recent chunks, and temporal RoPE within each $\mathcal{H}_i^m$ is rebased as outdated FIFO entries are discarded, so that $\mathcal{H}_i = (\mathcal{H}_i^v, \mathcal{H}_i^a)$ keeps a consistent positional layout throughout streaming generation. This cache is sufficient for local audio-visual continuity under the few-step sampler. It is not sufficient once a chunk has left the FIFO. Those tokens are no longer available, even if the sink still holds the opening of the stream. When the camera returns, or when a model trajectory revisits a previously observed place, the generator should not treat the scene as new merely because the corresponding history has been evicted.

To retain structure beyond the sliding window without changing the causal mask, the few-step sampler, or the variant-specific conditions $\mathcal{D}_i$ in Eq. (40), we introduce a compact recurrent state $\mathcal{M}_i$ that is written only from the past. Let C_i denote a short leak-free prefix taken strictly before the start of chunk i , and let $\psi \subset \Theta$ collect the memory parameters inside the generator, including the trajectory pathway ϕ when the world-model variant is used. The next synchronized chunk and the updated memory are

$$(V_i, A_i) \sim p_{\Theta}(\cdot \mid \mathcal{H}_i, \mathcal{M}_{i-1}, \mathcal{D}_i), \quad \mathcal{M}_i = \mathcal{F}_{\psi}(\mathcal{M}_{i-1}, C_i, V_i). \quad (41)$$

Thus $\mathcal{H}_i$ continues to bound attention, while $\mathcal{M}_i$ stores what should remain after the FIFO forgets.

Leak-free prefix. The construction of C_i follows the same constraint as Eq. (28). Every index in C_i satisfies $t < t_i^{\text{start}}$, so the write never observes the current clean target or any future chunk. The prefix is ordered from oldest to newest and is concatenated in front of the noisy target, so a left-to-right scan matches the chunk order of streaming generation. First-chunk windows that do not contain such a history are dropped during training. Retaining them forces a first-frame fallback and would allow $\mathcal{F}_{\psi}$ to condition on a target that is still being denoised, which is the analogue of letting $\mathcal{B}_{i,\text{noisy}}$ attend to $\mathcal{B}_{i,\text{clean}}$.

Target-first rebase. As the FIFO slides, we already rebase temporal RoPE on $\mathcal{H}_i^m$. Camera and trajectory encodings that travel with C_i are rebased in the same way, using the first pose of the current chunk as the origin rather than the origin of chunk $i - 1$. Without this rebase, $\mathcal{M}_{i-1}$ and C_i would remain in the previous chunk’s coordinate system, and the trajectory pathway ϕ would observe an artificial jump at every chunk boundary even when the underlying trajectory is smooth.

Block-wise recurrent write. $\mathcal{F}_\psi$ is realized as a sparse state-space scan attached to selected DiT blocks, rather than as a global memory token in text cross-attention. For each spatial token trajectory the module maintains a state s_t and updates

$$s_t = \sigma(\delta) \odot s_{t-1} + (1 - \sigma(\delta)) \odot \tilde{u}_t, \quad y_t = g_t \odot s_t, \quad (42)$$

where δ is a learned per-channel decay, $\tilde{u}_t$ is a bounded input projection of the normalized block features, and g_t is a sigmoid write gate. The scan runs over the concatenated prefix and target, so the residual at the start of chunk i already contains a summary of C_i . The scan output is added through a residual whose scale is confined to a small interval. The lower bound prevents the recurrence from collapsing to a no-op, and the upper bound prevents it from overwriting the pretrained backbone. After V_i is generated, we retain only the last few latent-aligned steps as C_{i+1} and take the terminal scan state as $\mathcal{M}_i$. Long-video variants continue to place text and optional image conditions in $\mathcal{D}_i$, and world-model variants continue to condition on ϕ . Causal memory does not replace those inputs. It is the component of the history that should remain available after $\mathcal{H}_i$ slides.

The three components admit natural alternatives. C_i can be replaced by a suffix, a nearest-first tail, or FoV retrieval. Actions can keep the previous chunk’s coordinates. $\mathcal{F}_\psi$ can be omitted, given an unbounded residual, or trained on first-chunk windows that have no real prefix. We compared these variants under the same backbone, chunk length, and SGF schedule. The configuration above, leak-free prefix, target-first rebase, bounded block-wise write, and history-bearing training windows, is the one that worked best. It does not change Eq. (40). It only specifies what $\mathcal{H}_i$ contains and how much of that history is written into $\mathcal{M}_i$ before the sink-plus-FIFO window slides.

6 Agentic Planning and Control

An overview of our Director Agent workflow is shown in Fig. 4, covering short-video generation and the shot-wise long-video generation loop with prompt enhancement, memory retrieval, review, and memory admission.

Long-form video creation begins with an open-ended user intention, whereas a video generator requires concrete descriptions of what happens in each shot. A short request rarely specifies a complete narrative, stable character identities, scene transitions, dialogue, camera progression, or which earlier visual evidence should be reused. We therefore introduce a Director Agent that organizes generation as a persistent, shot-wise process. The agent develops the user’s intention into a screenplay, applies prompt enhancement (PE), manages visual-audio memory, and coordinates generation and local revision. The video model remains responsible for synthesis, while the agent determines what should be generated and which context should condition it.

A central goal is autonomous completion. After receiving a brief and any optional reference media, the agent can understand the request, plan the story, prepare every shot, generate the synchronized long audio-video sequence, check the intermediate results, update continuity memory, and assemble the final film without waiting for a human decision at every shot. This is the default autonomous path. The same path exposes human-in-the-loop checkpoints: a user can act as a critic, edit the screenplay or current prompt in real time, reselect a memory, and resume generation from the affected shot.

The workflow is organized around a persistent project workspace. The workspace contains the user request, the evolving screenplay, a compact story profile, the ordered shot plan, generated artifacts, the character and scene memory banks, and the decisions made during review. It also records the relation between neighboring shots and the references supplied to each generation request. This state is important for long videos: a later shot is conditioned on accepted visual evidence and on the user’s latest decisions, rather than on a

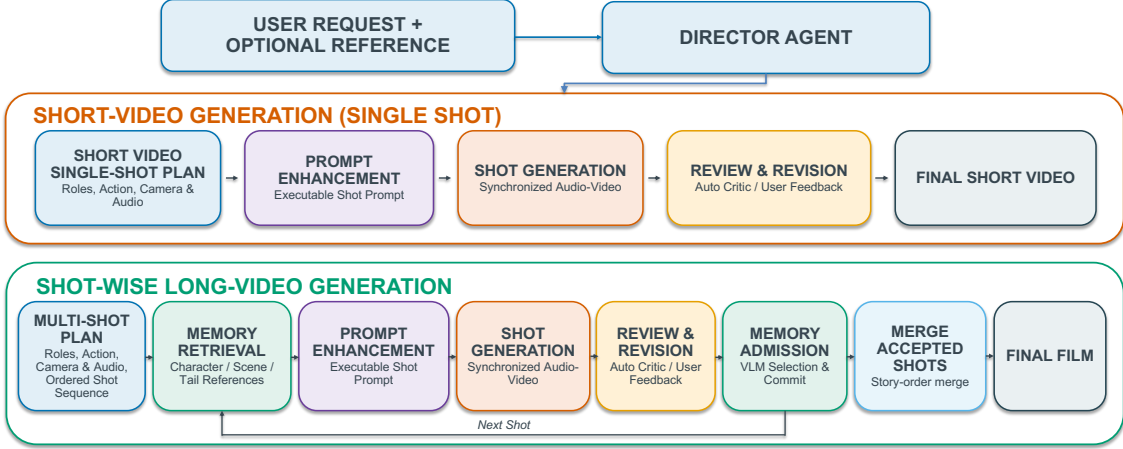


Figure 4 Overview of the Director Agent workflow. Given a user request and optional reference, the agent performs shot planning and prompt enhancement (PE), and supports both single-shot short-video generation and iterative long-video generation through memory retrieval, shot-level review and revision, and memory admission.

reconstruction of the conversation or on an unverified plan. The workspace can therefore be resumed from an intermediate shot, while a revision changes only the affected shot and its downstream state.

Recent agentic workflows have been explored for long-video generation [15]. Our workflow separates three responsibilities that are easy to conflate in a long-video pipeline. Planning decides what the story should contain and how it is divided into observable beats. Generation turns one prepared beat into synchronized audio and video under the selected references. Review and memory admission decide which result is accepted and which visual evidence is safe to expose to later shots. Keeping these decisions explicit makes the workflow inspectable: a failed generation can be revised without silently changing the screenplay, and an unsuitable memory proposal can be replaced without altering the accepted shot.

6.1 Planning and Prompt Enhancement

Read the brief and develop the screenplay. Given a user instruction u , the agent first resolves the central event, characters, locations, temporal progression, and intended ending. It produces a complete screenplay before the first generation call, but keeps the individual shot descriptions editable. The screenplay establishes the narrative order, one main dramatic job for each outer shot, and the relation between adjacent shots. An outer shot is the unit submitted to the video generator; any internal camera changes needed to expose one beat remain part of that shot.

The planning record contains a compact story profile in addition to prose. It gives every recurring subject a stable identity, appearance, wardrobe, and voice anchor, and records the scene, important props, viewpoint, and relevant sound context. Stable subject IDs are reused across all later captions. This allows a later mention of a character to refer to an established identity while still permitting the shot to specify a new pose, action, or camera relationship. The profile also distinguishes persistent facts from shot-local facts. For example, a character’s coat and voice can persist, whereas a wet sleeve or a temporary facial reaction belongs only to the shot in which it is observed.

Prepare the shot plan. For every adjacent pair, the planner marks one of three continuity relations: a scene transition, an ordinary edit within the same scene, or one continuous action. A scene transition changes the location or the established spatial state and starts with fresh scene context. An ordinary edit keeps the scene but changes the viewpoint or shot composition. A continuous action preserves the terminal state and motion direction of the previous shot. This relation is written into the shot plan before generation and is later used consistently when choosing a representative scene frame, a true tail frame, and the image references described in the caption.

Construct the production prompt. The screenplay and the PE captions are generated as one planning stage:

$$C = \{c_t\}_{t=1}^T = \text{PE}(\mathcal{D}(u)). \quad (43)$$

Here $\mathcal{D}$ produces the narrative plan, c_t is the enhanced caption for outer shot t , and T is the number of planned shots. We use $S_t = (V_t, A_t)$ for the synchronized shot, following Sec. 3.1. The equation describes the interface between planning and generation rather than a new learning objective: the agent writes the conditions and the video model samples the corresponding shot.

Structured caption fields. PE treats a caption as a set of coordinated fields rather than as an undifferentiated paragraph. For shot t , we write the organization abstractly as

$$c_t = [r_t^{\text{story}}; r_t^{\text{visual}}; r_t^{\text{camera}}; r_t^{\text{audio}}; r_t^{\text{temporal}}], \quad (44)$$

where the five terms describe the narrative beat, visible subjects and environment, camera and composition, synchronized sound, and the ordered changes within the shot. The fields are written as natural language in the final prompt; the notation only makes their roles explicit.

The story field contains the subjects, action, dialogue, and required ending. The visual field records appearance, clothing, props, scene geometry, lighting, and style. The camera field specifies shot scale, viewpoint, composition, and motivated movement, while the audio field aligns dialogue, Foley, ambience, and music with the visual event. The temporal field gives the initial state, trigger, preparation, continuous action, completion, reaction, and final hold. The fields preserve our identity anchors, scene relations, and audio-visual memory roles while keeping the production plan readable at each stage.

Executable detail and grounding. PE preserves the hard invariants of the request—subjects, identities, props, dialogue, language, duration, and outcome—and expands each beat as initial state $\rightarrow$ trigger $\rightarrow$ action $\rightarrow$ completion $\rightarrow$ reaction $\rightarrow$ final hold. It adds only observable evidence that improves execution, such as contact mechanics, environment response, or material-specific sound; it does not introduce a new event, subject, or outcome merely to make the caption longer.

Camera, sound, and references. The camera field is chosen to expose the evidence: an establishing view gives geography, a closer view gives expression or manipulation, and a motivated reframe reveals a consequence. Dialogue, Foley, ambience, and music follow the same timeline. Identity references provide appearance and voice cues, scene references provide layout and lighting, and an opening tail provides the starting state for continuous action. PE names only references supplied to the current shot, keeping text and media conditions consistent.

Initial and image-aware PE. An initial reference image grounds the screenplay in its visible people, objects, composition, viewpoint, and lighting. Immediately before an image-conditioned generation, a second PE pass describes how that state evolves into the planned beat. It preserves subject IDs, dialogue, language, and reference roles while adapting the opening pose and motion. The same rule is used for a user-provided first frame and for the tail frame of an accepted shot. The user may revise the screenplay and shot count before generation; afterward, the confirmed anchors constrain local prompt edits.

6.2 Short-Video Generation

Short-video generation uses the same planning, PE, optional image conditioning, generation, and user revision procedure as the long-video workflow, but treats the request as a single-shot screenplay ($T = 1$). The agent still resolves the event, starting state, ending state, camera observation, and synchronized sound; the difference is that no later shot needs to inherit the result. For a fresh text-to-audio-video request, the enhanced caption is sent directly to the generator. The generated audio and video are sampled together so that the action, speech, Foley, and ambience follow one timeline.

With a user-provided first-frame reference, a multimodal model interprets the visible pose, viewpoint, environment, lighting, and terminal action state. It performs a local image-aware rewrite of the enhanced

caption: the opening pose and state-dependent action are adjusted to follow naturally from the frame, while the story beat, subject identities, dialogue, language, and reference roles remain fixed. The first frame is then passed as the image condition. A fresh shot has no image condition; an image-conditioned shot begins from the provided visual state. The same interface supports the audio-visual model’s text, image, and memory conditioning without introducing a separate short-video prompt format.

In autonomous mode, the agent completes the short-video path using its default review and delivery policy. In human-in-the-loop mode, the rendered clip is shown to the user as a critic checkpoint. The user can provide a concrete correction—for example, an appearance, action, dialogue, camera, sound, or continuity change—and the agent rewrites the current caption before regenerating it. Since there is no following shot, the accepted short clip does not update a cross-shot character bank. This keeps the single-shot path simple while preserving the same real-time revision semantics used by the long-video path.

6.3 Shot-Wise Long-Video Generation

Autonomous execution. The agent executes the confirmed screenplay in timeline order and prepares the next shot only after the current shot has been resolved. Auto Mode runs this loop from the brief to the final merge with default review and memory policies. The interactive mode exposes the same states to a human critic, who can pause the current shot, modify its prompt or memory choice, and resume from there.

Memory and shot condition. The character memory bank is partitioned by stable character ID, $B_{t-1} = \{B_{t-1}^{(k)}\}_k$, and stores an approved visual observation, paired audio evidence, and its source shot. A separate full-frame scene record preserves same-scene layout; a true tail frame is reserved for continuous action. Scene transitions start with fresh scene context.

Before generating shot t , the agent retrieves the entries for the identities in c_t . Let $m_{t-1}^{(k)}$ denote the selected entry for character k , and let P_t denote the same-scene representative; $P_t = \emptyset$ after a scene transition. The active memory is

$$M_{t-1} = \left\{ m_{t-1}^{(k)} \mid k \in \text{IDs}(c_t) \right\} \cup P_t. \quad (45)$$

Here $\text{IDs}(c_t)$ lists the current subjects. Character entries provide identity and speaker cues, while P_t provides scene context. If the user requests another existing memory, the agent rebuilds M_{t-1} and regenerates only the current shot.

Image-aware shot preparation. The agent supports fresh generation, first-frame generation, same-scene reference generation, and continuous generation. Let I_t be the optional image condition: a user first frame, the accepted previous tail, or $\emptyset$ for a fresh shot. Image-aware PE rewrites only the opening state needed to connect c_t to I_t and preserves the story beat, subject IDs, dialogue, and reference roles. The shot is sampled as

$$S_t \sim p_\theta(\cdot \mid \tilde{c}_t, M_{t-1}, I_t), \quad S_t = (V_t, A_t), \quad (46)$$

where $\tilde{c}_t$ is the image-aware caption, M_{t-1} is active memory, and I_t fixes the starting visual state.

Layered review and real-time intervention. After generation, Auto Mode continues with its default policy. In the interactive path, the user can accept the clip or provide critic feedback on appearance, action, dialogue, camera, sound, or continuity. The agent edits the current caption or active memory and regenerates only that shot; rejected artifacts do not enter later context.

Memory admission after shot acceptance. After an accepted shot, candidate frames and synchronized audio are collected. A VLM proposes a subject crop and a scene representative; Auto Mode resolves the proposal by default, while the user may approve or reselect it. Character crops enter the corresponding bank, the scene frame remains continuity-only, and the true tail is kept for continuous generation.

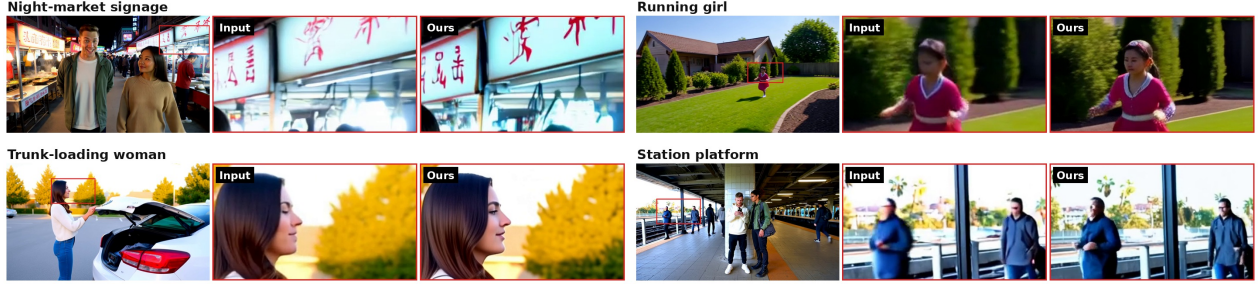


Figure 5 Qualitative effect of the SR stage on generator outputs, concentrated on the two content types that degrade first at the generator’s native resolution: small text and small faces. Smeared strokes on distant storefront signage become legible glyphs, and faces only a few dozen pixels across—including under motion—recover defined features rather than the coarse blobs the generator produces. Each cell shows the generator output with the inspected region boxed (left), followed by that region magnified before and after the SR stage.

Let $\hat{m}_t^{(k)}$ denote the resolved character memory for identity k after the automatic policy or the user’s review. The bank update can then be written as

$$B_t^{(k)} = \begin{cases} \text{Update}(B_{t-1}^{(k)}, \hat{m}_t^{(k)}), & \hat{m}_t^{(k)} \text{ is admitted,} \\ B_{t-1}^{(k)}, & \text{otherwise,} \end{cases} \quad (47)$$

where admission is decided automatically or by the user after re-selection. An unresolved proposal leaves the previous bank unchanged.

Assemble and deliver. After memory resolution, the agent prepares the next shot and repeats the loop until all planned shots are accepted. It then merges them in screenplay order with synchronized audio. Auto Mode delivers the completed film directly; interactive mode permits a final whole-film review before delivery.

7 Arbitrary-Step Super-Resolution

The generators of the preceding sections produce synchronized audio-visual content at the 720-class resolutions of the training corpus (Sec. 2.1.2); a dedicated super-resolution (SR) stage then restores fine texture and suppresses generation artifacts at the delivery resolution. This factorization concentrates the dominant inference cost in the SR stage, which evaluates a large backbone at full output resolution on every frame; with resolution and duration both fixed by the deliverable, the number of network evaluations is the only lever left. Few-step distillation pulls that lever at training time, baking a single step count into the weights [34, 45, 46], while training-free ODE solvers [21, 51] stay flexible but degrade sharply in the very-few-step regime that matters most here. We instead distill a flow-matching SR teacher [18, 38] with MeanFlow [8] into a model of the average velocity over an interval, turning the step count into an inference-time dial: one set of weights serves N -step sampling for any N , with no retraining and no external solver. Fig. 5 illustrates the effect of this stage on generator outputs.

7.1 SR as Conditional Generation

Conditioning by latent concatenation. We cast video SR as conditional restoration in the latent space of the Wan2.1 VAE [38], following the architecture-preserving T2V-to-TV2V super-resolution paradigm of Luxury et al. [25]. The low-resolution (LR) condition occupies the same latent grid as the target, and its latent z^{LR} is concatenated with the noisy target latent along the channel dimension at the network input, so that the model’s task reduces to re-synthesizing the high-frequency content the condition does not contain. To discourage the trivial shortcut of copying the condition, the condition is stochastically corrupted,

$$y = (1 - \alpha)z^{\text{LR}} + \alpha\epsilon_y, \quad \alpha \sim \mathcal{U}(0.3, 0.6), \quad \epsilon_y \sim \mathcal{N}(0, \mathbf{I}), \quad (48)$$

with α redrawn on every forward pass; the same perturbation is applied at training and sampling.

AIGC-oriented degradation. Because this SR stage is cascaded after a generator rather than applied to camera footage, its condition distribution is dominated by generation artifacts rather than sensor noise. Classical blind-SR pipelines [1, 39] synthesize blur, noise, and compression independently per frame and are therefore mismatched, whereas diffusion-generated video fails through temporally structured modes such as flicker, motion jitter, and color shift. We adopt the AIGC-oriented degradation design of Luxury et al. [25], prepending temporally coherent degradations whose parameters follow low-frequency trajectories in time, and annealing the overall strength toward a weak final setting matched to the upstream generator. On the target side, an in-batch perceptual-quality gate keeps supervision clean by replacing rejected clips with temporally flipped copies of retained ones.

Text-condition annealing. The teacher is adapted from a text-to-video (T2V) backbone into a text-and-video-to-video (TV2V) restorer, and we want the resulting model to super-resolve faithfully without relying on a caption at inference: a delivery-time SR stage should not require the upstream prompt to be available, nor should its fidelity depend on how well that prompt is written. Removing the caption from step 0 is nonetheless the wrong way to obtain this, because the backbone’s cross-attention and modulation pathways were pretrained exclusively on caption-conditioned generation; forcing them out of that regime at the same moment the newly added condition pathway is being learned makes two adaptations compete. We therefore adopt the prompt-annealing schedule of Lu et al. [22]: the probability of using the real caption decays linearly to zero over the early phase of training, after which a single fixed restoration prompt serves every sample. Annealing separates the two adaptations in time—the text pathway stays in its pretrained regime while gradient signal concentrates on the condition projection, and only then does the text distribution collapse onto the fixed prompt. The resulting model is caption-free at inference while retaining the generative prior that supplies detail the LR input does not contain.

Teacher objective. The teacher is trained with the standard conditional flow-matching objective [18], consistent with the convention of Eq. (27): for a clean latent x_0 , noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, and noise level $t \in (0, 1)$, the noisy latent is $x_t = (1 - t)x_0 + t\epsilon$ and the network regresses the straight-path velocity,

$$\mathcal{L}_{\text{SR}} = \mathbb{E}_{x_0, y, t, \epsilon} \left[\|v_\theta(x_t, t, y) - (\epsilon - x_0)\|_2^2 \right], \quad (49)$$

where t is drawn from a logit-normal distribution warped by a time-shift transform that concentrates training density at higher noise levels, y is the noise-perturbed condition of Eq. (48), and the text condition follows the annealing schedule above. The teacher is obtained by approximately 10.4k optimizer steps of SR fine-tuning through a staged curriculum over degradation strength and resolution, ending on multi-aspect-ratio buckets that match the intended delivery resolutions. At inference, this teacher is a conventional multi-step flow model: producing a single output requires tens of solver steps [51], each a full forward pass at the output resolution, which is precisely the cost that the next subsection removes.

7.2 Distilling an Arbitrary-Step Sampler

Average velocity. MeanFlow [8] replaces the instantaneous velocity $v(x_t, t)$ with the average velocity over an interval $[r, t]$,

$$u(x_t, r, t) = \frac{1}{t - r} \int_r^t v(x_\tau, \tau) d\tau, \quad 0 \leq r < t \leq 1, \quad (50)$$

so that the exact solution of the probability-flow ODE over $[r, t]$ is the single subtraction $x_r = x_t - (t - r)u(x_t, r, t)$. A network $u_\theta(x_t, r, t)$ that models this quantity is an interval integrator by construction: it does not approximate a step of an ODE solver, it is the step, for any interval it is queried on.

MeanFlow identity and training objective. Differentiating Eq. (50) along the trajectory yields the MeanFlow identity

$$u(x_t, r, t) = v(x_t, t) - (t - r) \frac{d}{dt} u(x_t, r, t), \quad \frac{du}{dt} = \partial_x u \cdot v + \partial_t u, \quad (51)$$

which converts the intractable interval integral into a local regression target: the known conditional velocity $\epsilon - x_0$ corrected by a Jacobian–vector product (JVP) of the network itself, evaluated under stop-gradient. We train with an explicit-gradient form of this regression,

$$\mathcal{L}_{\text{MF}} = \mathbb{E} \left[\|u_\theta - \text{sg}(u_\theta + w \Delta)\|_2^2 \right], \quad \Delta = (\epsilon - x_0) - u_\theta - \lambda(t - r) \widehat{\frac{du}{dt}}, \quad (52)$$

where $\text{sg}(\cdot)$ denotes stop-gradient and $\widehat{du/dt}$ is a stop-gradient estimate of the total derivative, so that the gradient with respect to u_θ is $-2w\Delta$ and the update direction is set directly. The per-sample weight $w = 1/(\|\Delta\| + c)$ is computed under stop-gradient—a gradient-carrying weight would open a backpropagation path that rewards enlarging the loss to shrink the weight—and down-weights high-noise samples where the JVP estimate is least stable. The tangent norm entering w is rescaled to be invariant to spatial and temporal tensor shape, which keeps a single c valid across all resolution buckets. The factor λ linearly ramps from 0 to 1 over the first 2,000 steps, interpolating smoothly from pure flow matching to the full MeanFlow objective while the network acquires its interval dependence. The pair (t, r) is drawn as the maximum and minimum of two independent samples from the same shifted logit-normal family used for teacher training (time-shift 2.0 in the distillation stage), so no new timestep schedule is introduced at distillation time and every sample trains the genuine interval objective ($t > r$).

Jacobian–vector products by finite differences. Exact forward-mode JVPs are unavailable in our training stack: fused attention kernels do not implement forward-mode rules, non-reentrant activation checkpointing does not propagate tangents through the recomputed forward, and the sharded data-parallel backward does not support them. We therefore estimate du/dt by a central finite difference along the ODE trajectory: perturbing (x_t, t) jointly by $\pm h$ in the direction $(v, 1)$ captures both the $\partial_x u \cdot v$ and $\partial_t u$ terms of Eq. (51) in a single difference. The two probe evaluations run without gradient tracking, with network outputs kept in full precision and the differencing performed in double precision, since half-precision mantissas are destroyed by the near-cancelling subtraction; both probes share an identical random state and a pinned condition tensor y , so that the stochastic conditioning noise of Eq. (48) cannot leak into the difference. One-sided differences are substituted per sample when the interval boundary would be crossed. The step size h was calibrated by sweeping against a higher-precision reference evaluation of the production model—an internal engineering measurement at a single operating point rather than an ablation—and the chosen value sits in the regime where truncation and round-off errors balance; the adaptive weight and warmup in Eq. (52) absorb the residual magnitude error, as the objective chiefly consumes the tangent’s direction. Measured on our configuration, the estimator costs $2.0\times$ the step time of plain flow matching and no additional activation memory, since peak memory remains set by the single gradient-carrying forward.

Arbitrary-step sampling. The distilled model needs no scheduler, no solver, and no per-budget retraining. Given any budget of K network evaluations, we partition the noise axis $1 = t_0 > t_1 > \dots > t_K = 0$ by warping a uniform grid through the same time-shift transform used in training,

$$t_k = \frac{s(1 - \frac{k}{K})}{1 + (s-1)(1 - \frac{k}{K})}, \quad k = 0, \dots, K, \quad (53)$$

with shift $s = 2.0$ and the endpoints fixed to the maximal training noise level and to zero, which concentrates the few available evaluations in the high-noise region where training density was placed, and iterate the exact interval update

$$x_{t_{k+1}} = x_{t_k} - (t_k - t_{k+1}) u_\theta(x_{t_k}, t_{k+1}, t_k), \quad k = 0, \dots, K-1. \quad (54)$$

The sampler consists solely of the grid of Eq. (53) and the subtraction in Eq. (54); no scheduler object exists. Setting $K = 1$ gives one-step generation; larger K trades computation for fidelity along a continuum, and during training we render $K \in \{1, 2, 4\}$ side by side from the same checkpoint at every validation. This is the deployment property we set out to obtain: the SR cost per output pixel becomes K forward passes with K chosen at request time, so a single model serves latency-critical interactive use at one step and quality-leaning offline rendering at higher step counts, replacing the tens of solver steps required by the teacher. We emphasize what this is not. It is not a fixed-step distilled student in the sense of Secs. 3.2 and 5.2, whose

DMD objectives [45, 46] commit to a specific sampler at training time; and it is not a model to be driven by a multi-step corrector [21, 51], which would be both redundant—the subtraction in Eq. (54) is already the exact interval integral—and incorrect, since such correctors assume the network predicts the instantaneous velocity v , whereas u_θ predicts the average velocity $u \neq v$.

7.3 Warm-Starting Without Loss

The degenerate diagonal. MeanFlow contains flow matching as its boundary case: when $r = t$, Eq. (50) reduces to $u(x_t, t, t) = v(x_t, t)$ and the target in Eq. (52) degenerates exactly to the flow-matching target of Eq. (49). The converged multi-step teacher therefore already solves the MeanFlow objective on the $r = t$ diagonal of the (r, t) plane, and distillation amounts to extending a function the teacher has learned on a line to the triangle above it—not to training a different model that must first re-approach the teacher.

Zero-initialized interval branch. To realize this extension without disturbing the teacher, the student architecture augments the backbone with a second conditioning branch that mirrors the timestep pathway: a sinusoidal embedding of the interval length $t - r$, passed through its own MLP and its own modulation projection, whose outputs are added to the corresponding timestep-branch activations. Feeding $t - r$ rather than r makes the degenerate case correspond to a fixed zero input, and keeping the projections separate—each branch projected before summation—decouples the new branch’s gradients from the pretrained modulation weights. The terminal linear layer of each new pathway is initialized to zero, so at step 0 the branch contributes exactly nothing and the student’s forward pass reproduces the teacher’s by construction; a start-of-training self-check verifies that the loss pipeline degenerates to flow matching when $r = t$, confirming the wiring. The warm start is thus lossless: no adapter distillation, no feature matching, and no re-convergence phase are needed.

Curriculum and confound control. Distillation is full-parameter and inherits the teacher’s entire conditioning stack unchanged—the latent concatenation and the condition noising of Eq. (48), the degradation pipeline, the paired HR/LR data, and the per-stage resolution buckets—so that any behavioral difference between student and teacher is attributable to the objective rather than to a data-distribution shift. Training proceeds in two stages: a first stage on moderate-resolution buckets initialized from the converged teacher, followed by a continuation on the teacher’s own high-resolution multi-aspect buckets with the optimizer state carried over; the shape-invariant tangent normalization of Sec. 7.2 is what allows this resolution change without retuning the loss weighting. The optimizer follows the teacher’s settings except for a raised second-moment coefficient (β_2 from 0.95 to 0.99), accommodating the higher gradient variance of the finite-difference target, and the high-resolution stage adds a cross-rank finiteness gate that skips optimizer steps on non-finite gradients before clipping, preserving the Adam state through transient instabilities and closing a failure mode in which a NaN gradient silently bypasses standard norm clipping because the comparison `nan < 1.0` evaluates to false. Combined with the tangent warmup, the overall trajectory is a continuous deformation of the teacher: the student begins as the teacher, is annealed onto the MeanFlow objective, and ends as a single set of weights for which Eq. (54) holds at every step budget.

8 Results

We evaluate JoyAI-Echo-1.5 from two complementary perspectives: long-horizon audio-visual generation and action-conditioned world modeling. For world modeling, we report results on two public benchmarks, WBench and SANA-WM-Bench, together with a large-scale human preference study.

8.1 Long-Horizon Audio-Visual Generation

Experimental Setup. We evaluate long-horizon audio-visual generation on a curated benchmark of 100 stories and 3,000 shots, following the protocol of JoyAI-Echo-1.0 [17]. We consider four complementary aspects:

- **Cross-shot consistency.** We report pairwise ViCLIP similarity [40] for visual-semantic consistency, Self-CIDS for character identity consistency using GroundingDINO [20], Torchreid [52], and FaceNet [31], and 3DSpeaker similarity [3] for speaker consistency.

Table 2 Quantitative comparison of baseline methods. ViCLIP and Self-CIDS assess cross-shot visual and identity consistency, Voice assesses speaker consistency, and Recall measures speech-content coverage. All metrics are higher-is-better. Boldface and underlining indicate the best and second-best results, respectively.

Method	Cross-Shot Consistency			Video Quality		Text Consistency	Speech
	ViCLIP $\uparrow$	Self-CIDS $\uparrow$	Voice $\uparrow$	Aesthetic $\uparrow$	Imaging $\uparrow$	CLIP $\uparrow$	Recall $\uparrow$
JavisDiT++ [19]	0.6955	0.5621	0.7933	0.5019	0.6084	0.2522	0.0951
LTX-2.3 [10]	0.7047	0.6135	0.6847	0.5436	0.4751	0.2559	0.9304
LTX-2.5 [10]	0.7775	0.6640	0.7603	0.6019	0.5804	<u>0.2683</u>	<u>0.9545</u>
ShotStream+MMAudio [24]	0.7988	0.7308	0.7945	0.5255	0.6044	0.1998	0.0297
StoryMem+MMAudio [50]	0.7897	0.7491	0.7643	0.5265	0.6320	0.2368	0.0301
HappyOyster (Directing)	0.7535	0.6940	0.7705	0.5606	0.6701	0.2544	0.1085
JoyAI-Echo-1.0 [17]	<u>0.8026</u>	<u>0.7793</u>	<u>0.8129</u>	0.5679	<u>0.7058</u>	0.2658	0.9489
JoyAI-Echo-1.5	0.8264	0.7937	0.8524	<u>0.5705</u>	0.7467	0.2868	0.9674

- **Video quality.** Following VBench-Long [14], we concatenate all generated shots into a long video, split it into 2-second clips, and evaluate aesthetic and imaging quality.
- **Text consistency.** We measure alignment between the generated visual content and the input story script using CLIP score [28].
- **Speech-content fidelity.** We transcribe generated speech with Whisper [29] and compute word-level recall against the reference dialogue. We use recall instead of the accuracy metric in JoyAI-Echo-1.0 and re-evaluate all methods under the same protocol.

Unless otherwise specified, all JoyAI-Echo-1.5 results are generated with the Director Agent enabled for shot-level planning and memory selection.

Baselines. We compare JoyAI-Echo-1.5 with three classes of audio-visual generation systems:

- **Joint audio-visual generation:** JarvisDiT++ [19], LTX-2.3, and LTX-2.5 [10]. Each method generates approximately 10-second shots from the corresponding structured prompts, which are then concatenated for long-form evaluation. For LTX-2.3 and LTX-2.5, we enable prompt enhancement and condition each shot after the first on the final frame of the preceding shot.
- **Cascaded generation:** ShotStream [24]+MMAudio [4] and StoryMem [50]+MMAudio. The video model first generates a silent 10-second shot, after which MMAudio synthesizes audio from the corresponding shot-level audio description. The resulting shots are concatenated using the same story scripts and shot durations.
- **Native long-form generation:** HappyOyster in Directing mode,¹ which directly generates continuous audio-visual sequences from narrative scripts.

Quantitative Comparison. Tab. 2 compares JoyAI-Echo-1.5 with native audio-visual generators, cascaded long-video systems, and JoyAI-Echo-1.0. JoyAI-Echo-1.5 achieves the best performance on six of the seven metrics and ranks second on aesthetic quality. In particular, it obtains the highest ViCLIP (0.8264), Self-CIDS (0.7937), and Voice (0.8524) scores, improving over JoyAI-Echo-1.0 by 0.0238, 0.0144, and 0.0395, respectively. These results demonstrate substantially improved cross-shot visual and speaker consistency. Meanwhile, JoyAI-Echo-1.5 achieves the best Imaging (0.7467), CLIP (0.2868), and speech Recall (0.9674), while remaining competitive in aesthetic quality. This indicates that the gains in long-range consistency are achieved without sacrificing local generation quality, text alignment, or speech fidelity.

User Study. We further conduct a human evaluation comparing the long-video variant of JoyAI-Echo-1.5 with HappyOyster in Directing mode through randomized and anonymized side-by-side comparisons. Following the Good-Similar-Bad (GSB) protocol, annotators evaluate story-instruction following, dialogue/background-audio following, memory/identity consistency, audio-visual synchronization, and audio quality. For each

¹Details available at <https://www.happyoyster.cn/docs>

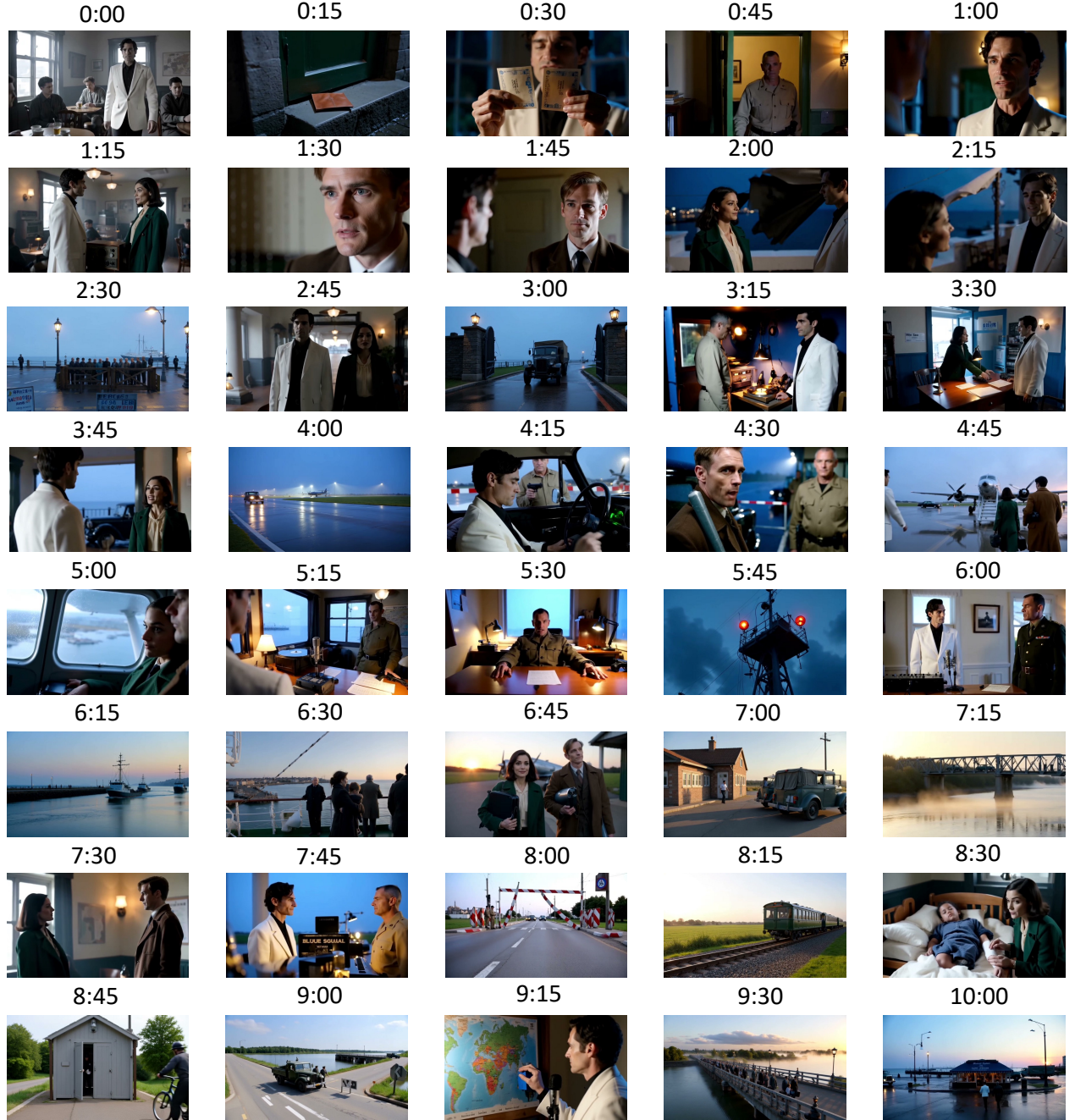


Figure 6 Qualitative visualization of a 10-minute long audio-visual sequence generated under the Director Agent. The 10-minute example follows a wartime evacuation story set in a neutral port city. A radio operator helps two civilians escape as military control tightens and the city gradually descends into chaos. Their escape becomes increasingly entangled with checkpoints, shifting alliances, and the growing danger faced by civilians left behind. The radio operator eventually turns his broadcast network into a broader evacuation effort, connecting scattered groups and helping them find a route out of the city.

dimension, they determine whether JoyAI-Echo-1.5 is better, the two results are similar, or HappyOyster is better.

As shown in Tab. 3, JoyAI-Echo-1.5 receives a higher preference rate than HappyOyster across all five



Figure 7 Qualitative comparison between JoyAI-Echo-1.5 and JoyAI-Echo-1.0 under the same long-form story. Frames sampled from temporally separated shots demonstrate the improved character consistency, scene stability, and visual rendering of JoyAI-Echo-1.5.

dimensions. The largest advantage is observed in audio-visual synchronization, where JoyAI-Echo-1.5 is preferred in 48.4% of the comparisons, compared with 20.3% for HappyOyster. JoyAI-Echo-1.5 also demonstrates clear advantages in story-instruction following (38.6% vs. 17.7%) and dialogue/background-audio following (39.2% vs. 23.5%).

For memory/identity consistency, JoyAI-Echo-1.5 is preferred in 24.8% of the comparisons, compared with 15.7% for HappyOyster, while 59.5% are judged similar. This indicates that both systems maintain identity adequately in many cases, but JoyAI-Echo-1.5 is more frequently preferred when perceptible differences occur. JoyAI-Echo-1.5 also achieves a higher preference rate in audio quality (44.4% vs. 33.3%). Overall, the user study demonstrates that the improvements measured by the automatic metrics remain perceptible in complete long-video generation, particularly in instruction following, audio-visual synchronization, and long-range consistency.

Qualitative Comparison. Fig. 7 compares JoyAI-Echo-1.5 with JoyAI-Echo-1.0 using the same long-form story and shot-level instructions. Compared with JoyAI-Echo-1.0, JoyAI-Echo-1.5 more reliably preserves the facial structure, hairstyle, clothing cues, and speaker identity of recurring characters across temporally distant shots. It also exhibits more stable exposure and color balance, with less accumulation of oversaturation, contrast artifacts, and scene drift over extended generation. These improvements are particularly visible after multiple scene changes, camera scale variations, and the introduction of new characters. In addition to the benchmark evaluation, we present a separate 10-minute full-system case study in Fig. 6. Despite the extended temporal horizon and frequent shot transitions, the generated sequence maintains recognizable character identities, coherent scene evolution, and stable visual quality across distant timestamps. This example demonstrates the system’s ability to sustain consistent long-form generation far beyond the duration of individual shots.

These examples illustrate the complementary roles of the model and the long-video generation pipeline. At the generator level, unified audio-visual Memory conditioning provides persistent identity and speaker

Table 3 Human evaluation based on Good–Similar–Bad (GSB) pairwise comparisons between JoyAI-Echo-1.5 (long-video variant) and HappyOyster for long-horizon audio-visual generation. Numbers denote the percentage of user preferences and may not sum to exactly 100% because of rounding.

Aspect	JoyAI-Echo-1.5	Similar	HappyOyster
Story-instruction following	38.6%	43.8%	17.7%
Dialogue/background-audio following	39.2%	37.3%	23.5%
Memory/identity consistency	24.8%	59.5%	15.7%
Audio-visual synchronization	48.4%	31.4%	20.3%
Audio quality	44.4%	22.2%	33.3%

cues while still allowing each target shot to follow its own textual description. The high-quality training corpus improves local rendering quality and broadens the model’s spatial and linguistic coverage, while transition-aware multi-shot captions strengthen its ability to represent changes in subject, action, camera view, and scene context around shot boundaries. At the system level, the Director Agent selects grounded Memory observations and revises the shot-level plan before subsequent generation, reducing the propagation of incorrect characters, actions, or scene attributes. Compared with JoyAI-Echo-1.0, JoyAI-Echo-1.5 therefore exhibits improvements at both local and long-range scales. Individual shots show clearer structure and more balanced rendering, while the full sequence exhibits less character drift, scene drift, and color accumulation over time. The qualitative results are consistent with the improvements in imaging quality, CLIP alignment, cross-shot visual consistency, and speaker consistency reported in Tab. 2. A more complete long-horizon visualization with densely sampled timestamps is provided in the supplementary material.

8.2 Action-Conditioned World Modeling

We evaluate two world-model variants of JoyAI-Echo-1.5: the original undistilled bidirectional multi-step model, denoted as **JoyAI-Echo-1.5**, and its distilled causal four-step variant, denoted as **JoyAI-Echo-1.5-Causal**. We benchmark these models on two public benchmarks, WBench [48] and SANA-WM-Bench [53], together with a subjective evaluation on our self-built World-Model Benchmark (WMB). The former evaluates broad interactive-world behavior, while the latter focuses on camera-trajectory following and long-horizon visual persistence.

8.2.1 WBench

Evaluation settings and metrics. We evaluate both JoyAI-Echo-1.5 and JoyAI-Echo-1.5-Causal on the official 158-case Navigation split of WBench [48], which covers first- and third-person interactive navigation. Following the official multi-turn evaluation protocol, we report five dimensions—Quality, Setting, Interaction, Consistency, and Physical—together with their overall Average score. Quality aggregates perceptual and temporal video-quality metrics; Setting evaluates scene and subject adherence; Interaction measures compliance with the requested interaction, including navigation trajectory following; Consistency aggregates spatial, geometric, photometric, perspective, subject, and segment-level consistency; and Physical measures visual plausibility and causal fidelity.

Quantitative Results. As shown in Tab. 4, both variants of our model rank at the top of the WBench Navigation leaderboard. The undistilled bidirectional multi-step JoyAI-Echo-1.5 achieves the highest overall Average score of 81.7, while the distilled four-step JoyAI-Echo-1.5-Causal ranks second with 81.0. This small performance gap indicates that causal few-step distillation largely preserves the interactive world-modeling capability of the original multi-step model.

JoyAI-Echo-1.5 achieves a Consistency score of 89.8, the second highest among all evaluated models, demonstrating strong world-state preservation across multi-turn interactions. Notably, JoyAI-Echo-1.5-Causal attains the highest Interaction score of 87.9, slightly surpassing the undistilled model at 87.2. Despite the transition from bidirectional multi-step generation to four-step causal autoregressive inference, the distilled model therefore retains strong control responsiveness while remaining competitive in overall generation quality.

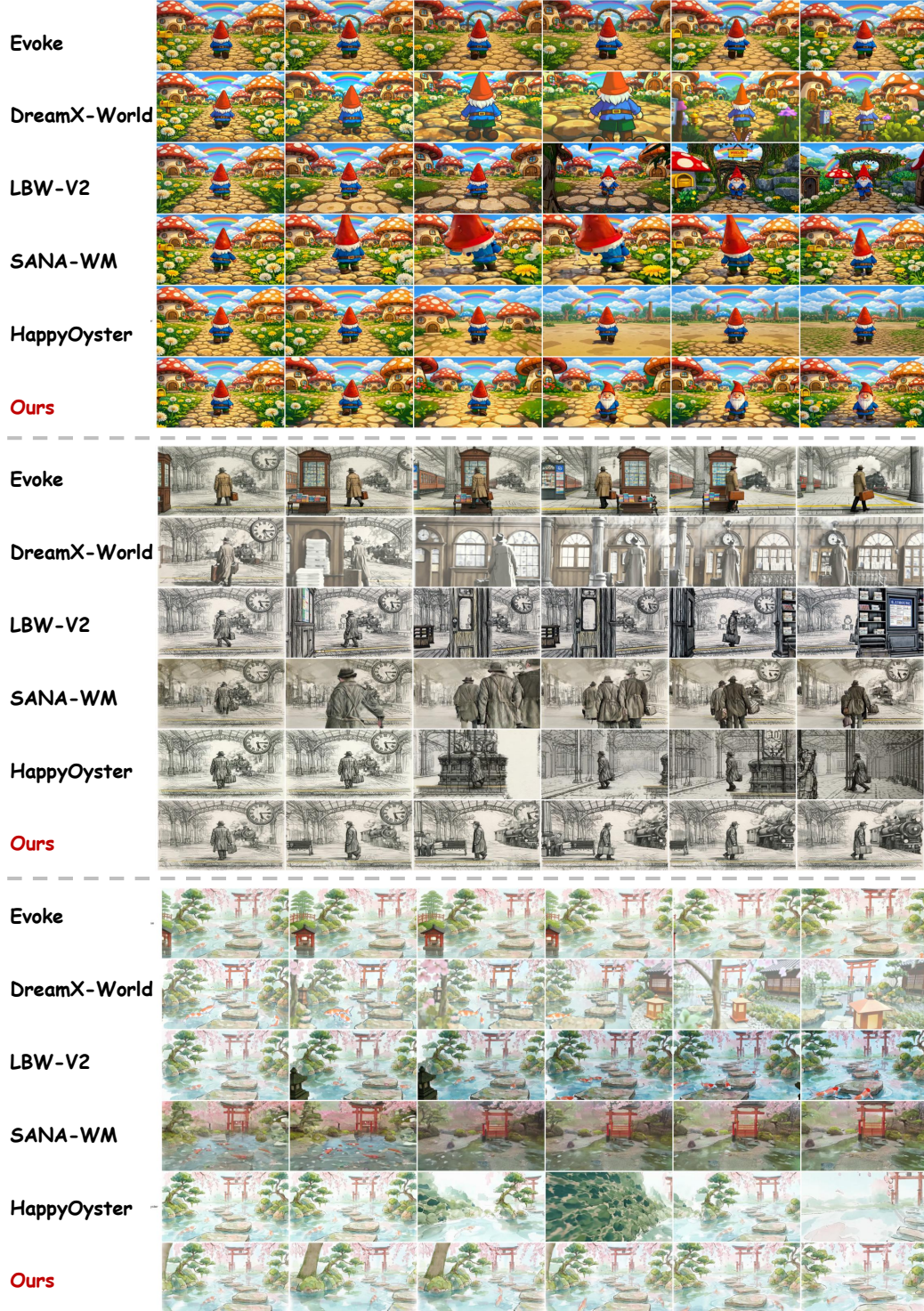


Figure 8 Qualitative comparison of the undistilled bidirectional multi-step JoyAI-Echo-1.5 on representative WBench Navigation cases across diverse visual domains. Given the same initial observation and navigation condition, our model better preserves scene identity and structure while producing more coherent viewpoint evolution over successive interaction steps.



Figure 9 Multi-turn comparison of JoyAI-Echo-1.5-Causal against representative few-step autoregressive world models on WBench Navigation. All models start from the same initial observation and follow the prescribed multi-turn action sequence. The three cases use (1) $A \rightarrow A \rightarrow A \rightarrow A$, (2) $\text{Left} \rightarrow \text{Left} \rightarrow \text{Right} \rightarrow \text{Right}$, and (3) $\text{Left} \rightarrow W \rightarrow \text{Right}$. Here, Left, Right, Up, and Down denote camera rotations, while W, S, A, and D denote camera translations. JoyAI-Echo-1.5-Causal exhibits stronger action adherence and better preserves scene style, subject identity, and world structure across successive turns, with less accumulated autoregressive drift.

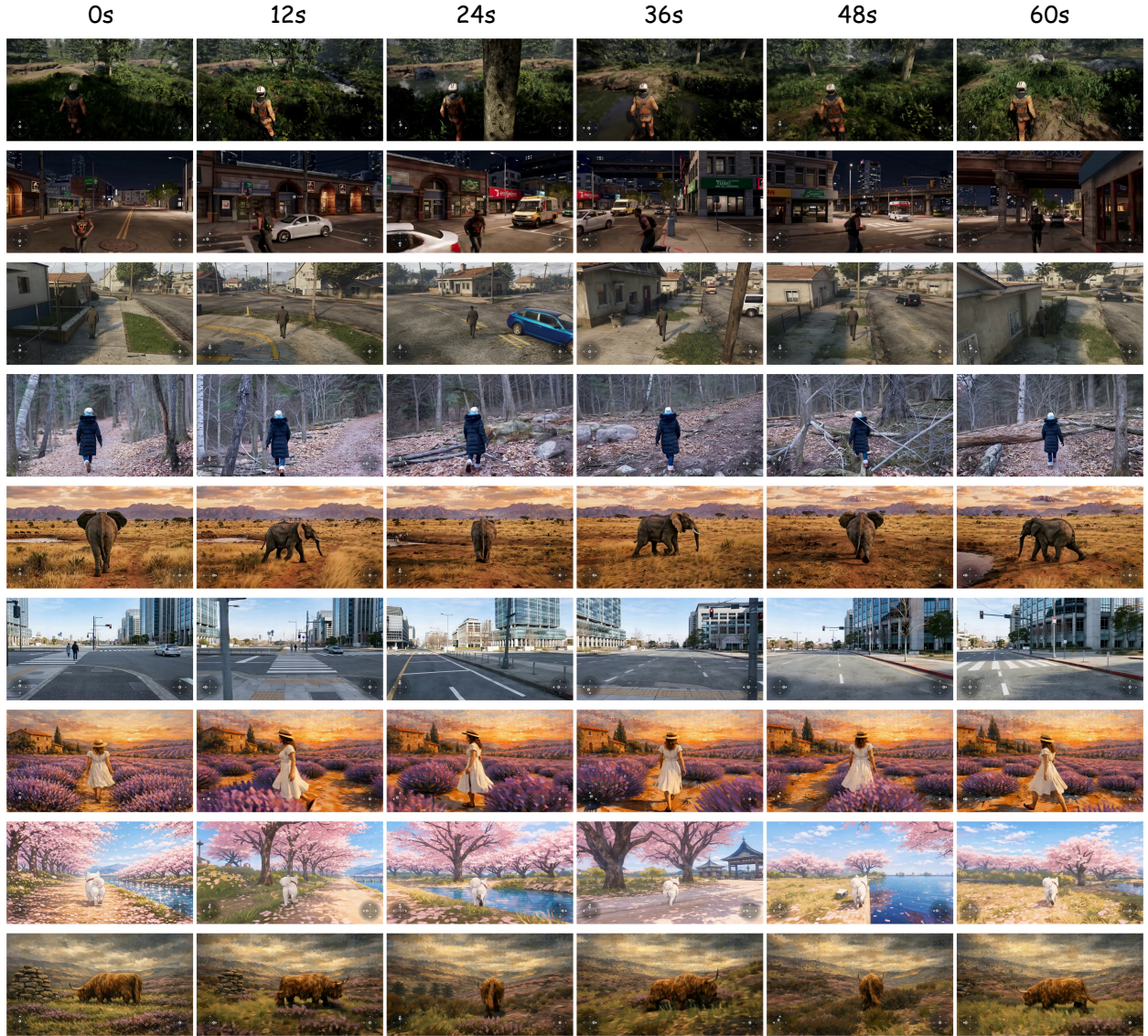


Figure 10 Causal long-horizon generation of JoyAI-Echo-1.5-Causal on representative cases. Each row shows a 60-second causal rollout sampled at different timestamps. Despite prolonged autoregressive generation, the distilled four-step causal model remains responsive to navigation controls while preserving scene identity, subject appearance, visual style, and major spatial structure, demonstrating robustness to long-horizon error accumulation.

Qualitative Results. Fig. 8 first compares the undistilled bidirectional multi-step JoyAI-Echo-1.5 with representative world models on WBench Navigation cases. Given the same initial observation and navigation condition, competing methods often exhibit abrupt viewpoint changes, unintended zooming, subject-scale variation, layout drift, or loss of scene identity. In contrast, JoyAI-Echo-1.5 produces more coherent world evolution across diverse visual domains, with stronger navigation adherence and better preservation of scene structure and appearance. These observations are consistent with its strong Interaction and Consistency scores in WBench.

We next compare causal few-step world models in Fig. 9, including DreamX-World, LingBot2-Fast, SANA-WM Streaming with and without Causal Refiner, Evoke, and our JoyAI-Echo-1.5-Causal. We evaluate three representative multi-turn interaction cases: $A \rightarrow A \rightarrow A \rightarrow A$, $Left \rightarrow Left \rightarrow Right \rightarrow Right$, and $Left \rightarrow W \rightarrow Right$, respectively. Here, Left, Right, Up, and Down denote camera rotations, whereas W, S, A, and

Table 4 Comparison on the 158-case WBench Navigation split [48]. JoyAI-Echo-1.5 denotes our undistilled bidirectional multi-step model, while JoyAI-Echo-1.5-Causal denotes its distilled four-step causal variant. Results are sorted by Average. Best and second-best results are shown in bold and underlined, respectively.

Model	Avg.	Quality	Setting	Interaction	Consistency	Physical
JoyAI-Echo-1.5	81.7	81.5	79.4	<u>87.2</u>	<u>89.8</u>	70.6
JoyAI-Echo-1.5-Causal	<u>81.0</u>	81.1	88.3	87.9	77.5	70.1
HiDream-O1-World [48]	80.7	81.9	81.9	79.5	88.0	72.1
LingBot-World (fast v2) [30]	79.4	81.8	76.8	82.8	86.5	69.1
Kling 3.0 [48]	79.0	81.4	91.0	69.4	83.7	69.3
LingBot-World (base-camera) [30]	78.5	78.9	72.6	80.1	89.9	71.2
Wan 2.7 [38]	78.1	81.5	<u>91.4</u>	64.4	81.6	<u>71.8</u>
HY-World 1.5 (AR-distill) [36]	78.1	78.1	72.2	86.8	86.9	66.3
HY-Video 1.5 [43]	77.9	77.6	85.6	71.4	87.4	67.4
LingBot-World (fast) [30]	77.4	79.4	77.9	79.2	84.9	65.7
HappyOyster [48]	76.8	77.3	74.2	84.9	84.3	63.5
Lyra 2.0 (4-step AR) [33]	76.4	77.1	73.2	85.6	79.3	66.7
Seedance 1.5 [32]	76.2	82.1	82.9	66.3	81.3	68.4
Cosmos3-Super [27]	76.0	76.4	91.6	58.0	85.0	69.2
SANA-WM (4-step AR) [53]	76.0	79.3	76.1	82.2	80.7	61.9
DreamX-World (5B AR) [6]	75.0	77.5	80.8	78.6	74.9	63.3
ABot-World [16]	74.7	76.8	71.4	84.0	79.5	61.7
Cosmos 2.5 [26]	74.6	72.9	83.3	63.1	86.5	67.4
Cosmos3-Nano [27]	74.2	77.4	87.3	59.1	83.7	63.6
LTX-2.3 [10]	74.2	77.1	85.2	66.4	77.2	64.9
Genie 3 [9]	73.9	75.2	72.5	73.4	82.6	65.7

D denote translational movements. Among the representative baselines, LingBot2-Fast and Evoke remain relatively competitive across multiple interactions, yet still exhibit noticeable action deviation together with gradual degradation in subject identity and visual style as the rollout progresses. Other baselines show more pronounced viewpoint drift, structural changes, or accumulated appearance inconsistency. In contrast, JoyAI-Echo-1.5-Causal follows the prescribed controls more accurately while better preserving scene layout, subject identity, and visual style across successive turns, demonstrating stronger multi-turn extrapolation and improved robustness to accumulated autoregressive errors.

We further examine the long-horizon extrapolation capability of the distilled causal model in Fig. 10. Over approximately 60 seconds of continuous causal streaming, JoyAI-Echo-1.5-Causal remains responsive to navigation controls while largely preserving scene identity, subject appearance, visual style, and major spatial structure across successive generated segments. The sustained coherence over this substantially extended horizon further indicates that the causal generator is robust to autoregressive error accumulation well beyond the training horizon.

8.2.2 SANA-WM-Bench

Evaluation settings and metrics. We evaluate SANA-WM-Bench under three complementary settings: (1) a short-horizon evaluation on the first 241 frames of each trajectory, (2) the undistilled multi-step model under the official 961-frame (60-second at 16 FPS) long-horizon protocol, and (3) the distilled causal model at 961 frames. The first setting provides a short-horizon reference for trajectory following and visual quality, the second follows the official long-horizon protocol to characterize the 60-second generation capability of the undistilled model, and the third evaluates whether the distilled few-step causal generator can retain strong long-horizon performance after distillation. For trajectory fidelity and visual quality, we follow the benchmark protocol and report rotation error (R , in degrees), relative translation error (T), camera-motion consistency

Table 5 Short-horizon comparison on SANA-WM-Bench using the first 241 frames of each trajectory. Each split contains 80 scenes. Best and second-best results within each split are shown in bold and underlined, respectively.

Method	Trajectory and Visual Quality			
	VBench $\uparrow$	$R\downarrow$	$T\downarrow$	CMC $\downarrow$
<i>Simple-Trajectory Split</i>				
Matrix-Game 3 [41]	79.41	4.474	0.406	0.453
LingBot-World-v2 [30]	81.13	3.092	0.282	0.310
HY-WorldPlay [36]	82.05	2.256	0.284	0.298
DreamX-World [6]	<u>82.42</u>	1.839	0.261	0.273
SANA-WM [53]	81.75	0.489	<u>0.194</u>	<u>0.195</u>
JoyAI-Echo-1.5	83.91	<u>0.523</u>	0.160	0.162
<i>Hard-Trajectory Split</i>				
Matrix-Game 3 [41]	79.37	12.819	0.583	0.717
LingBot-World-v2 [30]	81.67	6.367	0.357	0.426
HY-WorldPlay [36]	81.24	6.346	0.387	0.437
DreamX-World [6]	<u>82.03</u>	8.804	0.467	0.556
SANA-WM [53]	81.79	1.248	0.199	0.206
JoyAI-Echo-1.5	83.96	<u>1.697</u>	<u>0.256</u>	<u>0.266</u>

Table 6 Long-horizon comparison on the full SANA-WM-Bench trajectory under the official 961-frame setting (60 seconds at 16 FPS). For JoyAI-Echo-1.5, we report the undistilled multi-step model. Best and second-best results within each split are shown in bold and underlined, respectively.

Method	Trajectory and Visual Quality				Revisit Consistency			Temporal Quality
	VBench $\uparrow$	$R\downarrow$	$T\downarrow$	CMC $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	$\Delta IQ\downarrow$
<i>Simple-Trajectory Split</i>								
Infinite-World [44]	79.18	16.55	1.98	2.08	12.60	0.284	0.595	6.72
LingBot-World [30]	81.82	10.47	2.01	2.05	<u>14.59</u>	0.366	0.394	<u>0.04</u>
HY-WorldPlay [36]	68.82	17.89	2.36	2.45	12.83	0.321	0.616	23.59
Matrix-Game 3 [41]	78.53	12.96	1.83	1.92	12.29	0.326	0.553	2.41
SANA-WM [53]	79.29	<u>7.59</u>	<u>1.59</u>	<u>1.63</u>	14.16	0.333	0.458	3.79
JoyAI-Echo-1.5	<u>81.36</u>	3.22	0.918	0.927	15.10	<u>0.335</u>	<u>0.451</u>	0.02
<i>Hard-Trajectory Split</i>								
Infinite-World [44]	79.51	41.31	2.49	2.84	12.04	0.248	0.617	4.16
LingBot-World [30]	81.89	18.99	<u>1.65</u>	1.81	14.08	<u>0.332</u>	0.436	<u>0.58</u>
HY-WorldPlay [36]	70.46	35.46	2.34	2.64	13.72	0.328	0.654	25.88
Matrix-Game 3 [41]	78.79	18.79	1.67	1.82	12.17	0.317	0.556	0.32
SANA-WM [53]	79.60	10.02	1.66	<u>1.72</u>	<u>14.10</u>	0.327	<u>0.469</u>	3.09
JoyAI-Echo-1.5	<u>81.72</u>	<u>12.05</u>	1.265	1.358	14.58	0.333	0.482	0.93

error (CMC), and VBench Overall on a 0–100 scale. For the 961-frame evaluations, we additionally measure viewpoint revisit consistency using PSNR, SSIM, and LPIPS between generated frame pairs that revisit approximately the same camera pose, thereby evaluating long-term scene persistence without reference to ground-truth frames. We further report the temporal quality degradation $\Delta IQ = q_1 - q_6$, where q_1 and q_6 denote the VBench Imaging Quality scores of the first and last 10-second windows, respectively.

Quantitative Results. As shown in Tab. 5, in the 241-frame short-horizon setting, JoyAI-Echo-1.5 achieves the highest VBench Overall on both trajectory splits, reaching 83.91 on Simple and 83.96 on Hard. It also obtains the lowest translation and CMC errors on the Simple split, while remaining close to SANA-WM in rotation accuracy. On the Hard split, SANA-WM achieves lower trajectory errors, whereas JoyAI-Echo-1.5 maintains the highest overall visual quality.

As shown in Tab. 6, under the official 961-frame long-horizon setting, the undistilled multi-step JoyAI-Echo-1.5 remains highly competitive in both trajectory following and scene persistence. On the Simple split, it achieves a VBench Overall of 81.36, the lowest translation and CMC errors, and the highest revisit PSNR of 15.10dB. Although trajectory errors become more pronounced over the long rollout, the model maintains strong visual

Table 7 Long-horizon comparison of distilled causal world models on SANA-WM-Bench under the 961-frame setting (60 seconds at 16 FPS). JoyAI-Echo-1.5-Causal denotes ours 4-step causal model. Best and second-best results within each split are shown in bold and underlined, respectively.

Method	Trajectory and Visual Quality				Revisit Consistency			Temporal Quality
	VBench $\uparrow$	$R \downarrow$	$T \downarrow$	CMC $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	$\Delta IQ \downarrow$
<i>Simple-Trajectory Split</i>								
Evoke [47]	<u>80.11</u>	<u>9.859</u>	<u>2.159</u>	<u>2.190</u>	<u>13.50</u>	<u>0.2774</u>	<u>0.5010</u>	<u>-0.49</u>
LingBot-World-v2 [30]	79.20	22.516	2.391	2.541	11.90	0.1715	0.5906	2.59
SANA-WM + Causal Refiner [53]	79.55	17.875	2.222	2.338	14.41	0.2660	0.5748	-0.75
SANA-WM w/o Refiner [53]	78.21	15.359	2.288	2.393	9.19	0.1661	0.6321	-0.12
JoyAI-Echo-1.5-Causal	80.13	5.009	1.592	1.611	14.41	0.2998	0.4802	2.09
<i>Hard-Trajectory Split</i>								
Evoke [47]	<u>80.75</u>	20.097	<u>1.924</u>	<u>2.060</u>	13.06	0.2592	<u>0.5292</u>	<u>0.30</u>
LingBot-World-v2 [30]	78.74	34.416	2.125	2.413	12.34	0.2183	0.5732	3.84
SANA-WM + Causal Refiner [53]	79.60	23.794	1.932	2.124	<u>13.89</u>	<u>0.2605</u>	0.5696	-0.24
SANA-WM w/o Refiner [53]	78.27	<u>18.717</u>	2.006	2.142	9.26	0.1729	0.6077	2.06
JoyAI-Echo-1.5-Causal	81.06	14.221	1.735	1.829	14.04	0.2910	0.4893	1.30

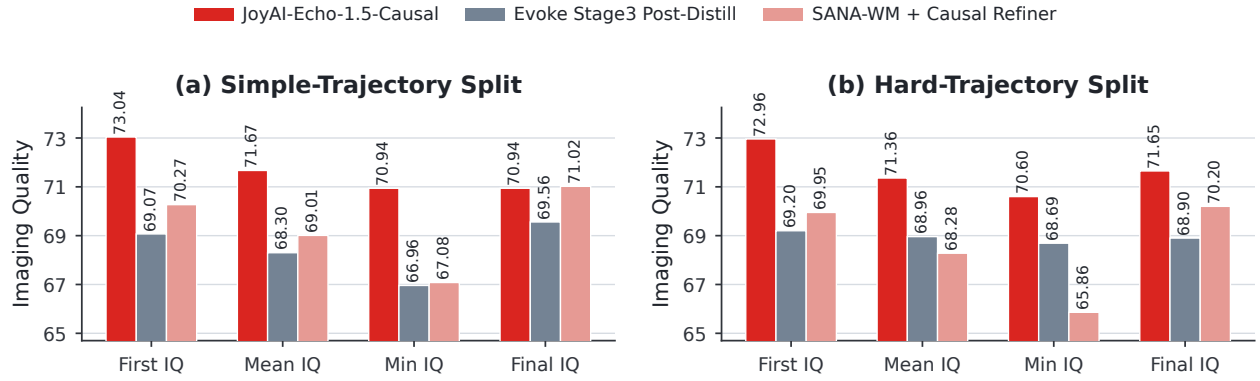


Figure 11 Imaging quality comparison on the Simple- and Hard-Trajectory splits. We report the first-window, mean, minimum, and final-window IQ. Our method achieves consistently higher mean and minimum IQ than Evoke and SANA-WM with Causal Refiner. Although the competing methods exhibit smaller IQ drops, their overall IQ remains lower, indicating that a small drop can also result from consistently low imaging quality rather than stronger long-horizon quality preservation.

quality and revisit consistency over the full 60-second trajectory. On the Simple split, its rotation error reaches 3.22° at 961 frames, while on the Hard split it increases to 12.05° , indicating that accumulated pose drift remains a key challenge under difficult long-horizon trajectories.

Tab. 7 further compares the distilled causal model with recent state-of-the-art causal distillation methods under the same 961-frame horizon. JoyAI-Echo-1.5-Causal achieves the highest VBench scores on both Simple and Hard splits, with 80.13 and 81.06, respectively, together with the best trajectory accuracy and strongest revisit consistency among the evaluated causal models. As further illustrated in Fig. 11, it achieves substantially higher first-window, mean, and minimum IQ than Evoke and SANA-WM with Causal Refiner, while maintaining competitive final-window IQ. Although these baselines exhibit smaller ΔIQ , their lower initial and overall IQ suggest that a small drop may also result from consistently low imaging quality rather than stronger preservation of a high-quality initial state. Compared with the undistilled 961-frame model, causal few-step distillation introduces a moderate degradation in trajectory accuracy and revisit consistency while largely preserving long-horizon visual quality, demonstrating robust 60-second autoregressive rollout capability.

Qualitative Results. Fig. 12 compares approximately 60-second causal rollouts generated from the same initial observation under matched camera trajectories. As the rollout horizon increases, competing methods

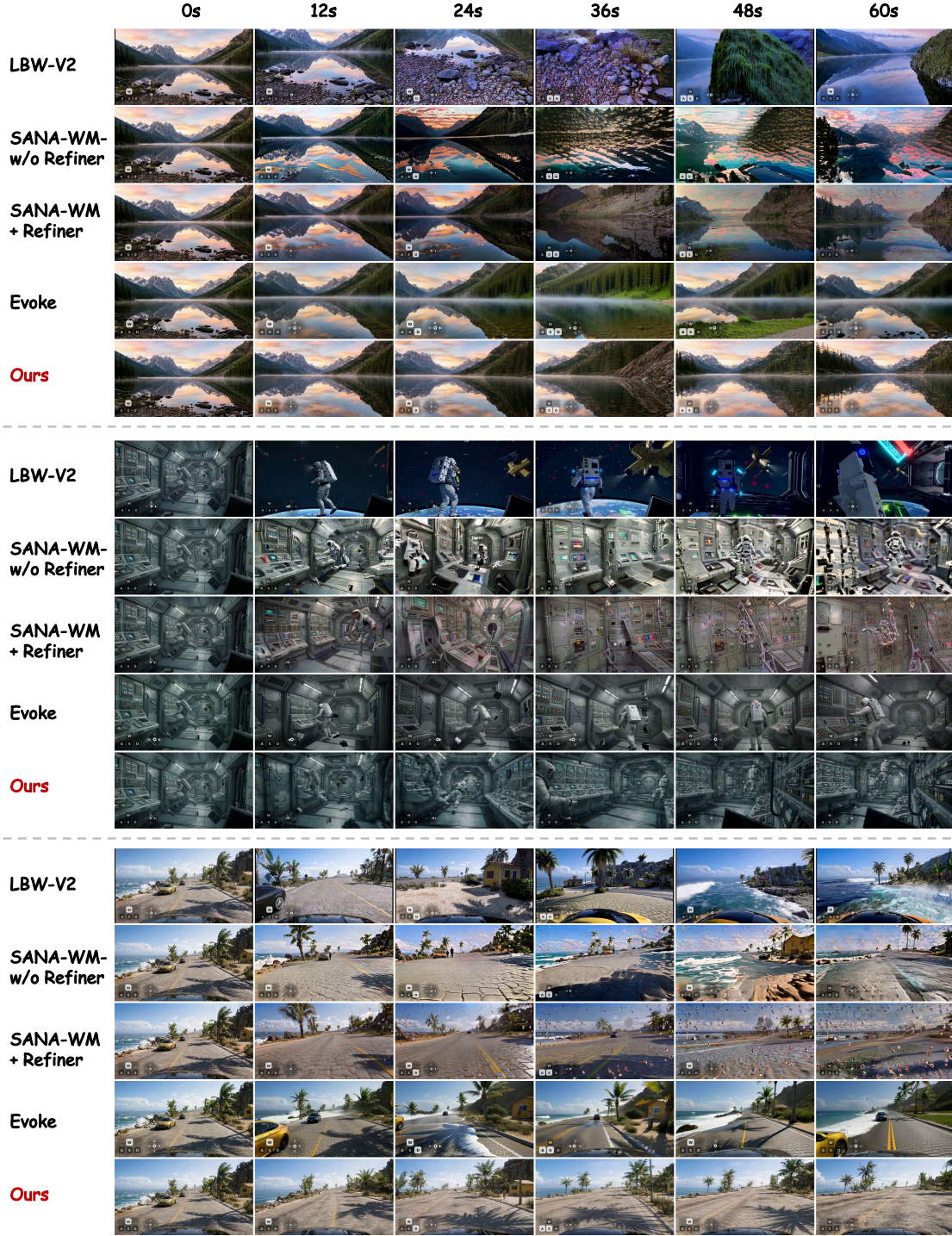


Figure 12 Qualitative comparison of 961-frame (60-second at 16 FPS) causal rollouts on representative SANA-WM-Bench trajectories. We compare JoyAI-Echo-1.5-Causal with LingBot-World-v2 (LBW-V2), SANA-WM with and without Causal Refiner, and Evoke. Frames are sampled every 12 seconds from the same initial observation and camera-trajectory condition. Across the extended rollout, JoyAI-Echo-1.5-Causal exhibits less accumulated autoregressive drift, more consistent trajectory following, and stronger preservation of scene identity, visual style, and spatial structure.

Table 8 Human evaluation based on Good–Similar–Bad (GSB) pairwise comparisons between JoyAI-Echo-1.5 (world model variant) and Happy Oyster for world modeling. Numbers denote the percentage of user preferences.

Aspect	JoyAI-Echo-1.5	Similar	Happy Oyster
Semantic following	13.63%	56.94%	29.44%
Initial-world consistency	39.50%	45.06%	15.44%
Motion	18.94%	56.51%	24.56%
Spatiotemporal consistency	57.69%	31.88%	10.44%
Visual aesthetics	52.88%	23.32%	23.81%
Overall	63.13%	9.82%	27.06%

Table 9 Human evaluation based on Good–Similar–Bad (GSB) pairwise comparisons between JoyAI-Echo-1.5 (world model variant) and LingBot-World-v2 for world modeling. Numbers denote the percentage of user preferences.

Aspect	JoyAI-Echo-1.5	Similar	LingBot-World-v2
Semantic following	34.00%	50.86%	15.14%
Initial-world consistency	14.64%	77.65%	7.71%
Motion	25.29%	54.71%	20.00%
Spatiotemporal consistency	32.64%	52.65%	14.71%
Visual aesthetics	32.50%	47.57%	19.93%
Overall	46.50%	31.92%	21.57%

exhibit different forms of cumulative autoregressive drift. LingBot-World-v2 progressively alters the scene composition and viewpoint, while SANA-WM without the Causal Refiner can undergo substantial appearance and style shifts. The Causal Refiner improves local temporal stability but does not fully prevent long-term structural and appearance drift. Evoke generally maintains coherent short-term appearance, but deviations in viewpoint evolution and scene configuration become increasingly apparent over longer horizons.

In contrast, JoyAI-Echo-1.5-Causal exhibits the strongest long-horizon consistency across these examples. In both natural and structured indoor environments, it better preserves visual style, scene and subject identity, and major geometric structures while evolving the viewpoint according to the prescribed trajectory. These properties remain comparatively stable at later timestamps instead of progressively degrading with the accumulated autoregressive history. The qualitative observations are consistent with our trajectory-following, temporal-degradation, and revisit-consistency metrics, supporting the improved robustness of our causal model to long-horizon error accumulation and its stronger action-conditioned world persistence.

8.2.3 User Study

We additionally conduct a subjective evaluation on a self-built World-Model Benchmark (WMB) containing 200 cases. The benchmark covers game, humanoid-robot, natural outdoor, household indoor, and driving scenes, with 132 third-person and 68 first-person cases. We compare JoyAI-Echo-1.5 with LingBot-World-v2 [30] and HappyOyster through randomized and anonymized side-by-side comparisons. Annotators evaluate semantic following, initial-world consistency, spatiotemporal consistency, motion quality, and visual aesthetics, together with an overall preference.

Against LingBot-World-v2, JoyAI-Echo-1.5 receives higher explicit preference across all five dimensions, with the largest margins in semantic following (34.0% vs. 15.1%) and spatiotemporal consistency (32.6% vs. 14.7%). The overall preference is 46.5% for JoyAI-Echo-1.5 versus 21.6% for LingBot-World-v2.

Against HappyOyster, JoyAI-Echo-1.5 shows substantial advantages in spatiotemporal consistency (57.7% vs. 10.4%), visual aesthetics (52.9% vs. 23.8%), and initial-world consistency (39.5% vs. 15.4%). HappyOyster is preferred more often in semantic following and motion quality, while JoyAI-Echo-1.5 achieves a substantially higher overall preference (63.1% vs. 27.1%). These results complement the public benchmarks, showing that

the gains in visual quality and spatiotemporal consistency are also reflected in human preference.

9 Conclusion

We presented a framework for extending audio-visual generation beyond isolated clips toward persistent stories and interactive worlds. The long-video system uses composable cross-shot audio-visual memory to preserve character appearance, speaker identity, and narrative continuity under flexible text, image, and memory conditioning. The world-model system translates heterogeneous navigation inputs into calibrated metric 6-DoF trajectories, enabling precise and controller-agnostic interaction across diverse environments and viewpoints. Together with few-step distillation, causal audio-visual teacher forcing, and short- and long-horizon Self-Gradient Forcing, these designs allow the models to generate synchronized audio and video efficiently while remaining stable under self-generated histories. Experiments demonstrate improved cross-shot consistency, instruction following, speech fidelity, trajectory control, and long-horizon world persistence, with leading results on both long-form generation evaluations and public world-model benchmarks. These findings suggest that memory, geometric control, and rollout-aware training are complementary foundations for video systems that can remember past events, respond to user intent, and continue evolving over time. Accumulated rotational drift on difficult long-horizon trajectories remains a key limitation, motivating future work on stronger geometric consistency, persistent world-state representations, and more reliable open-ended interaction.

10 Authors

10.1 Core Contributors

Nan Duan, Haoyang Huang, Weiyang Jin, Haoran Li, Yaowei Li, Yuming Li, Yijun Liu, Xin Lu, Xiaoxiao Ma, Yanwen Ma, Yaofeng Su, Yilang Sun, Haoyu Wang, Zeyue Xue, Songchun Zhang, Junhao Zhuang

10.2 Contributors

Bi Cheng, Jiayi Deng, Zuopeng Dong, Libing Fang, Tong He, Xintong Liu, Ruofan Lv, Jiaqi Shi, Zihan Su, Jiaqi Wang, Meihui Wang, Pan Wang, Haoyu Wu, Shiyi Zhang, Xuying Zhang, Mingchen Zhong

References

- [1] Kelvin C. K. Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Investigating tradeoffs in real-world video super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 5962–5971, 2022.
- [2] Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. Advances in Neural Information Processing Systems, 37:24081–24125, 2024.
- [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] 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.
- [5] Juechu Dong, Boyuan Feng, Driss Guessous, Yanbo Liang, and Horace He. Flexattention: A programming model for generating fused attention variants. In Proceedings of Machine Learning and Systems, volume 7, 2025.
- [6] DreamX Team, Yancheng Bai, Rui Chen, Xiangxiang Chu, Ruqing Dang, Hao Dou, Bingjie Gao, Qiwen Gu, Siyu Hong, Jiachen Lei, et al. DreamX-World 1.0: A general-purpose interactive world model. arXiv preprint arXiv:2606.16993, 2026.
- [7] Gunnar Farneback. Two-frame motion estimation based on polynomial expansion. In Scandinavian Conference on Image Analysis (SCIA), pages 363–370. Springer, 2003.
- [8] Zhengyang Geng, Mingyang Deng, Xingjian Bai, Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. In Advances in Neural Information Processing Systems, volume 38, pages 75460–75482, 2025.
- [9] Google DeepMind. Genie 3: A new frontier for world models, 2025. URL <https://deepmind.google/discover/blog/genie-3-a-new-frontier-for-world-models/>.
- [10] Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. LTX-2: Efficient joint audio-visual foundation model. arXiv preprint arXiv:2601.03233, 2026.
- [11] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685, 2021.
- [12] Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, et al. ViPE: Video pose engine for 3D geometric perception. arXiv preprint arXiv:2508.10934, 2025.
- [13] Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. Advances in Neural Information Processing Systems, 38:167283–167308, 2026.
- [14] 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.
- [15] HyCreator Team. HyCreator: An agent harness for long video generation, 2026. URL <https://hycreator.tencent.com/>. Multimodal Reinforcement Learning Team, Tencent HY.
- [16] Fan Jiang, Zhaoxu Sun, Mengchao Wang, Ziyu Zhu, Chiyu Wang, Yunpeng Zhang, Wenlin Liu, Yun Wang, Xue Zheng, Rui Sun, et al. ABot-World-0: Infinite interactive world rollout on a single desktop GPU. arXiv preprint arXiv:2607.19191, 2026.
- [17] Haoran Li, Fredreic Li, Shichen Ma, Jie Huang, Yijun Liu, Jiaqi Shi, and Yanwen Ma. Joyai-echo: Pushing the frontier of long audio-visual generation. 2026.
- [18] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In The Eleventh International Conference on Learning Representations, 2023.

- [19] 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*, 2026.
- [20] 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. In *European Conference on Computer Vision*, 2023.
- [21] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In *Advances in Neural Information Processing Systems*, volume 35, pages 5775–5787, 2022.
- [22] Xin Lu, Zihao Fan, Mingchen Zhong, Jie Huang, Xueyang Fu, and Zheng-Jun Zha. OmniVR: Joint video-audio conditional generation for restoring degraded historical films. *arXiv preprint arXiv:2608.04224*, 2026.
- [23] 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.
- [24] 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.
- [25] Luxury, Jie Huang, Zihao Fan, Xiaoxiao Ma, Yuming Li, Jun-hao Zhuang, Zeyue Xue, Siming Fu, Haoran Li, Mingchen Zhong, Guohui Zhang, Shichen Ma, Yijun Liu, Jiaqi Shi, Yanwen Ma, Yaofeng Su, Haoyu Wang, Yaowei Li, Songchun Zhang, Weiyang Jin, Yuxuan Bian, Shiyi Zhang, Haojun Xu, Shuai Lu, Xin Han, Wei Tang, Haoyang Huang, and Nan Duan. Ultra Flash: Scaling real-time streaming video generation to high resolutions. *arXiv preprint arXiv:2606.09150*, 2026.
- [26] NVIDIA. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025.
- [27] NVIDIA. Cosmos 3: Omnimodal world models for physical AI. Technical report, NVIDIA, 2026.
- [28] 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.
- [29] 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>.
- [30] Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, et al. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026.
- [31] 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.
- [32] Team Seedance, Heyi Chen, Siyan Chen, Xin Chen, Yanfei Chen, Ying Chen, Zhuo Chen, Feng Cheng, Tianheng Cheng, Xinqi Cheng, et al. Seedance 1.5 pro: A native audio-visual joint generation foundation model. *arXiv preprint arXiv:2512.13507*, 2025.
- [33] Tianchang Shen, Sherwin Bahmani, Kai He, Sangeetha Grama Srinivasan, Tianshi Cao, Jiawei Ren, Ruilong Li, Zian Wang, Nicholas Sharp, Zan Gojcic, et al. Lyra 2.0: Explorable generative 3D worlds. *arXiv preprint arXiv:2604.13036*, 2026.
- [34] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 32211–32252. PMLR, 2023.
- [35] 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.
- [36] Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. WorldPlay: Towards long-term geometric consistency for real-time interactive world modeling. *arXiv preprint arXiv:2512.14614*, 2025.

- [37] Zeyue Tian, Lei Ke, Zhaoyang Liu, Ruibin Yuan, Liუმeng Xue, Yujiu Yang, Weijia Chen, Xu Tan, Qifeng Chen, Wei Xue, and Yike Guo. Audiox-turbo: A unified framework for efficient anything-to-audio generation. [arXiv preprint arXiv:2606.12555](#), 2026.
- [38] Wan Team, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. [arXiv preprint arXiv:2503.20314](#), 2025.
- [39] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 1905–1914, 2021.
- [40] Yi Wang, Yanan 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*.
- [41] Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixun Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, et al. Matrix-Game 3.0: Real-time and streaming interactive world model with long-horizon memory. [arXiv preprint arXiv:2604.08995](#), 2026.
- [42] Ronald J Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280, 1989.
- [43] Bing Wu, Chang Zou, Changlin Li, DuoJun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, et al. HunyuanVideo 1.5 technical report. [arXiv preprint arXiv:2511.18870](#), 2025.
- [44] Ruiqi Wu, Xuanhua He, Meng Cheng, Tianyu Yang, Yong Zhang, Zhuoliang Kang, Xunliang Cai, Xiaoming Wei, Chunle Guo, Chongyi Li, and Ming-Ming Cheng. Infinite-World: Scaling interactive world models to 1000-frame horizons via pose-free hierarchical memory. [arXiv preprint arXiv:2602.02393](#), 2026.
- [45] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. In *NeurIPS*, 2024.
- [46] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *CVPR*, 2024.
- [47] Yuanyang Yin, Gongxuan Wang, Yifan Zhan, Chuanhao Li, Kaipeng Zhang, and Feng Zhao. Alaya-evoke: From linear-scaling supervision to endless world. [arXiv preprint arXiv:2608.13546](#), 2026.
- [48] Kaining Ying, Hengrui Hu, Siyu Ren, Jiamu Li, Fengjiao Chen, Ziwen Wang, Xuezhi Cao, Xunliang Cai, and Henghui Ding. WBench: A comprehensive multi-turn benchmark for interactive video world model evaluation. [arXiv preprint arXiv:2605.25874](#), 2026.
- [49] Cheng Zhang, Boying Li, Meng Wei, Yan-Pei Cao, Camilo Cruz Gambardella, Dinh Phung, and Jianfei Cai. Unified camera positional encoding for controlled video generation. [arXiv preprint arXiv:2512.07237](#), 2025.
- [50] 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.
- [51] Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. UniPC: A unified predictor-corrector framework for fast sampling of diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pages 49842–49869, 2023.
- [52] Kaiyang Zhou and Tao Xiang. Torchreid: A library for deep learning person re-identification in pytorch. [arXiv preprint arXiv:1910.10093](#), 2019.
- [53] Haoyi Zhu, Haozhe Liu, Yuyang Zhao, Tian Ye, Junsong Chen, Jincheng Yu, Tong He, Song Han, and Enze Xie. SANA-WM: Efficient minute-scale world modeling with hybrid linear diffusion transformer. [arXiv preprint arXiv:2605.15178](#), 2026.
- [54] Junhao Zhuang, Shiyi Zhang, Yuxuan Bian, Yaowei Li, Yawen Luo, Yijun Liu, Weiyang Jin, Songchun Zhang, Xianglong He, Xuying Zhang, Haoran Li, Haoyang Huang, Zeyue Xue, and Nan Duan. Self Gradient Forcing: Native long video extrapolation. [arXiv preprint arXiv:2607.20368](#), 2026.