



EchoWM: Open and Enterable Omnimodal World Models

Songchun Zhang^{1,3,*}, Yaowei Li^{2,3,*}, Junhao Zhuang^{3,*}, Weiyang Jin^{4,3,*},
 Haoyu Wang^{3,5}, Xin Lu^{3,6}, Shiyi Zhang⁵, Haoran Li³, Xiaoxiao Ma^{3,6}, Yuming Li^{3,4},
 Yijun Liu^{3,5}, Yaofeng Su^{3,7}, Yanwen Ma^{3,8}, Haoyu Wu^{3,4}, Zihan Su^{3,5}, Yue Ma¹,
 Lvmin Zhang⁹, Haoyang Huang³, Zeyue Xue^{3,‡}, Anyi Rao^{1,†}, Nan Duan^{3,†}

¹HKUST, ²PKU, ³Joy Future Academy, JD, ⁴HKU, ⁵THU, ⁶USTC, ⁷FDU, ⁸Beihang University,
⁹Stanford University

*These authors contributed equally., [†]Corresponding authors., [‡]Project lead.

Abstract

We present ECHOWM, an omnimodal world model for enterable generative media that responds to continuous navigation while jointly generating 720p video, environmental sound, music and speech. We organize interaction around camera intent: in first-person scenes, it specifies observer motion, while in third-person scenes, camera-character dynamics are learned from data without view-specific controllers. Discrete commands and continuous poses are mapped to a shared metric-scale relative 6-DoF trajectory, with dataset-level calibration preserving motion magnitude across heterogeneous data. To jointly learn audio-visual generation and trajectory control, we construct a complementary data engine and adopt progressive training followed by autoregressive post-training for long-horizon generation. Extensive evaluations show that ECHOWM achieves strong trajectory following and high visual quality on public world-model benchmarks, supporting both first- and third-person interaction across varied subjects, and maintaining synchronized environmental sound and speech over long-horizon generation.

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

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





Figure 1 EchoWM turns audio-visual generation into an enterable world. Given a reference observation, a structured world description, and a user control sequence represented as a relative 6-DoF trajectory, one model supports first- and third-person interaction, synchronized environmental sound and speech, and multi-turn continuation.

1 Introduction

Generative video models are rapidly evolving into expressive audio-visual media engines [7, 9, 24], yet user interaction remains predominantly prompt-based and passive. World models invert this relationship by generating future states that respond to continuous user or agent inputs [1, 12, 14, 25, 35, 44, 54]. However, most existing systems produce silent visual rollouts or rely on rigid, subject-specific action spaces. In an interactive world, audio is essential: footsteps and collisions reveal physical feedback, ambient sounds convey spatial dynamics, and speech connects visible subjects to social and narrative context. While recent work explores joint audio-visual generation in robotics [23], enabling continuous, multimodal participation across general and game scenes remains an open challenge.

To address this gap, we introduce the concept of an enterable omnimodal world: a generated environment that users can continuously traverse with synchronized visual, environmental sound, and speech output. Here, interaction focuses on navigation and viewpoint evolution representable by continuous trajectories, rather than unrestricted control over arbitrary subject behaviors. Our key insight is to organize this interaction around camera intent: a subject-agnostic interaction interface grounded in relative 6-DoF trajectories that captures how an observer intends to navigate and view the generated scene. Operating independently of specific actor dynamics, vehicle controls, or camera-rig constraints, camera intent generalizes seamlessly across diverse subjects and perspectives: in first-person scenes, it directly guides observer motion; in third-person scenes, the same trajectory naturally orchestrates coordinated character locomotion, vehicle displacement, camera tracking, and scene framing. By learning this observer-subject relationship directly from data, a single interface spans diverse perspectives and domains without requiring viewpoint-specific controllers.

Table 1 Reported native capabilities of representative generative video and world models. A checkmark denotes explicit native support, whereas “—” means that the capability is not reported. Pose trajectory and discrete navigation denote continuous and discrete viewpoint-related interaction, respectively.

Model	Output	Participation			Audio feedback		
	High-res. video	FPP + TPP	Continue traj.	Discrete keyboard	Env. sound	Background music	Spoken speech
<i>Interactive world models</i>							
LingBot-World [27]	✓	✓	✓	—	—	—	—
DreamX-World 1.0 [4]	✓	✓	✓	—	—	—	—
SANA-WM [54]	✓	✓	✓	—	—	—	—
Matrix-Game 3.0 [37]	✓	✓	—	✓	—	—	—
Matrix-Game 3.5 [26]	✓	✓	✓	—	—	—	—
Genie 3 [6]	✓	✓	—	✓	—	—	—
Happy Oyster [48]	✓	✓	—	✓	✓	✓	—
<i>Generative and omnimodal models</i>							
HunyuanVideo 1.5 [39]	✓	—	—	—	—	—	—
LTX-2.3 [9]	✓	—	—	—	✓	✓	✓
Cosmos 3 [23]	✓	—	—	—	✓	—	—
EchoWM	✓	✓	✓	✓	✓	✓	✓

Realizing an enterable omnimodal world under this unified interface requires resolving three coupled geometric and distributional mismatches, which we address through three dedicated technical designs:

- **Architecture (Unifying Viewpoint Semantics & Scale):** Heterogeneous video sources differ fundamentally in metric scale and control semantics; naive normalization can erase displacement magnitude or degrade control consistency. We design a unified architecture that maps both discrete commands and continuous poses to a shared relative 6-DoF trajectory with a fixed dataset-level metric scale, while encoding pairwise camera geometry via relative UCPE. Combined with viewpoint context, the learned world prior interprets this intent as either first-person observer motion or third-person camera–subject evolution.
- **Data (Bridging AV Richness & Control Precision):** Real-world videos offer rich acoustics but messy motion, whereas simulations provide precise geometry with simple semantics. We build a world-data engine from four complementary sources (internal gameplay, web gameplay, UE simulation, and general web video) and structure them into AV-rich, control-clean, and balanced high-quality mixtures, allowing the model to jointly learn multimodal quality and control responsiveness without relying on a single ideal data source.
- **Training (Decoupling Multimodal Quality & Control Responsiveness):** Unifying audio-visual synthesis with precise control induces severe training distribution conflicts. We therefore use a progressive four-stage curriculum in which Audio-Visual Continued Pretraining (AV-CPT) establishes visual dynamics and acoustic priors, Action Fine-Tuning (Action-SFT) isolates control sensitivity by freezing the backbone and training only a lightweight trajectory pathway, Joint Fine-Tuning (Joint-FT) consolidates both capabilities with a low learning rate on balanced data, and Autoregressive Post-Training adapts the model to self-generated context for long-horizon inference.

We instantiate these principles in ECHOWM, an omnimodal world model for enterable generative media. Given a reference observation and a textual world description, ECHOWM responds to continuous navigation by jointly generating 720p video, environmental audio, and speech with multi-turn continuation. As shown in Table 1 and Figure 1, ECHOWM covers real indoor/outdoor scenes, stylized game worlds, artistic content, and diverse subjects. Quantitative evaluations demonstrate state-of-the-art performance across public benchmarks: ECHOWM ranks first on WBench Navigation and achieves top-tier visual quality on SANA-WM-Bench across both simple and challenging trajectory splits, while maintaining robust visual and control consistency over long-horizon rollouts.

Our main contributions are summarized as follows:

- **Enterable Omnimodal World Modeling.** We formulate an enterable generative-media setting that combines continuous user navigation with native joint generation of video, environmental sound, and speech.

The formulation supports both first- and third-person interaction and extends beyond a single rollout through synchronized multi-turn continuation.

- **Unified Camera-Intent Interface.** We organize navigation around camera intent and represent heterogeneous discrete commands and continuous camera poses with a shared relative 6-DoF trajectory. Dataset-level metric calibration and relative UCPE provide consistent geometric conditioning across data sources, while the learned observer–subject relationship allows the same interface to support first-person observer motion and third-person camera–character evolution without view-specific controllers or camera rigs.
- **Capability-Aligned World Data Engine.** We build a scalable data engine from internally collected gameplay, human-played Internet gameplay, Unreal Engine simulation, and general Internet video, combining complementary audio-visual and geometric signals. The processed data are organized into AV-rich, control-clean, and balanced high-quality mixtures so that audio-visual generation and trajectory following are trained from the subsets in which their supervision is most reliable.
- **Progressive Training.** We introduce a progressive training strategy that first adapts joint audio-visual generation, then learns trajectory conditioning with the backbone frozen, and subsequently fine-tunes both components together. Autoregressive post-training further exposes the model to self-generated audio-visual histories, enabling stable long-horizon and multi-turn interactive generation.

2 Related Work

2.1 Interactive Generative World Models

Generative world models have increasingly moved from compact latent dynamics for imagination and control toward domain-specific action-conditioned simulation and open-ended interactive video generation [8, 10, 11]. Early systems such as UniSim and Genie learn visual futures conditioned on agent actions [1, 44], while recent game-oriented models demonstrate real-time or playable neural simulation [18, 21, 28]. More recent systems substantially broaden visual fidelity, domain coverage, and interaction horizon [14, 22, 35, 53]. Genie 3 supports real-time exploration of diverse 720p environments [6]; WorldPlay and Matrix-Game focus on streaming generation and long-term geometric consistency [13, 31, 37]; LingBot-World, SANA-WM, and DreamX-World further extend controllable world generation toward broader visual domains and longer trajectories [4, 27, 54]. Most of these models nevertheless remain primarily visual world simulators, with interaction interfaces often tailored to specific domains, environments, or control schemes. Our work instead focuses on the complementary problem of building a single camera-centric interaction interface that spans first- and third-person navigation while natively generating video, environmental sound, and speech across general and game scenes.

2.2 Action and Camera Conditioning for Interactive Generation

Interactive generative models commonly expose control through either discrete actions or continuous camera geometry. Action-conditioned world models such as ABot-World and Hunyuan-GameCraft condition future generation on keyboard or controller inputs, supporting navigation and character control over time [17, 18]. Such interfaces are intuitive, but their action semantics are typically tied to a particular controller or environment. Camera-controllable video models instead condition generation on poses or trajectories. A common approach represents camera rays with Plücker coordinates: CameraCtrl [12], for example, processes dense Plücker features with a dedicated camera encoder, while the LingBot-World family injects Plücker-based camera conditions through feature modulation such as AdaLN [27]. MotionCtrl, CamCo, and GEN3C extend camera control to flexible motion and 3D-consistent generation [25, 36, 42]. More recent methods such as UCPE [50] incorporate relative camera geometry directly into attention, reducing reliance on a standalone camera encoder. These lines of work differ in both user-facing control spaces and internal geometric representations. EchoWM maps discrete navigation commands and continuous metric poses to a shared relative 6-DoF trajectory, using the same geometric interface across first- and third-person interaction.

2.3 Joint Audio-Visual and Omnimodal Generation

Audio generation for video has traditionally been studied as a video-to-audio problem, where sound is synthesized after the visual stream [3, 20, 52]. More recent generative models instead produce visual and acoustic content jointly, including Movie Gen, Veo 3, SyncFlow, and LTX-based audio-visual diffusion models [7, 9, 19, 24]. OmniForcing further converts joint audio-visual diffusion into a streaming autoregressive generator while preserving cross-modal synchronization [30]. In parallel, omnimodal world models such as Cosmos 3 incorporate video, audio, language, and action for embodied-agent and Physical-AI settings [23]. These advances establish strong native audio-visual generation, but most do not expose continuous geometric control over the generated world. EchoWM focuses on combining native environmental sound and speech with a shared trajectory interface for first- and third-person interaction across general and game-oriented generative media.

2.4 Autoregressive and Long-Horizon Video Generation

Autoregressive video diffusion replaces bidirectional attention with causal computation. Diffusion Forcing adapts full-sequence generators to causal prediction [2], and Self Forcing trains on self-generated rollouts [16]. Long context is then treated as memory: FramePack bounds the token budget, and Longlive keeps an attention sink of early frames [45, 51]. Interactive worlds go further with explicit retrieval. WorldMem and Context-as-Memory index past views by pose or field of view [41, 49]; Infinite-World, WorldPlay, and Matrix-Game 3.0 carry hierarchical or camera-aligned memories across long rollouts [31, 37, 40]. OmniForcing and Self Gradient Forcing extend this lineage to synchronized audio-video streams and native long-video extrapolation [28, 30, 55]. We apply a similar causal post-training strategy to trajectory-conditioned joint audio-visual world generation. Unlike explicit retrieval, which incurs additional computation for searching and matching against long-term history, the generator maintains a bounded sink-plus-FIFO history and writes a compact state from past chunks only, allowing long-horizon interaction to reuse native memory even after recent tokens leave the active window.

3 World Data Engine

Training an enterable omnimodal world requires four properties that rarely coexist in a single data source: diverse appearance, natural audio, interactive motion, and reliable camera geometry. We therefore combine four complementary sources: internally collected game recordings, human-played Internet game recordings, Unreal Engine (UE) simulation, and general Internet video. Together, these sources provide complementary supervision for audio-visual world modeling and trajectory-conditioned interaction. Internally collected gameplay provides controlled interaction, native audio, and action logs; human-played gameplay contributes natural control timing and richer speech; UE provides ground-truth metric geometry and controlled motion coverage; and general Internet video broadens appearance and acoustic diversity beyond game-rendered worlds. We process them through two complementary paths: an audio-visual path that prioritizes appearance, speech, and environmental sound, and a geometry path that preserves long-range temporal continuity for metric camera recovery. The two paths are joined after processing into temporally aligned examples with structured audio-visual annotations and reliable metric camera trajectories wherever geometry can be recovered. The resulting examples are organized into three stage-aligned mixtures for AV-CPT, Action-SFT, and Joint-FT. Table 2 summarizes the complementary signals and limitations of the four sources, while figure 4 illustrates the overall source-to-mixture pipeline.

3.1 Complementary Data Sources

Internally collected gameplay. Our internally collected game recordings form a major source of trajectory-conditioned data and span first- and third-person play across open-world exploration, urban driving, and equestrian traversal. Programmatic scripts control the characters, providing explicit control inputs together with cleaner action-observation correspondence than uncontrolled Internet video. A ReShade-based plug-in suppresses heads-up displays and interface overlays before capture, preserving the original field of view without post-hoc cropping. We synchronously record high-resolution RGB video, the native stereo game mix, script-issued action logs, and environment metadata. The native mix retains dialogue, music, ambience, and

Table 2 Complementary strengths of the four data sources. Filled versus hollow dots encode relative source strength; checkmarks indicate availability; dashes indicate absence. Scene dynamics refers to subject and environment motion beyond the commanded camera or character trajectory.

Source	Audio-visual content				Action signal			Coverage	
	Visual quality	Audio richness	Speech	Scene dynamics	Action clarity	Logs	Pose reliability	Human operation	Non-game domain
Collected gameplay	●●●	●●●	✓	●●●	●●●	✓	●●○	—	—
Internet gameplay	●●○	●●●	✓	●●●	●●○	—	●●○	✓	—
UE renders	●●○	—	—	—	●●●	✓	●●●	—	—
General Internet video	●●●	●●●	✓	●●●	●○○	—	●○○	—	✓

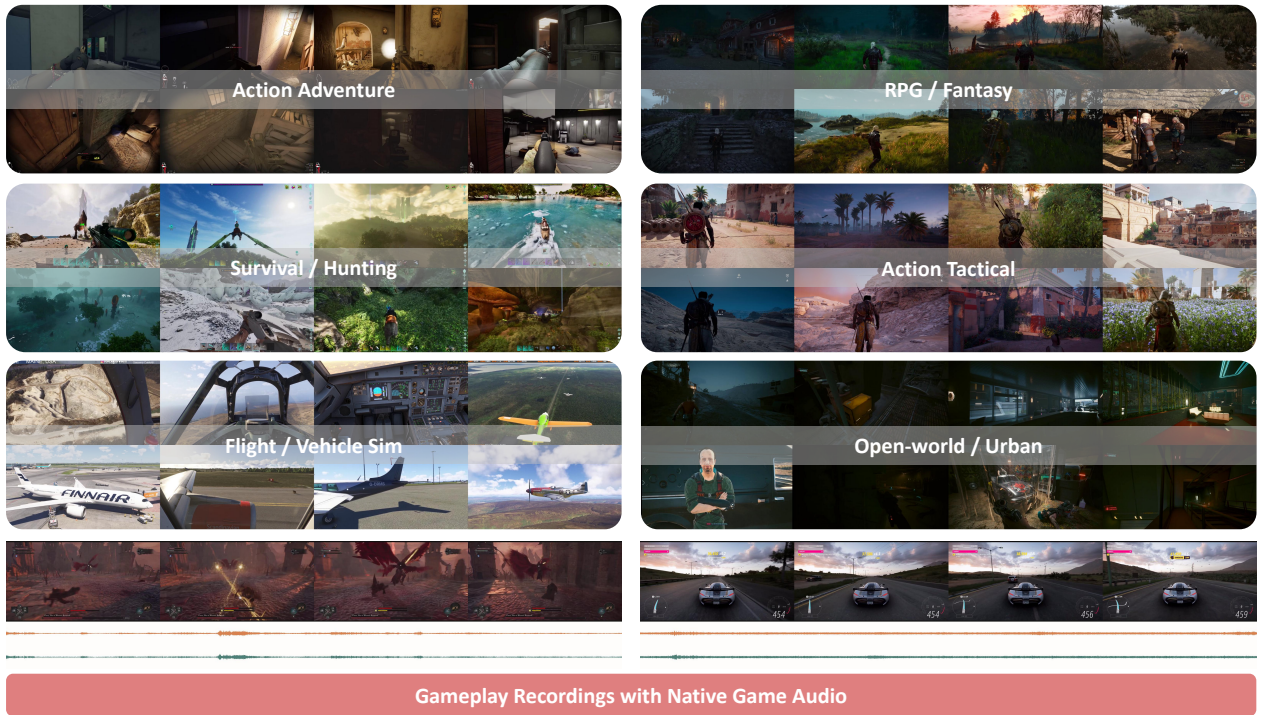


Figure 2 Gameplay recordings with native game audio. Representative frames from the internally collected gameplay corpus span six activity categories, with synchronized native-game audio waveforms shown below the visual sequences.

interaction sounds such as footsteps, vehicles, and collisions. Representative visual sequences and synchronized native-game audio waveforms are shown in [figure 2](#).

Importantly, native action logs and realized camera trajectories capture complementary aspects of interaction. An issued control does not necessarily induce the corresponding camera displacement. For example, a forward command may produce little or no translation when the character is blocked by geometry, constrained by terrain, or affected by game physics. The action stream therefore records control intent, whereas the camera trajectory records the motion actually realized by the environment. We preserve both signals in the dataset, while using the recovered camera trajectory as the model’s unified motion condition. Because camera poses are not exported from these games, we recover metric trajectories from video as described in [section 3.3](#).

Human-played Internet gameplay. Human-played Internet game recordings complement internally

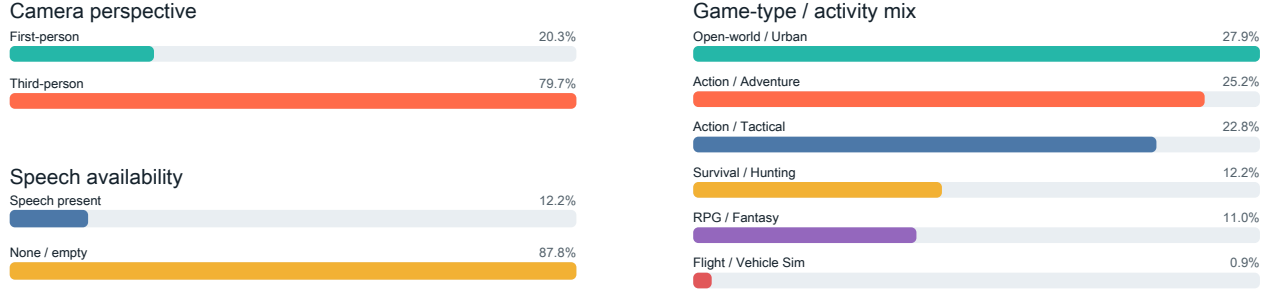


Figure 3 Game-video dataset statistics. The combined training corpora contain predominantly third-person clips, a broad mix of game activities, and a smaller speech-present subset. Percentages are normalized within the reported collection.

collected gameplay with less scripted and more context-dependent interaction. Although source-native action logs are unavailable, these recordings exhibit natural control timing and contain reactions to scene events, in-game dialogue, player commentary, and more diverse human speech. Continuous clips with reliable recovered camera trajectories contribute to trajectory-conditioned training, while the broader pool strengthens audio-visual and speech modeling.

Unreal Engine simulation. UE complements game recordings with ground-truth metric poses, synchronized action logs, and controlled translation and rotation coverage, but does not provide natural audio. We drive characters in the Play-In-Editor world using local-frame navigation commands and engine physics rather than navigation-mesh paths, allowing realized motion to reflect collisions and environmental constraints. When a character encounters an obstacle, the runtime controller adjusts its heading and continues the rollout.

Each character carries one head-mounted first-person camera and four character-relative third-person cameras that remain oriented toward and move with the character. At every step, we export synchronized frames, keyboard states, ground-truth camera poses and intrinsics, and scene metadata for all five views. Training randomly selects one view together with its corresponding pose and viewpoint metadata from each rollout. Both signals are retained, but the current model is conditioned only on the realized camera trajectory. The capture rig itself is never exposed to the model.

General Internet video. General Internet video broadens the training distribution beyond game-rendered worlds with diverse real-world and cinematic visual and acoustic content. It contributes realistic appearance, environmental sound, speech, richer composition, and camera-subject relationships such as tracking, orbiting, and reframing. Because these clips provide neither source-native action logs nor standardized navigation, they primarily expand the audio-visual prior and improve cross-domain generalization. A smaller subset of temporally continuous clips that passes metric pose recovery and motion-quality filtering is additionally included in trajectory-conditioned training. The resulting game-video distribution is summarized in [figure 3](#).

3.2 Temporal Segmentation and Canonicalization

We process the data through separate audio-visual and geometry paths because the two objectives impose different requirements on temporal continuity.

Audio-visual path. For audio-visual training, shot-boundary detection first isolates continuous spans in Internet and cinematic footage. AV-eligible sources are then sliced directly into short model-ready clips, since learning visual appearance, speech, and environmental sound does not require long-range geometric reconstruction.

Geometry path. For non-UE data used in trajectory-conditioned training, we instead preserve continuous windows of approximately one minute before producing the final short clips. Reconstructing the long window first ensures that neighboring training clips inherit a shared geometric reconstruction rather than being estimated in separate coordinate systems. The longer window also provides broader temporal support and

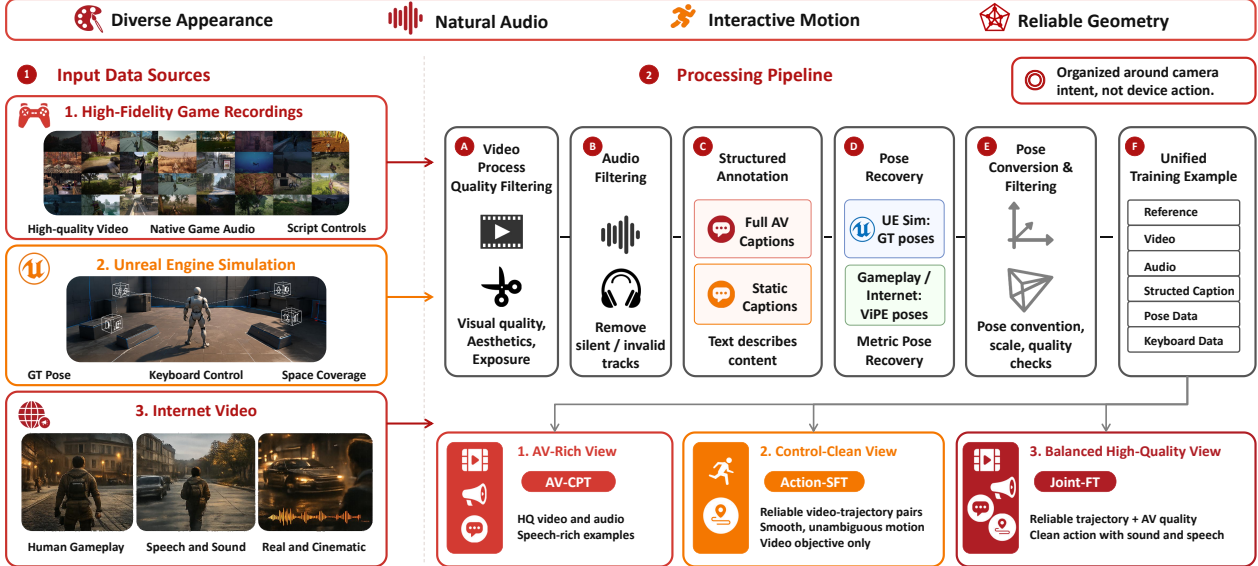


Figure 4 World data engine. We combine internally collected game recordings, human-played Internet game recordings, Unreal Engine renders, and general Internet video to complement action logs, natural audio, interactive motion, reliable poses, and cross-domain appearance. A shared pipeline constructs synchronized clips, filters and annotates video and audio, recovers reliable metric poses, and forms AV-rich, control-clean, and balanced high-quality mixtures for AV-CPT, Action-SFT, and Joint-FT, respectively.

more persistent correspondences for geometric optimization. Metric pose recovery and trajectory-level quality control are therefore performed on these long windows before they are sliced into model-ready clips.

Canonicalization. All retained media are re-encoded into canonical video and audio formats for stable decoding and batched training, while preserving stereo channels when present.

3.3 Metric Pose Recovery and Quality Filtering

Trajectory-conditioned training requires camera trajectories that are both temporally consistent and metrically scaled across heterogeneous data sources. UE provides camera-to-world poses directly in metric units, whereas the remaining trajectory-eligible sources require pose recovery from video.

3.3.1 Long-Sequence Metric Pose Recovery

Following the robust annotation principle of SANA-WM [54], we separate pose estimation from subsequent trajectory quality control rather than directly exposing raw estimator outputs to training.

We use ViPE [15] as the optimization framework and combine VGGT-Omega [34] with MoGe-2 as complementary geometric backends. VGGT-Omega provides temporally coherent long-sequence geometry, while MoGe-2 supplies per-frame metric-depth constraints for metric-scale geometric optimization. ViPE jointly integrates these cues to optimize temporally consistent metric-scale camera poses and camera intrinsics. We additionally adapt ViPE to support per-frame intrinsic optimization to accommodate zoom, stabilization, cropping, and imperfectly known intrinsics in Internet videos. Together, these components produce metric camera trajectories and aligned intrinsics for long videos in the wild.

We perform pose recovery on approximately one-minute continuous windows rather than independently on the final short training clips. This shared reconstruction keeps neighboring clips in a common geometric coordinate system while providing broader temporal support for geometric optimization.

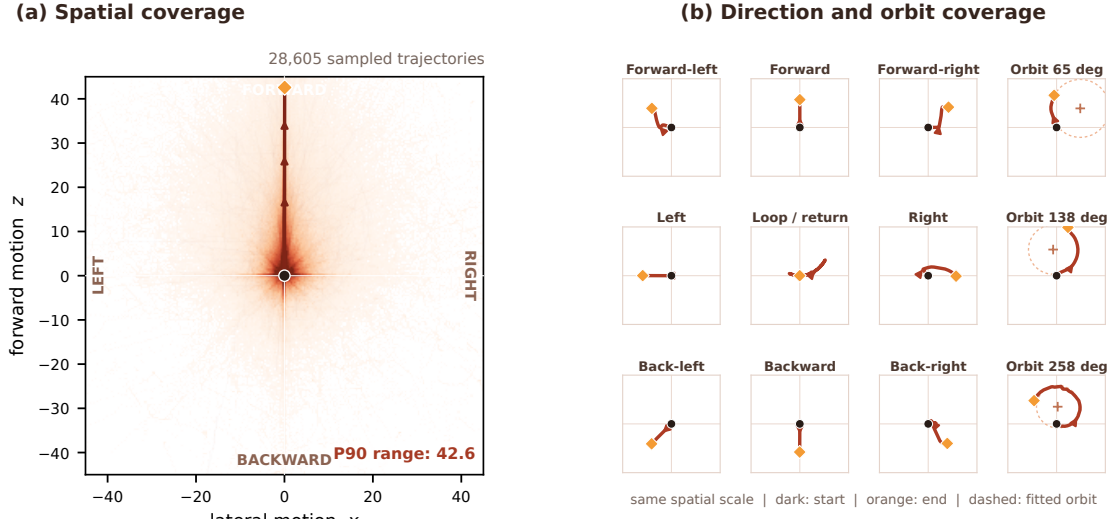


Figure 5 Trajectory coverage of the control-clean mixture. (a) Aggregate lateral–forward motion coverage over 28,605 sampled trajectories; the marked P90 range summarizes the robust spatial extent. (b) Representative straight, diagonal, return-loop, and orbit trajectories shown at a common spatial scale. Black and orange markers denote trajectory start and end points, respectively, and dashed circles show fitted orbits.

3.3.2 Temporal Densification and Alignment

Because processing every frame of a long sequence is computationally expensive, we temporally subsample each window before pose estimation. The recovered sparse trajectory is then aligned back to the original video timestamps using spherical linear interpolation (SLERP) for camera rotations and linear interpolation for translations. This interpolation is used only to restore frame-aligned trajectory annotations at the original timestamps; the underlying video observations are not synthesized or modified. The resulting dense trajectories and intrinsics are converted to a common camera-to-world convention, axis orientation, metric length unit, and temporal sampling rate. Each long sequence is then sliced into short trajectory-conditioned examples using identical temporal boundaries for video, available stereo audio, pose, intrinsics, and native control streams. All timestamps are re-indexed relative to the beginning of each short clip. UE rollouts bypass pose reconstruction and interpolation because their camera poses are available directly from the renderer, but follow the same convention conversion, temporal alignment, and final clip-construction pipeline.

3.3.3 Trajectory Quality Filtering

We apply trajectory-level quality control before reconstructed sequences are admitted to trajectory-conditioned training. Rather than relying on a single estimator score, we evaluate three complementary aspects of trajectory quality: (1) reconstruction reliability and intrinsic-calibration stability; (2) temporal consistency, including excessive frame-to-frame translation or rotation jitter and abrupt orientation changes; and (3) motion plausibility, including implausible displacement, degenerate trajectories, and extreme scale outliers. Sequences that fail these checks are removed before final short-clip construction.

Ground-truth UE trajectories undergo the same coordinate-convention conversion and motion-range checks, while their original metric values are preserved exactly. The retained poses are subsequently converted into the relative trajectory representation and dataset-level translation normalization described in [sections 4.2.1](#) and [4.2.2](#).

3.4 Audio-Visual Filtering and Structured Annotation

Audio-visual filtering. We remove visually degenerate clips using perceptual quality, aesthetic quality, exposure, and brightness criteria. For sources with audio, we additionally reject invalid tracks and clips with

unusable loudness statistics, while separately identifying speech-containing examples for stage-wise sampling. Sources without an audio stream, such as UE renders, are treated as missing-audio examples rather than semantic silence and are not used as positive supervision for audio content.

Structured audio-visual annotation. We use Qwen3-Omni [43] and Gemini-3-Pro in the annotation pipeline to process synchronized video and audio and produce structured descriptions of both persistent scene properties and time-varying audio-visual events. Their outputs are consolidated into the common structured schema below. The source records may use field-specific names such as `static_scene_caption`, `character`, `perspective`, `narrative_caption`, and `sounds`; we normalize these to the canonical fields **scene**, **subject**, **viewpoint**, **narrative**, and **sound**, respectively. Generation metadata and the raw tagged response are retained for provenance but are not exposed as semantic conditioning fields.

- **Persistent context.** The **scene** field describes the environment, spatial layout, persistent geometry, and ambient state while excluding the controlled subject and individual actions. The **style** field captures rendering medium, realism, lighting, art direction, and color treatment. The **viewpoint** field specifies the initial first- or third-person observer–subject relationship without describing subsequent camera motion. The **subject** field describes the primary subject’s identity, appearance, clothing, and equipment without encoding subsequent motion or interaction.
- **Temporal audio-visual events.** The **narrative** field records chronological scene evolution, including subject behavior, interactions, and camera motion. The **speech** field records spoken content and, when identifiable, the speaker role, distinguishing visible speakers, off-screen sources, and player commentary. The **sound** field describes non-speech acoustics, including environmental ambience, music, and interaction sound effects.

Preventing textual motion leakage. For trajectory-conditioned training, we exclude the narrative field because it may explicitly describe subject motion and camera evolution. Otherwise, the model could infer the target motion from text rather than from the supplied trajectory. During Action-SFT, we therefore retain the scene, style, viewpoint, and subject fields while removing the narrative, speech, and sound descriptions; the static fields and the reference observation define the initial scene, leaving the trajectory as the only condition that specifies subsequent camera evolution. For AV-CPT, the full annotation is retained, including narrative, speech, and sound. For Joint-FT, the static fields are combined with speech and sound, while narrative remains excluded. This separation prevents explicit camera-motion leakage while preserving audio conditioning in the stages that optimize the audio-visual backbone.

3.5 Stage-Aligned Data Mixtures

After processing, each example retains the subset of synchronized signals available from its source, including video, structured text, viewpoint metadata, camera intrinsics, audio when present, metric camera trajectory when reliable, and native action logs when available. Action logs and camera trajectories are intentionally preserved as distinct signals: the former record issued control intent, whereas the latter describe motion actually realized under scene geometry and environment dynamics. The current model uses camera trajectories as its unified motion condition, while native action streams remain available for alignment, data analysis, and future interaction modeling.

Rather than sampling uniformly from the processed corpus, we construct three mixtures aligned with the three stages of progressive training. The curriculum progresses from broad audio-visual modeling, to clean trajectory control, and finally to joint consolidation. The AV path supplies the broad AV-rich pool, whereas the geometry path filters for reliable trajectories and supplies the control-clean pool; their high-quality intersection forms the balanced mixture for Joint-FT.

- **AV-rich mixture.** Internally collected gameplay, human-played Internet gameplay, and general Internet video are selected for high visual and acoustic quality, with speech-rich examples explicitly sampled. Reliable camera trajectories are not required, allowing AV-CPT to establish a broad prior over visual appearance, speech, environmental sound, and natural scene dynamics.

- **Control-clean mixture.** Internally collected gameplay, human-played Internet gameplay, UE simulation, and pose-eligible general Internet video are selected for reliable metric trajectories and smooth, interpretable motion. This mixture prioritizes control identifiability and supplies Action-SFT, with the motion-bearing narrative field removed to prevent textual motion leakage.
- **Balanced high-quality mixture.** A smaller intersection of reliable trajectory supervision and high-quality audio-visual content combines clear controllable motion with strong visual quality and, when available, environmental sound and speech. This mixture supports Joint-FT to consolidate trajectory following with the audio-visual prior.

Figure 5 characterizes the motion support supplied by the control-clean mixture. For the coverage analysis, we sample 28,605 trajectories from the control-clean mixture. Their aggregate distribution spans lateral and forward-backward displacement, while representative examples include straight, diagonal, return-loop, and orbiting motion. This coverage exposes Action-SFT to varied realized camera evolution rather than a narrow set of discrete command templates.

Together, the three mixtures form a data curriculum from audio-visual diversity, through control identifiability, to joint audio-visual and interactive modeling.

4 Method

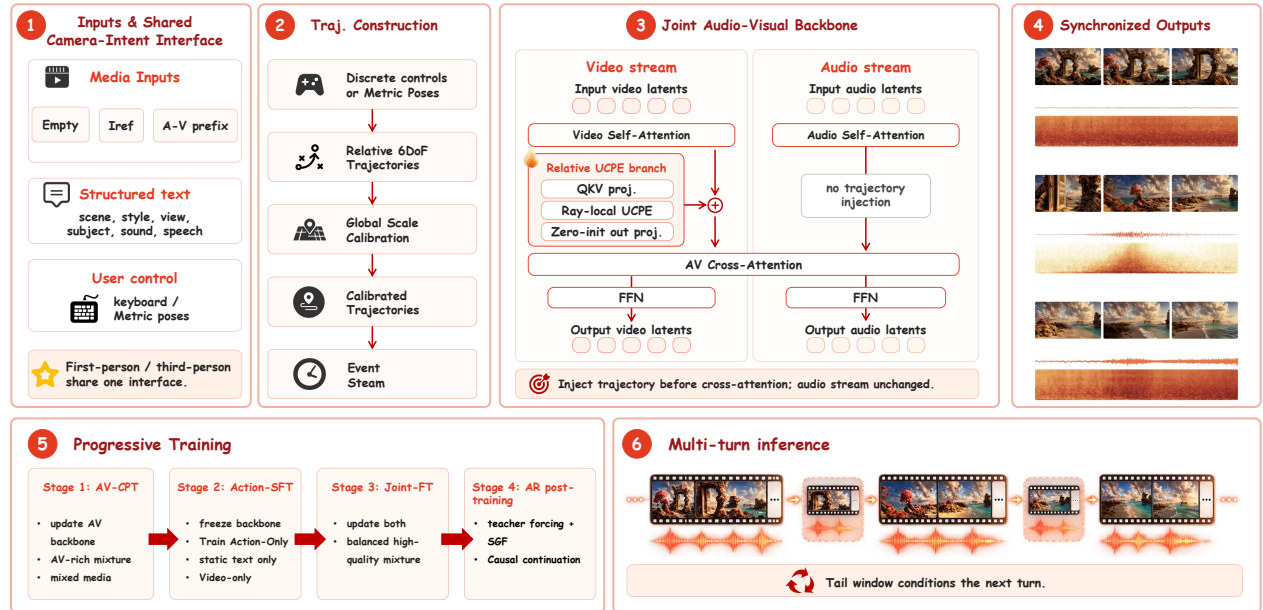


Figure 6 Overview of EchoWM. Media context, structured text, and user controls share a unified camera-intent interface. Discrete controls or metric poses are converted into globally calibrated relative 6-DoF trajectories and serialized as an event stream. A relative UCPE branch injects this trajectory into video self-attention before audio-visual cross-attention, while the audio stream receives no direct trajectory condition. Four progressive training stages yield synchronized audio-visual generation and support multi-turn inference through synchronized tail-window conditioning.

4.1 Overview and Problem Formulation

Given an optional media context $\mathcal{M}$, a structured text condition $\mathcal{C}$ following the annotation schema in section 3, and a user control sequence $U_{1:T}$, our goal is to jointly generate synchronized video and audio while following the requested camera motion. The media context may be empty, a reference observation I_{ref} , or a synchronized audio-visual prefix $(V_{1:\tau}, A_{1:\tau})$.

We convert $U_{1:T}$ into a calibrated relative 6-DoF camera trajectory $P_{1:T}$ and model

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

where θ denotes the joint audio-visual backbone and ϕ denotes the trajectory-conditioning branch. The static text fields specify the scene, style, subject, and initial viewpoint, while $P_{1:T}$ specifies the subsequent camera motion. Following WorldEvent [33], we keep persistent scene descriptions separate from time-varying motion descriptions and provide camera evolution explicitly through the geometric trajectory condition.

We build on a pretrained joint audio-visual diffusion transformer and extend it with metric camera-trajectory conditioning and autoregressive audio-visual generation. Heterogeneous user controls and training camera poses are first converted to a shared relative 6-DoF trajectory with dataset-level translation calibration. The resulting trajectory is injected into the video backbone through a UCPE-based camera branch shared by first- and third-person data. We train the model progressively, first adapting the audio-visual backbone, then learning trajectory conditioning with the backbone frozen, and finally jointly fine-tuning both parameter groups. Autoregressive audio-visual post-training further adapts the model to causal generation from its own history. Figure 6 summarizes the overall framework.

4.2 Metric Camera-Intent Conditioning

We first establish a camera-centric interface that separates user-facing navigation from the representation consumed by the generator. The resulting relative trajectory is then calibrated to a shared metric scale across heterogeneous sources and injected into the video backbone through relative camera geometry. This decomposition keeps control semantics, translation amplitude, and neural conditioning explicit while allowing the same pathway to serve both first- and third-person viewpoints.

4.2.1 Unified Camera-Intent Interface

We represent user navigation through the desired camera motion over time rather than through a system-specific action vocabulary. In first-person scenes, the trajectory directly describes observer ego-motion. In third-person scenes, it specifies camera evolution, while the associated character motion and camera-character coupling are learned from data. The model receives no explicit character trajectory, controller state, or camera-rig parameters. The reference observation and static viewpoint description determine the initial viewing configuration, while the relative trajectory specifies its subsequent evolution.

Let $T_t \in \text{SE}(3)$ denote the camera-to-world pose at frame t . We express camera motion relative to the initial observation as $\Delta T_t = T_0^{-1}T_t$. This representation removes dependence on the global rigid coordinate frame while preserving the continuous magnitude and direction of translation and rotation. Although users may interact through discrete navigation commands, we train the generator on camera trajectories rather than raw action labels for three reasons:

- **Broader data coverage.** Frame-aligned action logs are available only for instrumented gameplay and simulation, whereas camera trajectories can be recovered from a substantially broader range of videos after pose-quality filtering.
- **No controller-dependent inverse mapping.** Recovering pseudo-actions from observed camera motion is generally ambiguous: the same displacement may result from different user inputs, controller dynamics, or camera-smoothing rules. This ambiguity is particularly pronounced in third-person scenes, where character motion and camera-following behavior jointly determine the observed camera trajectory.
- **Continuous 6-DoF motion.** Keyboard inputs form a discrete and system-specific control vocabulary. In contrast, camera trajectories directly preserve continuous translation and rotation, including vertical motion, roll, motion magnitude, and translation-rotation coupling.

At inference time, users may still interact through discrete navigation commands. Let $\mathbf{u}_t$ denote the keyboard state at time t . A fixed mapping M , parameterized by the configured per-step translation and rotation

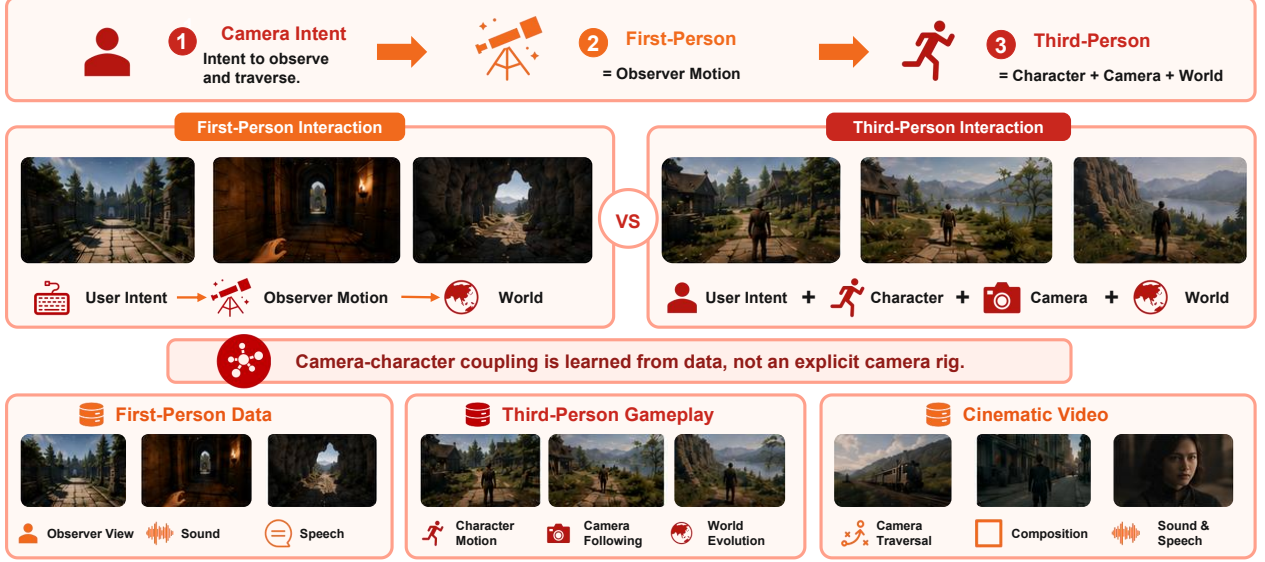


Figure 7 Camera intent unifies first- and third-person interaction. The same user intent is realized as observer motion in first-person scenes and as coordinated camera–character–world evolution in third-person scenes. The coupling is learned from heterogeneous data rather than specified by an explicit camera rig.

magnitudes, converts the active commands into a local 6-DoF increment

$$\xi_t = M \mathbf{u}_t, \quad \xi_t = [\mathbf{v}_t^\top, \boldsymbol{\omega}_t^\top]^\top, \quad (2)$$

where $\mathbf{v}_t$ and $\boldsymbol{\omega}_t$ denote translational and rotational increments in the current camera frame, respectively. We then integrate these local increments and express the resulting trajectory relative to its initial pose:

$$T_t = T_{t-1} \text{Exp}(\hat{\xi}_t), \quad \Delta T_t = T_0^{-1} T_t, \quad (3)$$

where $\hat{\xi}_t \in \mathfrak{se}(3)$ is the matrix representation of the 6-DoF increment. A continuous metric pose sequence bypasses the discrete mapping and integration, but undergoes the same coordinate conversion and relative transformation. Training trajectories recovered in [section 3.3](#) and inference-time navigation commands therefore share the same relative trajectory representation, which is subsequently calibrated as described in [section 4.2.2](#). This interface separates the user-facing control format from the geometric condition consumed by the generator, allowing different navigation interfaces and heterogeneous training sources to share the same trajectory representation. Its scope is limited to navigation and viewpoint-related controls whose effects can be represented by camera motion; semantic actor actions such as attacking, jumping, or object manipulation are not modeled by this interface.

4.2.2 Global Translation Scale Calibration

Relative poses remove dependence on the global rigid coordinate frame but do not resolve translation-scale differences across heterogeneous data sources. For trajectory conditioning, the same numerical displacement should correspond to a comparable physical camera motion across examples. Per-clip or per-chunk normalization would erase meaningful displacement differences and can introduce artificial scale changes at continuation boundaries. Using the largest displacement in the corpus is instead sensitive to reconstruction outliers. We therefore estimate one robust scale from the metric training trajectories.

For clip i , let $\Delta T_{i,k} = [\Delta \mathbf{R}_{i,k} \mid \Delta \mathbf{t}_{i,k}]$ denote the relative pose at frame k . We compute

$$m_i = \max_k \|\Delta \mathbf{t}_{i,k}\|_2, \quad s_{\text{global}} = Q_{0.9}(\{m_i\}_{i \in \mathcal{D}_{\text{train}}}), \quad (4)$$

where $Q_{0.9}$ denotes the 90th percentile over the training set. Because trajectory-conditioned training uses fixed-duration clips at a common temporal sampling rate, m_i provides a comparable measure of clip-level translation extent across sources. We retain trajectories whose maximum displacement does not exceed s_{global} rather than clipping larger trajectories, which would alter their displacement magnitude. For each retained trajectory,

$$\hat{\mathbf{t}}_{i,k} = \frac{\Delta \mathbf{t}_{i,k}}{s_{\text{global}}}, \quad P_{i,k} = \left[\Delta \mathbf{R}_{i,k} \mid \hat{\mathbf{t}}_{i,k} \right]. \quad (5)$$

Rotations remain unchanged, and the same s_{global} is used for all training and inference trajectories. Together with the common temporal sampling rate, this preserves displacement differences across examples and avoids scale-induced changes in motion speed at continuation boundaries.

4.2.3 Relative Trajectory Conditioning

Given the calibrated trajectory $P_{1:T}$, we next inject camera geometry into the video backbone. In our comparison setup, Plücker-based conditioning represents the world-space ray associated with spatial cell s at frame t as

$$\mathcal{P}_{t,s} = (\mathbf{d}_{t,s}, \mathbf{m}_{t,s}), \quad \mathbf{m}_{t,s} = \mathbf{o}_t \times \mathbf{d}_{t,s}, \quad (6)$$

where $\mathbf{d}_{t,s}$ is the ray direction and $\mathbf{o}_t$ is the camera center. Because the moment term explicitly contains $\mathbf{o}_t$, this absolute representation changes with the chosen world origin. The baseline additionally materializes dense ray features and uses a separate camera encoder to learn relations across frames. We instead use Unified Camera Positional Encoding (UCPE) [50], which introduces relative ray geometry directly inside attention. For a global rigid transformation G , $(GT_i)^{-1}(GT_j) = T_i^{-1}T_j$, so pairwise camera geometry is unchanged by a global rigid change of coordinates. Translation scale is handled separately by the calibration in section 4.2.2. Let $i = (t, s)$ index a video latent token at frame t and spatial cell s , with homogeneous image coordinate $\mathbf{p}_s$. We interpret the calibrated pose $P_t = [\mathbf{R}_t \mid \mathbf{o}_t]$ as a camera-to-reference transform, with the first camera defining the reference frame. Given this pose and intrinsics K_t , the corresponding reference-space ray direction is

$$\mathbf{d}_{t,s} = \mathbf{R}_t \text{norm}(K_t^{-1} \mathbf{p}_s). \quad (7)$$

Following UCPE, we construct a local ray frame using this direction and the camera image axes. Let $\mathbf{v}_t^{\text{down}}$ denote the image-down direction:

$$\begin{aligned} \mathbf{z}_{t,s} &= \mathbf{d}_{t,s}, \\ \mathbf{x}_{t,s} &= \text{norm}(\mathbf{v}_t^{\text{down}} \times \mathbf{z}_{t,s}), \\ \mathbf{y}_{t,s} &= \mathbf{z}_{t,s} \times \mathbf{x}_{t,s}. \end{aligned} \quad (8)$$

Together with camera center $\mathbf{o}_t$, this basis defines the ray-to-reference transform $\mathbf{T}_i^{\text{wr}}$. Its inverse $\mathbf{D}_i = (\mathbf{T}_i^{\text{wr}})^{-1}$ maps reference coordinates into the local ray frame. We inject this geometry through a lightweight camera-attention branch parallel to the original video self-attention. For a head of width d , half of the channels use the tiled ray transform $\mathbf{G}_i = \mathbf{I}_{d/8} \otimes \mathbf{D}_i$, while the remaining channels retain spatiotemporal RoPE. With branch-specific QKV projections, the camera attention is

$$\begin{aligned} \tilde{\mathbf{q}}_i^c &= (\mathbf{G}_i^\top \oplus \text{RoPE}_i) \mathbf{q}_i^c, \\ (\tilde{\mathbf{k}}_i^c, \tilde{\mathbf{v}}_i^c) &= (\mathbf{G}_i^{-1} \oplus \text{RoPE}_i) (\mathbf{k}_i^c, \mathbf{v}_i^c), \\ \mathbf{o}_i^c &= (\mathbf{G}_i \oplus \text{RoPE}_i^{-1}) \text{Attn}_{\text{cam}}(\tilde{\mathbf{q}}^c, \tilde{\mathbf{k}}^c, \tilde{\mathbf{v}}^c)_i. \end{aligned} \quad (9)$$

The resulting query-key interaction depends on the relative ray transform $\mathbf{G}_i \mathbf{G}_j^{-1}$ between tokens i and j . The camera-branch output is merged into the original video attention through a zero-initialized projection. The branch-specific QKV and output projections constitute ϕ . Zero initialization preserves the pretrained video pathway at the start of trajectory fine-tuning and lets the branch learn only the residual required for camera conditioning. The same branch is shared across first- and third-person data. The static viewpoint field specifies the initial observer-subject relationship, while $P_{1:T}$ specifies subsequent camera motion. No view-specific network or discrete action embedding is introduced. The camera branch operates only on video tokens; audio tokens receive no direct trajectory conditioning. We use the relative-ray component of UCPE and omit its optional Lat-Up encoding, since the trajectory is defined relative to the initial camera frame.

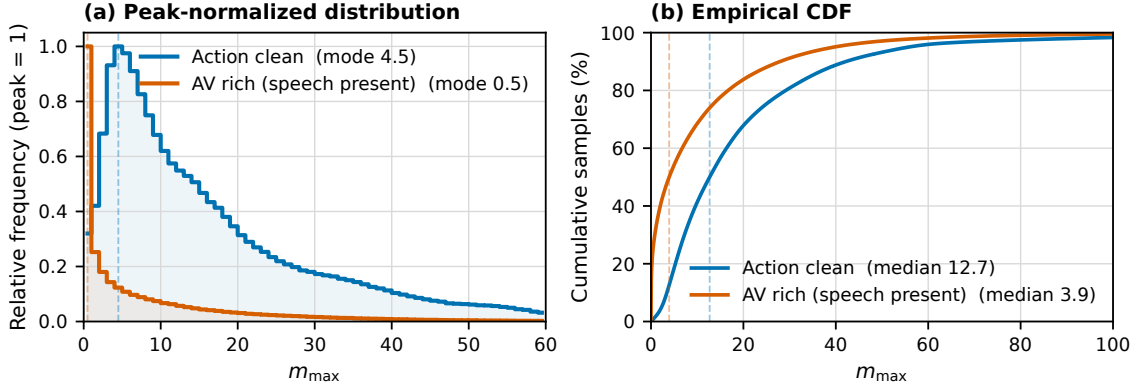


Figure 8 Pose-magnitude distributions for global scale calibration. The plots show the peak-normalized distribution and empirical CDF of the per-clip maximum translation magnitude m_i for the action-clean and AV-rich speech-present subsets. Action-clean clips contain broader motion magnitudes (mode 4.5, median 12.7), whereas speech-present AV-rich clips concentrate at smaller motions (mode 0.5, median 3.9). The heterogeneous support motivates one robust dataset-level percentile scale rather than independently normalizing each source.

4.3 Progressive Audio-Visual Control Training

Reliable audio-visual supervision and reliable trajectory supervision are concentrated in different parts of the training data. As shown in figure 8, AV-rich examples provide stronger audio supervision but typically cover a narrower range of camera motion, whereas control-clean examples contain more reliable and diverse trajectories. We therefore separate training into three stages. AV-CPT first adapts the joint audio-visual backbone, Action-SFT then learns the trajectory-conditioning branch with the backbone frozen, and Joint-FT finally updates both parameter groups on the balanced high-quality subset.

Stage 1: Audio-Visual Continued Pretraining. The pretrained model already supports joint audio-visual generation, but the interactive training data differ in visual appearance, motion statistics, speech, ambience, and interaction-related sound. We first adapt the complete backbone θ on the AV-rich mixture without trajectory conditioning. The full structured annotation is retained, including narrative, speech, and sound. For each example, the clean media context $\mathcal{M}$ is sampled from an empty condition, a reference frame I_{ref} , or a synchronized audio-visual prefix $(V_{1:\tau}, A_{1:\tau})$. These cases train text-conditioned generation, reference-frame-conditioned generation, and synchronized continuation, respectively. When a reference frame or prefix is provided, the exposed tokens remain clean and are excluded from the denoising loss. We optimize

$$\theta_{\text{AV}} = \arg \min_{\theta} \mathbb{E}_{(V,A,C_{\text{AV}}) \sim \mathcal{D}_{\text{AV}}, \mathcal{M} \sim q_{\text{cond}}(\cdot | V,A)} \left[\mathcal{L}_V(V | C_{\text{AV}}, \mathcal{M}; \theta) + \lambda_A \mathcal{L}_A(A | C_{\text{AV}}, \mathcal{M}; \theta) \right], \quad (10)$$

where both losses are evaluated only on regions selected for generation by the sampled media condition.

Stage 2: Action Fine-Tuning. We refer to this stage as Action-SFT because it trains responsiveness to the user-facing navigation interface, although supervision is provided by camera trajectories rather than raw keyboard labels. Starting from θ_{AV} , we freeze the audio-visual backbone and train only the camera-branch QKV projections and zero-initialized output projection, collectively denoted by ϕ . Training uses the control-clean mixture, which favors reliable metric trajectories and visually unambiguous camera motion. To prevent text from directly specifying the target motion, we retain only the static annotation fields $\mathcal{C}_{\text{static}} = \{c_{\text{scene}}, c_{\text{style}}, c_{\text{view}}, c_{\text{subject}}\}$. The narrative field and other motion-dependent descriptions are removed, so $P_{1:T}$ is the only condition that specifies the target camera evolution. We disable the audio loss and sample $\mathcal{M}_{\text{SFT}} \in \{\emptyset, I_{\text{ref}}\}$, with reference-frame-conditioned examples sampled more frequently than unconditional ones. The reference frame constrains scene appearance and the initial viewpoint, leaving the

trajectory to specify subsequent camera motion. We optimize

$$\phi_{\text{SFT}} = \arg \min_{\phi} \mathbb{E}_{\mathcal{M}_{\text{SFT}} \sim q_{\text{SFT}}} [\mathcal{L}_V(V \mid \mathcal{M}_{\text{SFT}}, \mathcal{C}_{\text{static}}, P_{1:T}; \theta_{\text{AV}}, \phi)]. \quad (11)$$

The trajectory is used only as a conditioning signal and is not itself a prediction target. No explicit action-to-audio objective is introduced in this stage.

Stage 3: Joint Fine-Tuning. AV-CPT and Action-SFT optimize different parameter subsets on different data mixtures. We therefore perform a final Joint-FT stage on the smaller balanced high-quality subset, for which both camera trajectories and audio-visual signals are reliable.

The text condition retains the static fields together with speech and sound, while the narrative field remains excluded. Starting from $(\theta_{\text{AV}}, \phi_{\text{SFT}})$, we update both parameter groups with a reduced learning rate:

$$(\theta^*, \phi^*) = \arg \min_{\theta, \phi} \mathbb{E} [\mathcal{L}_V(\theta, \phi; P_{1:T}) + \lambda_A \mathcal{L}_A(\theta, \phi)]. \quad (12)$$

The camera branch remains restricted to the video stream. The reduced learning rate allows the backbone and trajectory branch to be updated jointly while limiting disruption to the audio-visual and control capabilities learned in the preceding stages.

4.4 Streaming Audio-Visual Post-Training

We progressively adapt a pretrained bidirectional, multi-step audio-visual diffusion model into a high-throughput, low-latency streaming autoregressive generator. This transition requires two key transformations: converting bidirectional temporal attention into causal autoregressive computation for online generation of synchronized audio-video 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, 38] to initialize a causal streaming generator from the bidirectional backbone. We then introduce short-horizon Self-Gradient Forcing (SGF) [16, 55] 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. All autoregressive post-training stages retain the trajectory-conditioning branch and jointly update the generator parameters (θ, ϕ) unless otherwise stated.

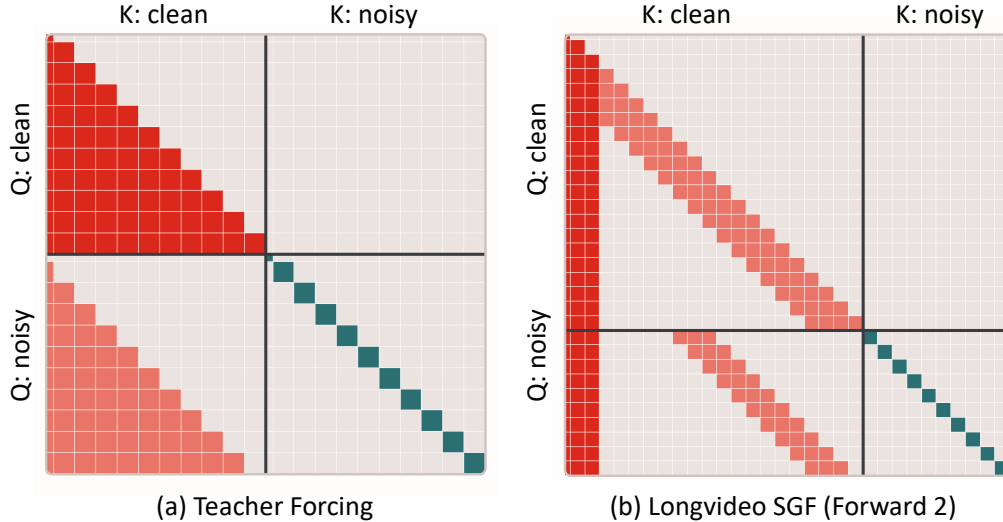


Figure 9 Chunk-level causal attention patterns used for streaming post-training. (a) The audio-visual teacher-forcing pattern used by short-horizon reconstruction: each noisy query attends to the preceding clean context and the noisy tokens of its current chunk, while clean targets and future chunks are excluded. (b) The mask used by Forward 2 of long-horizon SGF, where access to clean history is additionally constrained by the sink-plus-FIFO KV-cache policy.

4.4.1 Audio-Visual Teacher Forcing

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

$$x_{\sigma_i, i}^m = (1 - \sigma_i)x_i^m + \sigma_i\epsilon_i^m, \quad \epsilon_i^m \sim \mathcal{N}(0, I), \quad (13)$$

where ϵ_i^m is Gaussian noise and 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:

$$\mathcal{S}(\mathcal{B}_{i, \text{clean}}) = \{\mathcal{B}_{j, \text{clean}} : j \leq i\}, \quad \mathcal{S}(\mathcal{B}_{i, \text{noisy}}) = \{\mathcal{B}_{j, \text{clean}} : j < i\} \cup \{\mathcal{B}_{i, \text{noisy}}\}, \quad (14)$$

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-video cross-attention. As illustrated in Fig. 9(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.

We additionally condition the model on a global condition $\mathcal{C}$ and a temporally aligned trajectory sequence $P^{(1:N)}$, where $P^{(i)}$ denotes the trajectory segment associated with chunk i . Consistent with the autoregressive generation process, chunk i only uses the trajectory available up to the current chunk, denoted by $P^{(\leq i)}$.

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 model:

$$\mathcal{L}_{\text{TF}} = \sum_{m \in \{v, a\}} \lambda_m \mathbb{E} \left[\sum_{i=1}^N \left\| v_{\theta}^m(x_{\sigma_i, i}^m, \sigma_i \mid \mathcal{C}, P^{(\leq i)}) - (\epsilon_i^m - x_i^m) \right\|_2^2 \right], \quad (15)$$

where v_{θ}^m denotes the predicted flow velocity for modality m , λ_m balances the video and audio losses, and θ denotes the generator parameters. 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.

4.4.2 Short-Horizon Audio-Visual Self-Gradient Forcing

The causal teacher-forcing model provides an effective initialization for autoregressive generation, but two discrepancies remain within the temporal horizon of the base model. First, teacher-forcing training conditions the generator on ground-truth histories, whereas inference relies on its own generated outputs. Second, the original diffusion model requires multiple denoising steps for each audio-visual chunk. We address these discrepancies using Self-Gradient Forcing (SGF) [55], which trains on self-generated autoregressive histories while recovering gradients through their reconstructed causal representations. We additionally apply Distribution Matching Distillation (DMD) [46] to convert the original multi-step generator into a few-step sampler.

Forward 1: autoregressive self-rollout. We first run a complete chunk-wise rollout without gradient tracking. This rollout uses the same few-step sampler, causal attention configuration, and trajectory conditioning as autoregressive inference. When generating chunk i , the model is conditioned on the currently available trajectory $P^{(\leq i)}$. For instance, with a four-step sampler, the model performs four denoising steps to

produce each synchronized audio-visual chunk and then executes an additional clean-context forward pass at $t = 0$ to populate the causal KV cache used by subsequent chunks.

An exit step t_e is sampled during the rollout. For every chunk i , we record the noisy audio-visual state at this step, $r_i = (r_i^v, r_i^a)$, together with the final generated clean latents $\hat{x}_i = (\hat{x}_i^v, \hat{x}_i^a)$. All operations in this forward pass are performed under `no_grad`. Forward 1 therefore captures the autoregressive states encountered during few-step inference without retaining the computation graph of the sequential sampling process.

Forward 2: audio-visual context-gradient reconstruction. We subsequently replay the recorded exit-step computation in a differentiable parallel forward pass. The generated clean latents $\{\hat{x}_i\}_{i=1}^N$ are treated as stop-gradient inputs and evaluated at the clean context timestep $t = 0$, while the recorded noisy states $\{r_i\}_{i=1}^N$ retain their sampled timestep t_e . Clean-context tokens and noisy target tokens are then arranged using the same causal construction as Eq. (14).

The resulting attention pattern is shown in Fig. 9(a). A noisy query associated with chunk i can attend to the clean history from preceding chunks and to the noisy tokens of the current chunk. Clean targets, future clean chunks, and future noisy chunks remain inaccessible. Clean-context queries are likewise restricted to causal clean history. This mask reproduces the context–target dependencies of the serial rollout while allowing the exit-step predictions for all chunks to be reconstructed in parallel.

In our LTX-2.3-based audio-visual generator, this chunk-level causal mask is applied to all five attention operations that involve temporally indexed rollout information: video self-attention, audio self-attention, audio-to-video (A2V) attention, video-to-audio (V2A) attention, and UCPE attention. The video and audio self-attention masks enforce causality within each modality, while the A2V and V2A masks impose the same temporal constraint on cross-modal information exchange. The UCPE attention follows the corresponding chunk-level visibility pattern for trajectory-conditioned features. Text cross-attention is unchanged because the global text condition is available throughout the complete rollout.

Although the clean and noisy latents recorded in Forward 1 are detached, the generator recomputes their context representations with gradient tracking in Forward 2. In particular, the clean-context forward reconstructs the attention projections and K/V representations through which generated audio-visual chunks are encoded as causal history. A loss on a later chunk can therefore propagate through its causal attention operations into the reconstructed K/V representations of preceding generated chunks. This restores supervision for the audio-visual context-writing computation while avoiding backpropagation through the denoising decisions and serial cache updates of Forward 1. The reconstructed clean predictions are denoted by $\{\bar{x}_i\}_{i=1}^N$, where $\bar{x}_i = (\bar{x}_i^v, \bar{x}_i^a)$.

Distribution-matching objective. We apply DMD to the recovered predictions under the self-generated autoregressive histories. Unlike teacher-forcing supervision, these predictions are conditioned on the generator’s own preceding audio-visual outputs. The resulting objective therefore combines few-step diffusion distillation with adaptation to the history distribution encountered during autoregressive inference.

For a sampled diffusion timestep s , the recovered clean audio-visual latent is perturbed as

$$\bar{x}_{s,i} = (1 - s)\bar{x}_i + s\eta_i, \quad \eta_i \sim \mathcal{N}(0, I), \quad (16)$$

where η_i is independently sampled Gaussian noise. The generator is initialized from the causal model obtained after teacher forcing. Both the real-score model s_{real} and the fake-score model s_{fake} are initialized from the pretrained bidirectional diffusion model. The real-score model remains frozen and represents the reference distribution, whereas the fake-score model is trainable and tracks the evolving distribution of the few-step generator.

The generator is optimized using

$$\mathcal{L}_{\text{DMD}} = \mathbb{E}_{i,s,\eta_i} [w(s) \langle \text{sg}(s_{\text{fake}}(\bar{x}_{s,i}, s) - s_{\text{real}}(\bar{x}_{s,i}, s)), \bar{x}_i \rangle], \quad (17)$$

where $\text{sg}(\cdot)$ denotes stop-gradient and $w(s)$ is a diffusion-dependent weighting factor. For clarity, the global condition $\mathcal{C}$ and trajectory condition $P^{(\leq i)}$ are omitted from the score-model notation. The score difference directs the few-step generator toward the distribution represented by the pretrained multi-step model, while

the dependence of $\bar{x}_i$ on the differentiable reconstruction preserves gradients through the causal context representations.

Fake-score updates. Following each generator update, we perform multiple updates of the fake-score model. Generated samples are detached and used to optimize s_{fake} with the standard flow-matching objective, enabling it to follow the changing generator distribution.

Short-horizon audio-visual SGF thus exposes the generator to its own autoregressive histories, restores supervision through the reconstructed causal context, and establishes few-step generation. The resulting model serves as the initialization for long-horizon streaming training.

4.4.3 Long-Horizon Audio-Visual Self-Gradient Forcing

Short-horizon SGF adapts the generator to self-generated states within the temporal range of the base model. During sustained streaming generation, however, the model must operate on histories containing errors accumulated over a much longer sequence. We therefore continue self-rollout training on substantially longer synchronized audio-visual trajectories. The few-step sampler and DMD objective are retained from the short-horizon stage, while the extended rollout exposes the model to progressively shifted autoregressive states.

Long-horizon rollout. A long-horizon trajectory is formed by sequentially generating multiple temporal segments, each using the rollout length adopted during short-horizon SGF. To maintain continuity between adjacent segments, the terminal video frame of segment k is used as the initial frame of segment $k + 1$. The model does not restart from an independent visual state at each boundary. Instead, later segments are generated from the preceding self-generated trajectory and therefore inherit the prediction errors accumulated in earlier segments. This provides training exposure to the states encountered during sustained world-model rollout.

Bounded causal context. A full causal KV cache would grow continuously with the rollout duration. We instead maintain a sink-plus-FIFO cache for each modality. Let S_m denote the number of persistent sink chunks and R_m the number of recent chunks retained for modality m . When generating chunk i , the available historical context is

$$\mathcal{H}_i^m = \underbrace{(\hat{x}_1^m, \dots, \hat{x}_{\min(S_m, i-1)}^m)}_{\text{attention sink}} \parallel \underbrace{(\hat{x}_{\max(S_m+1, i-R_m)}^m, \dots, \hat{x}_{i-1}^m)}_{\text{recent FIFO history}}, \quad (18)$$

where $\parallel$ denotes ordered concatenation. The sink retains a fixed prefix of the generated trajectory, while the FIFO component preserves the most recent context. Consequently, the number of historical chunks directly available at each attention layer is bounded by $S_m + R_m$, regardless of the total rollout duration.

Forward 2 reconstructs the long trajectory using the attention pattern shown in Fig. 9(b). Relative to the short-horizon mask in Fig. 9(a), the clean causal history available to each query is further restricted by the sink-plus-FIFO policy. A noisy query attends to the retained clean sink and FIFO context, together with the noisy tokens of its current chunk. The same visibility pattern is used during autoregressive inference and during the differentiable reconstruction.

Layer-wise long-range context gradients. The sink-plus-FIFO window limits the K/V entries that a query accesses directly at a single Transformer layer, but it does not confine the effective context gradient to only those entries. A retained FIFO representation at an upper layer has already incorporated information from the causal context visible to that chunk in preceding layers. During backpropagation, a loss on a later chunk first reaches the sink and FIFO K/V representations visible at the current layer. The gradient can then continue through the lower-layer computations that produced these representations and reach earlier K/V representations through intermediate chunks. Stacking the causal attention layers therefore produces a multi-layer gradient path whose effective temporal reach can be longer than the direct sink-plus-FIFO window of any individual layer.

This layer-wise propagation does not introduce backpropagation through the serial rollout. The latents, denoising states, and KV-cache trajectory recorded in Forward 1 remain detached. Gradients are propagated

only through the attention and context representations recomputed in Forward 2. Thus, the cache policy keeps the direct attention fan-in bounded, while the reconstructed multi-layer computation allows later losses to provide supervision beyond the K/V entries directly visible at the current layer.

Score-model evaluation range. The generator retains its original RoPE configuration throughout long-horizon rollout, independently of the sink-plus-FIFO cache policy. Because the real-score and fake-score models preserve the bidirectional architecture of the pretrained fixed-length diffusion model, their DMD evaluations are restricted to the temporal range observed during pretraining.

Long-horizon DMD optimization. We apply DMD supervision to every rollout segment rather than only to the final part of the trajectory. Let $\mathcal{L}_{\text{DMD}}^{(k)}$ denote the DMD loss computed from the recovered audio-visual predictions of segment k . For a trajectory containing K rollout segments, the training objective is

$$\mathcal{L}_{\text{Long-SGF}} = \sum_{k=1}^K \mathcal{L}_{\text{DMD}}^{(k)}. \quad (19)$$

Each segment is consequently supervised at a different stage of the self-generated trajectory. Losses on later segments are evaluated under histories that already reflect the errors accumulated in preceding segments. Together with the reconstructed multi-layer context gradients, this training adapts the few-step generator to the progressively changing state distribution encountered during long-horizon world-model rollout.

4.5 Interactive Inference

Bidirectional multi-turn continuation. At inference time, multi-turn generation reuses synchronized video and audio from the preceding segment as clean context for the next turn. For turn $j > 1$, we extract

$$\mathcal{M}^{(j-1)} = \text{Tail}_w(V^{(j-1)}, A^{(j-1)}), \quad (20)$$

and generate

$$(V^{(j)}, A^{(j)}) \sim p_{\theta^{\text{PT}}, \phi^{\text{PT}}}(\cdot \mid \mathcal{M}^{(j-1)}, \mathcal{C}^{(j)}, P_{1:T}^{(j)}). \quad (21)$$

Each turn may specify a different navigation trajectory. Because the context window contains synchronized video and audio, visual and acoustic continuation are conditioned on the same recent history. The generated segment then provides the tail context for the following turn. This tail-reencoding procedure is the bidirectional continuation mode. We use a separate persistent-state procedure for causal streaming inference, described next, so the two modes differ in how they manage historical context rather than in the trajectory interface or model parameterization.

Causal streaming inference. We keep a persistent streaming KV cache instead of restarting from a decoded tail. At inference, we maintain separate streaming KV caches for video and audio, along with the camera-attention KV states associated with video tokens, each following the same sink-plus-local structure used during training. The sink retains a persistent long-term prefix, while the local window stores the most recent context. As old local tokens are removed, we rebase temporal RoPE and camera encodings over the retained cache to maintain a consistent positional layout.

Let $\mathcal{K}^{(t)}$ denote the retained cache state before the next block. Given text condition $\mathcal{C}$ and trajectory P , we sample

$$(V, A) \sim p_{\theta^{\text{PT}}, \phi^{\text{PT}}}(\cdot \mid \mathcal{K}^{(t)}, \mathcal{C}, P). \quad (22)$$

The persistent sink provides long-term context throughout the autoregressive rollout without requiring an external memory bank.

5 Evaluation

We evaluate ECHOWM on two public benchmarks for direct comparison with recent world models. WBench Navigation measures broad interactive-world performance across visual quality, setting adherence, interaction, consistency, and physical plausibility, whereas SANA-WM-Bench provides a more focused evaluation of short-

and long-horizon camera-trajectory following and visual persistence. Together, these benchmarks provide complementary evidence for generation quality, controllability, and long-horizon world consistency. We further supplement the quantitative results with qualitative analysis, human evaluation, and ablation studies.

5.1 WBench Navigation

Table 3 Comparison on the 158-case WBench Navigation split [48]. Results are sorted by Average.

Model	Avg.	Quality	Setting	Interaction	Consistency	Physical
ECHOWM	81.7	81.5	79.4	<u>87.2</u>	<u>89.8</u>	70.6
ECHOWM-Flash	<u>81.0</u>	81.1	88.3	87.9	77.5	70.1
HiDream-O1-World [48]	80.9	81.0	82.2	80.0	88.0	73.3
Alaya-EVOKE [48]	80.8	82.8	83.8	78.6	86.9	<u>72.1</u>
LingBot-World (fast v2) [27]	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) [27]	78.5	78.9	72.6	80.1	89.9	71.2
Wan 2.7 [32]	78.1	81.5	<u>91.4</u>	64.4	81.6	71.8
HY-World 1.5 (AR-distill) [31]	78.1	78.1	72.2	86.8	86.9	66.3
HY-Video 1.5 [39]	77.9	77.6	85.6	71.4	87.4	67.4
LingBot-World (fast) [27]	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) [29]	76.4	77.1	73.2	85.6	79.3	66.7
Seedance 1.5 [5]	76.2	<u>82.1</u>	82.9	66.3	81.3	68.4
Cosmos3-Super [23]	76.0	76.4	91.6	58.0	85.0	69.2
SANA-WM (4-step AR) [54]	76.0	79.3	76.1	82.2	80.7	61.9
DreamX-World (5B AR) [4]	75.0	77.5	80.8	78.6	74.9	63.3
ABot-World [17]	74.7	76.8	71.4	84.0	79.5	61.7
Cosmos 2.5 [22]	74.6	72.9	83.3	63.1	86.5	67.4
Cosmos3-Nano [23]	74.2	77.4	87.3	59.1	83.7	63.6
LTX-2.3 [9]	74.2	77.1	85.2	66.4	77.2	64.9
Genie 3 [6]	73.9	75.2	72.5	73.4	82.6	65.7

Evaluation protocol. We evaluate both ECHOWM and its distilled four-step causal variant, ECHOWM-Flash, on the official 158-case Navigation split of WBench [48], covering first- and third-person interactive navigation. We convert the supported control interfaces to our shared internal trajectory condition and follow the official multi-turn evaluation protocol. We report all five official dimensions and their overall Average: Quality measures perceptual video quality; Setting measures scene and subject adherence; Interaction measures navigation and action following; Consistency measures spatial, temporal, and viewpoint persistence; and Physical measures causal fidelity and visual plausibility. All scores follow the official 0–100 normalization and aggregation, with higher values indicating better performance.

Quantitative results. As shown in table 3, both variants rank at the top of the WBench Navigation benchmark, demonstrating strong interactive world-modeling capability before and after causal distillation. The undistilled multi-step ECHOWM achieves the highest overall Average score of 81.7 and a strong Consistency score of 89.8, showing navigation performance together with stable world-state preservation across multi-turn interactions. Building on this strong base model, ECHOWM-Flash closely follows with an Average score of 81.0 and achieves the highest Interaction score of 87.9, slightly surpassing the undistilled model at 87.2. Despite replacing bidirectional multi-step generation with causal four-step inference, ECHOWM-Flash therefore retains most of the original model’s overall capability while preserving, and even slightly improving, control responsiveness. Together, these results establish ECHOWM as a strong interactive world model and show that the proposed distillation effectively transfers its capability to the more efficient causal few-step variant.

Qualitative results. As illustrated in figures 10 and 12, competing world models often exhibit abrupt



Figure 10 WBench Navigation comparison.

viewpoint changes, unintended zooming, subject-scale variation, or scene-layout drift under the same initial observation and navigation condition. In contrast, ECHOWM more faithfully follows the prescribed navigation while better preserving scene structure, subject appearance, and visual style across successive interactions. As shown in figure 13, we further evaluate ECHOWM-Flash under repeated interactions and extended autoregressive rollout using three representative multi-turn cases. The action sequences are $A \rightarrow A \rightarrow A \rightarrow A$, $Left \rightarrow Left \rightarrow Right \rightarrow Right$, and $Left \rightarrow W \rightarrow Right$, respectively. Here, *Left*, *Right*, *Up*, and *Down* control camera rotation, while *W*, *S*, *A*, and *D* control camera translation. Compared with representative recent causal world models, ECHOWM-Flash follows these turn-wise control changes more consistently while exhibiting less accumulated drift in viewpoint, scene layout, subject identity, and visual style. Competitive baselines such as LingBot2-Fast and Evoke remain coherent over multiple turns, but still show noticeable action deviation and gradual degradation in subject identity or visual style as the rollout progresses. Overall, ECHOWM-Flash better preserves both controllability and visual consistency over extended autoregressive generation, indicating improved robustness to long-horizon error accumulation.

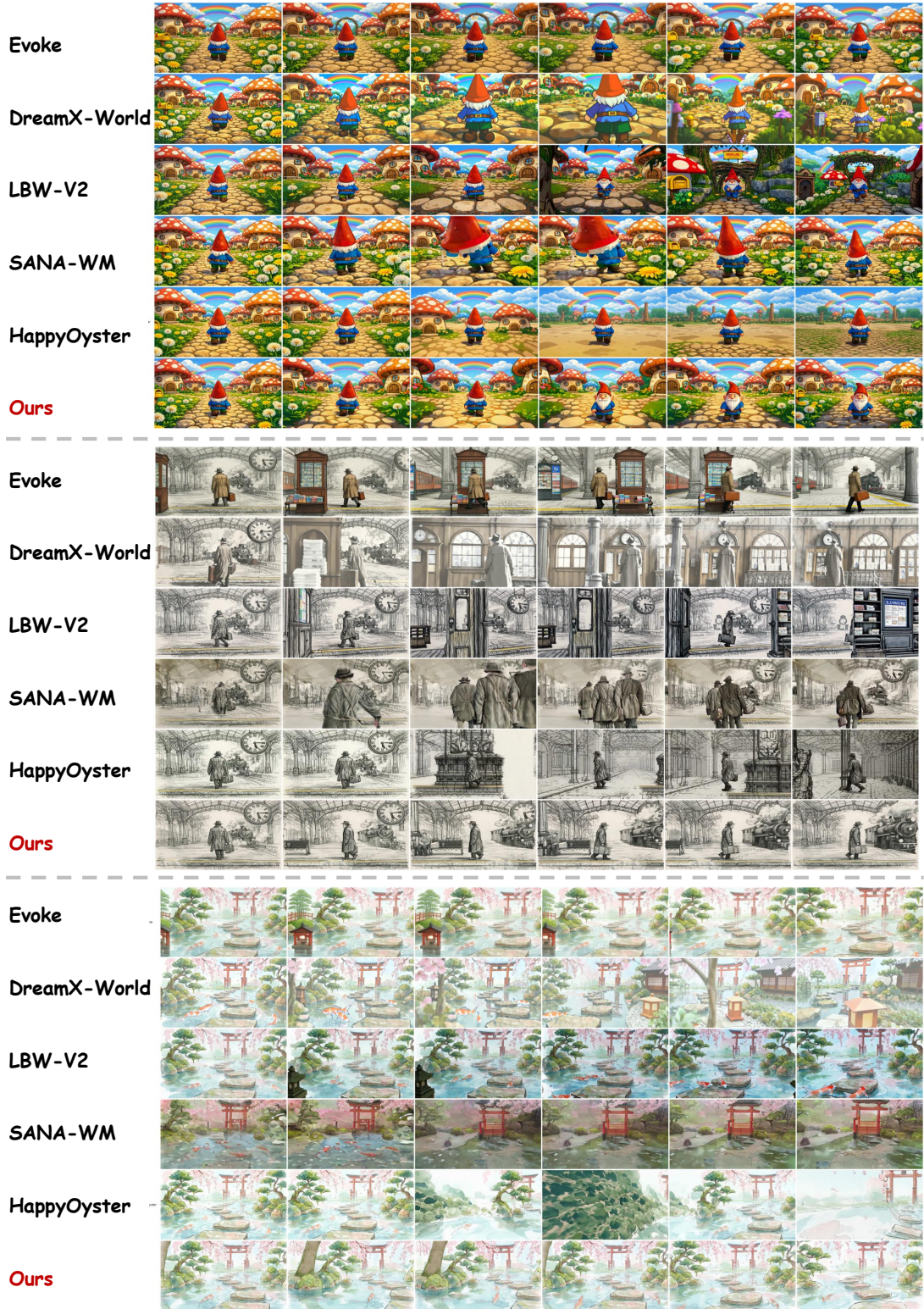


Figure 11 WBench Navigation comparison.



Figure 12 WBench Navigation comparison.



Figure 13 Qualitative comparison of causal world models on WBench Navigation. ECHO-WM-Flash shows better action adherence and multi-turn consistency in scene structure, subject identity, and visual style.

5.2 SANA-WM-Bench

Table 4 Short-horizon comparison on SANA-WM-Bench using the front 241 frames of each trajectory. Each split contains 80 scenes. VBench Overall follows the benchmark convention of reporting normalized visual quality on a 0–100 scale. Camera accuracy reports rotation error R in degrees, relative translation error T , and camera-motion consistency error CMC after similarity alignment.

Method	VBench Overall $\uparrow$	$R \downarrow$	$T \downarrow$	CMC $\downarrow$
<i>Simple-Trajectory Split</i>				
Matrix-Game 3 [37]	79.41	4.474	0.406	0.453
LingBot-World-v2 [27]	81.13	3.092	0.282	0.310
HY-WorldPlay [31]	82.05	2.256	0.284	0.298
DreamX-World [4]	<u>82.42</u>	1.839	0.261	0.273
SANA-WM [54]	81.75	0.489	<u>0.194</u>	<u>0.195</u>
ECHOWM	83.91	<u>0.523</u>	0.160	0.162
<i>Hard-Trajectory Split</i>				
Matrix-Game 3 [37]	79.37	12.819	0.583	0.717
LingBot-World-v2 [27]	81.67	6.367	0.357	0.426
HY-WorldPlay [31]	81.24	6.346	0.387	0.437
DreamX-World [4]	<u>82.03</u>	8.804	0.467	0.556
SANA-WM [54]	81.79	1.248	0.199	0.206
ECHOWM	83.96	<u>1.697</u>	<u>0.256</u>	<u>0.266</u>

Table 5 Long-horizon comparison on the full SANA-WM-Bench trajectory. Revisit consistency is measured on same-pose pairs. VBench Overall is the normalized Quality component. Temporal degradation is $\Delta IQ = IQ_{0-10} - IQ_{50-60}$ in percentage points; lower is better, and a negative value indicates improvement in the last window. Best and second-best results within each split are shown in bold and underlined, respectively.

Method	Trajectory and visual quality				Revisit consistency			Temporal
	VBench $\uparrow$	$R \downarrow$	$T \downarrow$	CMC $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	$\Delta IQ \downarrow$
<i>Simple-Trajectory Split</i>								
Infinite-World [40]	79.18	16.55	1.98	2.08	12.60	0.284	0.595	6.72
LingBot-World [27]	81.82	10.47	2.01	2.05	<u>14.59</u>	0.366	0.394	<u>0.04</u>
HY-WorldPlay [31]	68.82	17.89	2.36	2.45	12.83	0.321	0.616	23.59
Matrix-Game 3 [37]	78.53	12.96	1.83	1.92	12.29	0.326	0.553	2.41
SANA-WM [54]	79.29	<u>7.59</u>	<u>1.59</u>	<u>1.63</u>	14.16	0.333	0.458	3.79
ECHOWM	<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 [40]	79.51	41.31	2.49	2.84	12.04	0.248	0.617	4.16
LingBot-World [27]	81.89	18.99	<u>1.65</u>	1.81	14.08	<u>0.332</u>	0.436	<u>0.58</u>
HY-WorldPlay [31]	70.46	35.46	2.34	2.64	13.72	0.328	0.654	25.88
Matrix-Game 3 [37]	78.79	18.79	1.67	1.82	12.17	0.317	0.556	0.32
SANA-WM [54]	79.60	10.02	1.66	<u>1.72</u>	<u>14.10</u>	0.327	<u>0.469</u>	3.09
ECHOWM	<u>81.72</u>	<u>12.05</u>	1.265	1.358	14.58	0.333	0.482	0.93

Evaluation protocol. We evaluate SANA-WM-Bench [54] under three complementary settings: (1) a 241-frame short-horizon evaluation of the undistilled multi-step model, (2) a official 961-frame long-horizon protocol, where four overlapping 241-frame generations are sequentially chained, and (3) ECHOWM-Flash under the same 961-frame horizon using four-step causal streaming inference. The first setting provides a short-horizon reference for trajectory following and visual quality, while the latter two characterize long-horizon generation before and after causal few-step distillation under their respective inference schemes.

Following the benchmark protocol, we report rotation error (R , in degrees), relative translation error (T), camera-motion consistency error (CMC), and VBench Overall. For the 961-frame evaluations, we additionally measure viewpoint-revisit consistency using PSNR, SSIM, and LPIPS between frames that revisit approximately the same camera pose, and report temporal imaging-quality degradation, $\Delta IQ = q_1 - q_6$, where q_1 and q_6 denote the VBench Imaging Quality scores of the first and final 10-second windows, respectively.

Table 6 Long-horizon comparison of causal world models on SANA-WM-Bench under the 961-frame protocol. ECHO-WM-Flash denotes our four-step causal streaming model. Revisit consistency is measured on same-pose pairs, and ΔIQ measures temporal imaging-quality degradation. Best and second-best results within each split are shown in bold and underlined, respectively.

Method	Trajectory and visual quality				Revisit consistency			Temporal
	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>	-0.49
LingBot-World-v2 [27]	79.20	22.516	2.391	2.541	11.90	0.1715	0.5906	2.59
SANA-WM + Causal Refiner [54]	79.55	17.875	2.222	2.338	14.41	0.2660	0.5748	-0.75
SANA-WM w/o Refiner [54]	78.21	15.359	2.288	2.393	9.19	0.1661	0.6321	-0.12
ECHO-WM-Flash	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 [27]	78.74	34.416	2.125	2.413	12.34	0.2183	0.5732	3.84
SANA-WM + Causal Refiner [54]	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 [54]	78.27	<u>18.717</u>	2.006	2.142	9.26	0.1729	0.6077	2.06
ECHO-WM-Flash	81.06	14.221	1.735	1.829	14.04	0.2910	0.4893	1.30

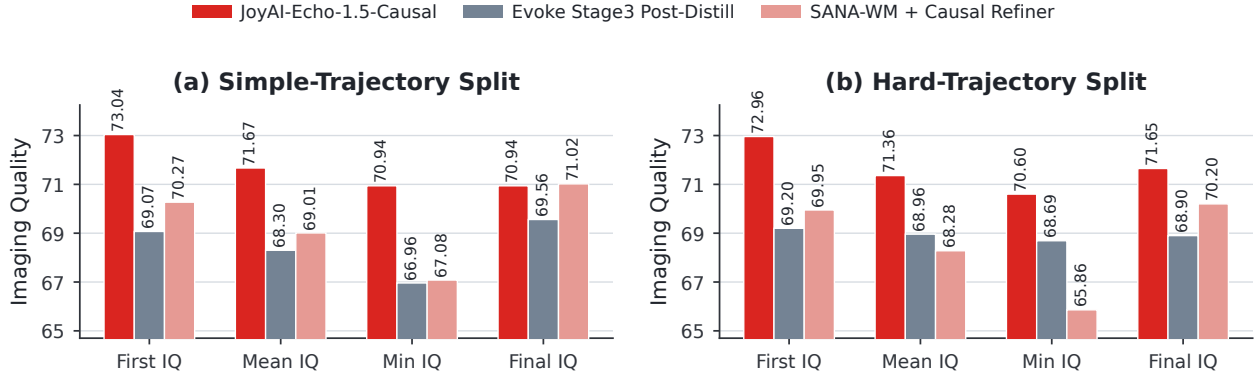


Figure 14 Imaging quality comparison on the Simple- and Hard-Trajectory splits. We report the first-window, mean, minimum, and final-window IQ. Our method achieves consistently higher mean and minimum IQ than Evoke and SANA-WM with Causal Refiner. Although the competing methods exhibit smaller IQ drops, 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.

Quantitative results. As shown in table 4, ECHO-WM achieves the highest 241-frame VBench Overall on both Simple and Hard splits, with 83.91 and 83.96, respectively. On Simple trajectories, it also achieves the lowest translation and CMC errors while remaining close to SANA-WM in rotation accuracy; on Hard trajectories, SANA-WM obtains lower pose errors, whereas ECHO-WM maintains the strongest visual quality.

Under the 961-frame long-horizon protocol (table 5), the undistilled multi-step ECHO-WM remains competitive in trajectory following while preserving strong visual quality and scene consistency. On the Simple split, it achieves a VBench Overall of 81.36, the lowest translation and CMC errors, and the highest revisit PSNR of 15.10 dB. Its rotation error increases from 0.523° at 241 frames to 3.223° at 961 frames, and further reaches 12.05° on the Hard split, indicating that accumulated pose drift remains the main challenge under difficult long-horizon trajectories.

We further compare ECHO-WM-Flash with recent causal world models under the same 961-frame horizon. As shown in table 6, ECHO-WM-Flash achieves the highest VBench Overall on both Simple and Hard splits, with 80.13 and 81.06, respectively, together with the strongest trajectory accuracy and revisit consistency among the evaluated causal methods. To further examine quality degradation over long rollouts, figure 14 summarizes the Imaging Quality (IQ) across temporal windows. ECHO-WM-Flash maintains substantially

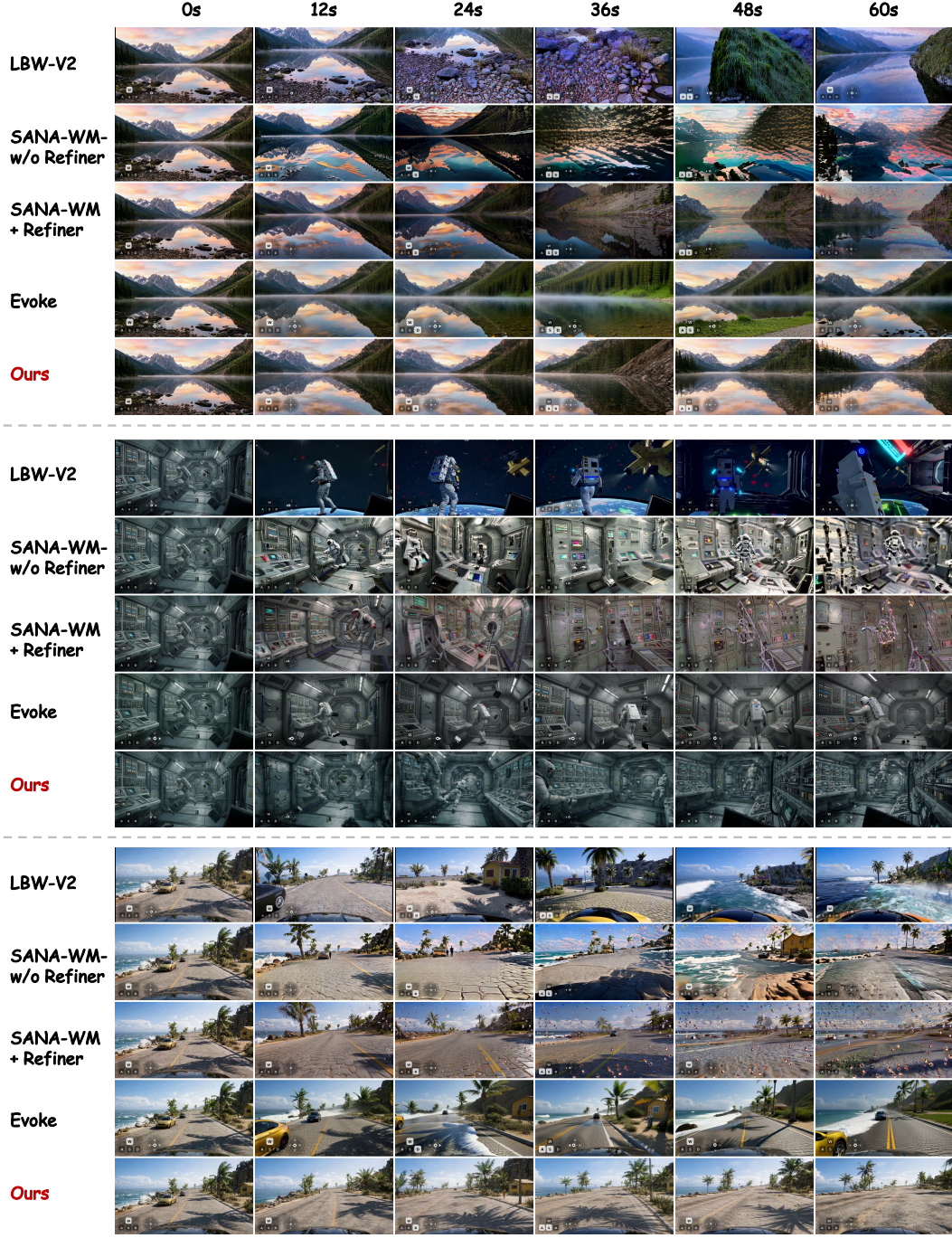


Figure 15 Qualitative comparison of long-horizon causal streaming on representative SANA-WM-Bench trajectories under the 961-frame protocol. We compare ECHO-WM-Flash with LingBot-World-v2 (LBW-V2), SANA-WM with and without Causal Refiner, and Evoke, using the same initial observation and camera-trajectory condition. Frames are shown at 12-second intervals. ECHO-WM-Flash exhibits less accumulated autoregressive drift while better preserving trajectory adherence, scene identity, visual style, and major spatial structure over the extended rollout.

higher initial, mean, and minimum IQ than competing methods while remaining competitive in the final window. Although some baselines exhibit a smaller ΔIQ , their lower absolute IQ suggests that temporal degradation should be interpreted jointly with the overall quality level. Overall, four-step causal distillation

introduces only moderate degradation relative to the undistilled model while largely preserving long-horizon visual quality and control fidelity.

Qualitative results. The long-horizon causal streaming comparisons in [figure 15](#) further highlight differences in autoregressive error accumulation. Competing causal models progressively exhibit viewpoint deviation, scene-layout drift, or changes in appearance and visual style as the streaming rollout proceeds. In contrast, ECHOWM-Flash better preserves scene and subject identity, visual style, and major geometric structures while following the prescribed camera trajectory. This sustained consistency at later timestamps is aligned with the quantitative trajectory, revisit-consistency, and temporal-quality results, demonstrating stronger robustness to long-horizon autoregressive drift.

5.3 User Study

Table 7 Aggregate pairwise user-study judgments over 200 cases. Each cell reports the percentage of judgments. “Both good” and “both bad” preserve the two distinct same-quality outcomes.

Competitor	Dimension	Competitor preferred	Both good	Both bad	ECHOWM preferred
LingBot-World-v2	Semantic following	15.14%	24.93%	25.93%	34.00%
	Initial-world consistency	7.71%	62.36%	15.29%	14.64%
	Motion	20.00%	13.71%	41.00%	25.29%
	Spatial-temporal consistency	14.71%	27.36%	25.29%	32.64%
	Visual aesthetics	19.93%	28.00%	19.57%	32.50%
	Overall preference	21.57%	4.21%	27.71%	46.50%
HappyOyster	Semantic following	29.44%	51.50%	5.44%	13.63%
	Initial-world consistency	15.44%	42.81%	2.25%	39.50%
	Motion	24.56%	11.88%	44.63%	18.94%
	Spatial-temporal consistency	10.44%	18.19%	13.69%	57.69%
	Visual aesthetics	23.81%	1.13%	22.19%	52.88%
	Overall preference	27.06%	2.19%	7.63%	63.13%

We conduct subjective evaluation on a self-built World-Model Benchmark (WMB) containing 200 cases. WMB is designed to measure overall world-model behavior under interactive generation.

Benchmark composition. The cases cover five scene groups: 100 game scenes, 50 humanoid-robot scenes, 20 natural outdoor scenes, 15 household indoor scenes, and 15 driving scenes. The viewpoint split contains 132 third-person cases and 68 first-person cases. This composition stresses both game-oriented interaction and generalization to embodied, natural, household, and driving environments.

Control protocol. WMB follows the WBench control convention [48]. For first-person cases, WASD translates the camera and the arrow keys rotate the viewpoint. For third-person cases, WASD moves the visible subject and the arrow keys orbit the camera around that subject. The same command semantics are used for every model under comparison, while the generated video is evaluated from the resulting reference, text, and control condition.

Subjective dimensions. Annotators score five dimensions: (1) semantic following, measuring adherence to the requested scene and behavior; (2) initial-world consistency, measuring preservation of the reference observation and initial world state; (3) spatial-temporal consistency, measuring geometry, identity, and layout stability over time; (4) motion quality, measuring smoothness, responsiveness, and physical plausibility; and (5) visual aesthetics, measuring composition, lighting, texture, and overall perceptual quality.

Table 7 shows that ECHOWM receives the largest share of overall-preference votes against both competitors: 46.50% against LingBot-World-v2 and 63.13% against HappyOyster. Against LingBot-World-v2, ECHOWM also leads semantic following (34.00%), spatial-temporal consistency (32.64%), and visual aesthetics (32.50%), whereas initial-world consistency is dominated by both-good judgments (62.36%) and motion most often receives both-bad judgments (41.00%). Against HappyOyster, the clearest advantages are spatial-temporal consistency (57.69%) and visual aesthetics (52.88%). Semantic following remains a weakness in this comparison:

Table 8 Scene-level user-study results against LingBot-World-v2 and HappyOyster. All values are percentages. The LingBot-World-v2 comparison uses seven annotators, while the HappyOyster comparison uses eight annotators.

Scene	Dimension	vs. LingBot-World-v2				vs. HappyOyster			
		LBW-v2	Both good	Both bad	ECHOWM	HappyOyster	Both good	Both bad	ECHOWM
Autonomous driving	Semantic following	30.48	29.52	17.14	22.86	18.33	44.17	13.33	24.17
	Initial-world consistency	3.81	66.67	15.24	14.29	10.83	50.00	0.00	39.17
	Motion	21.90	6.67	42.86	28.57	5.83	10.83	35.83	47.50
	Spatial-temporal consistency	10.48	31.43	29.52	28.57	10.00	24.17	12.50	53.33
	Visual aesthetics	11.43	34.29	21.90	32.38	19.17	3.33	17.50	60.00
	Overall preference	24.76	1.90	30.48	42.86	14.17	3.33	7.50	75.00
Embodied robot	Semantic following	12.90	9.45	42.17	35.48	42.94	41.53	6.85	8.67
	Initial-world consistency	7.37	60.60	21.66	10.37	21.37	42.14	2.02	34.48
	Motion	21.20	7.83	53.46	17.51	31.45	7.46	52.62	8.47
	Spatial-temporal consistency	15.90	25.58	34.10	24.42	10.89	17.74	18.15	53.23
	Visual aesthetics	22.81	28.34	24.42	24.42	30.44	1.01	29.64	38.91
	Overall preference	22.35	3.00	39.63	35.02	40.93	1.81	11.69	45.56
Multi-social	Semantic following	21.43	28.57	14.29	35.71	12.50	62.50	6.25	18.75
	Initial-world consistency	21.43	64.29	7.14	7.14	6.25	62.50	0.00	31.25
	Motion	14.29	42.86	14.29	28.57	0.00	31.25	43.75	25.00
	Spatial-temporal consistency	21.43	42.86	0.00	35.71	0.00	12.50	6.25	81.25
	Visual aesthetics	28.57	28.57	7.14	35.71	18.75	0.00	12.50	68.75
	Overall preference	28.57	7.14	7.14	57.14	12.50	0.00	6.25	81.25
Natural outdoor	Semantic following	17.86	16.43	37.86	27.86	46.88	43.75	6.88	2.50
	Initial-world consistency	6.43	67.14	13.57	12.86	21.88	34.38	1.25	42.50
	Motion	15.71	7.86	51.43	25.00	38.75	10.00	41.25	10.00
	Spatial-temporal consistency	15.71	21.43	26.43	36.43	13.13	14.37	8.13	64.38
	Visual aesthetics	14.29	28.57	22.86	34.29	28.75	0.63	21.88	48.75
	Overall preference	17.86	4.29	31.43	46.43	41.25	1.25	8.75	48.75

HappyOyster is preferred in 29.44% of judgments versus 13.63% for ECHOWM, with 51.50% of pairs judged both good.

The reported scene subset contains 99 cases across autonomous driving, embodied robot, multi-social, and natural outdoor settings. For overall preference, ECHOWM leads LingBot-World-v2 in all four groups, with vote shares ranging from 35.02% in embodied robot to 57.14% in multi-social scenes. Against HappyOyster, ECHOWM is strongly preferred in autonomous driving (75.00%) and multi-social scenes (81.25%), while the margins are narrower in embodied robot (45.56% versus 40.93%) and natural outdoor scenes (48.75% versus 41.25%). The dimension-level breakdown also exposes localized failure modes: against HappyOyster, ECHOWM receives only 8.67% of semantic-following votes in embodied-robot scenes and 2.50% in natural-outdoor scenes, despite leading those scenes in overall preference.

5.4 Qualitative Analysis

Quantitative benchmarks measure performance under predefined protocols, but do not fully reveal the breadth of an enterable omnimodal world. We therefore organize the qualitative analysis along five complementary capability axes: first-person generalization across worlds, third-person interaction across visible subjects, spatial structure recoverable from generated views, native synchronized generation of video and audio, and multi-turn preservation of audio-visual context. Together, these axes examine whether the same camera-intent interface remains effective as the world, viewpoint, subject, spatial evidence, output modality, and interaction horizon change.

First-Person World Generalization. We first examine whether one camera-intent interface generalizes across substantially different world distributions. Figure 16 covers game-rendered worlds, real-world scenes, fantastic environments, and stylized content under the same first-person trajectory representation. Each row shows temporally ordered output frames, allowing changes in world appearance and layout to be examined without changing the interaction interface. The examples demonstrate observer-centric navigation across diverse world distributions without a domain-specific controller.

Third-Person Subject and Camera Control. We next examine third-person interaction across different visible subject configurations. Figure 18 groups rollouts containing humans, animals, other foreground subjects, and humans together with additional subjects. Unlike an explicit actor trajectory, the condition



Figure 16 First-person world generalization. Temporally ordered rollouts cover game worlds, real-world scenes, fantastic worlds, and stylized worlds. Across varied layouts and rendering styles, ECHOWM uses the same camera-intent interface for observer-centric navigation.

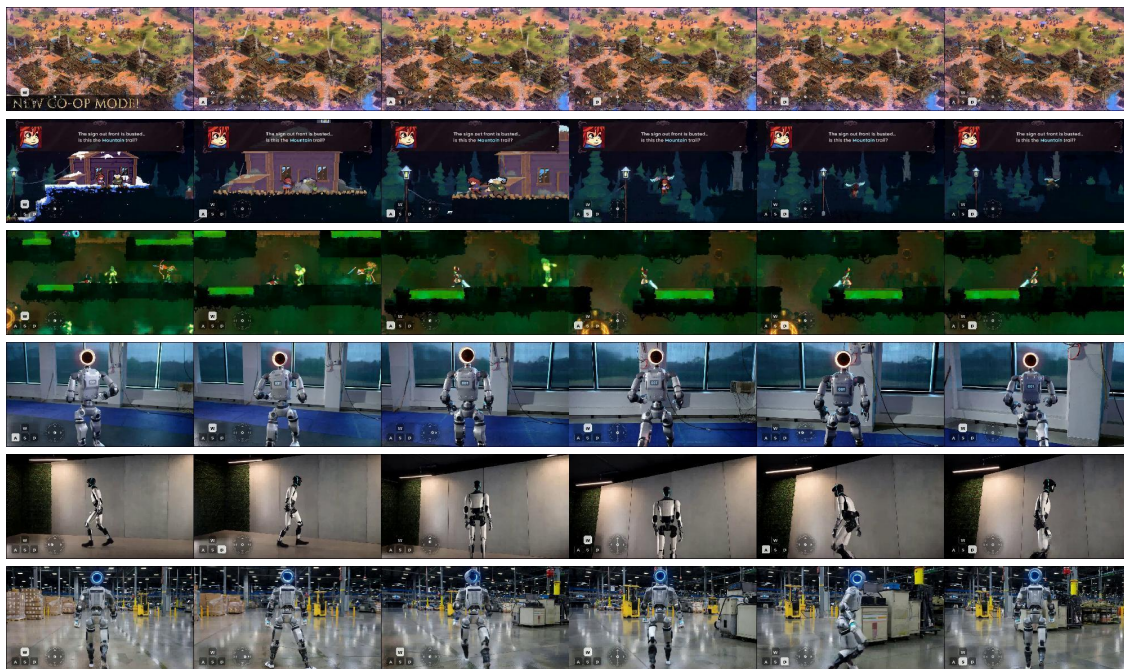


Figure 17 Generalization beyond the training domains. Our training mixture contains no 2D or 2.5D game data and no robot data. The top three rows show 2D/2.5D game scenes and the bottom three rows show robot environments; columns are temporally ordered frames under navigation controls. ECHOWM preserves coherent viewpoint evolution and scene or subject structure across these unseen domains.

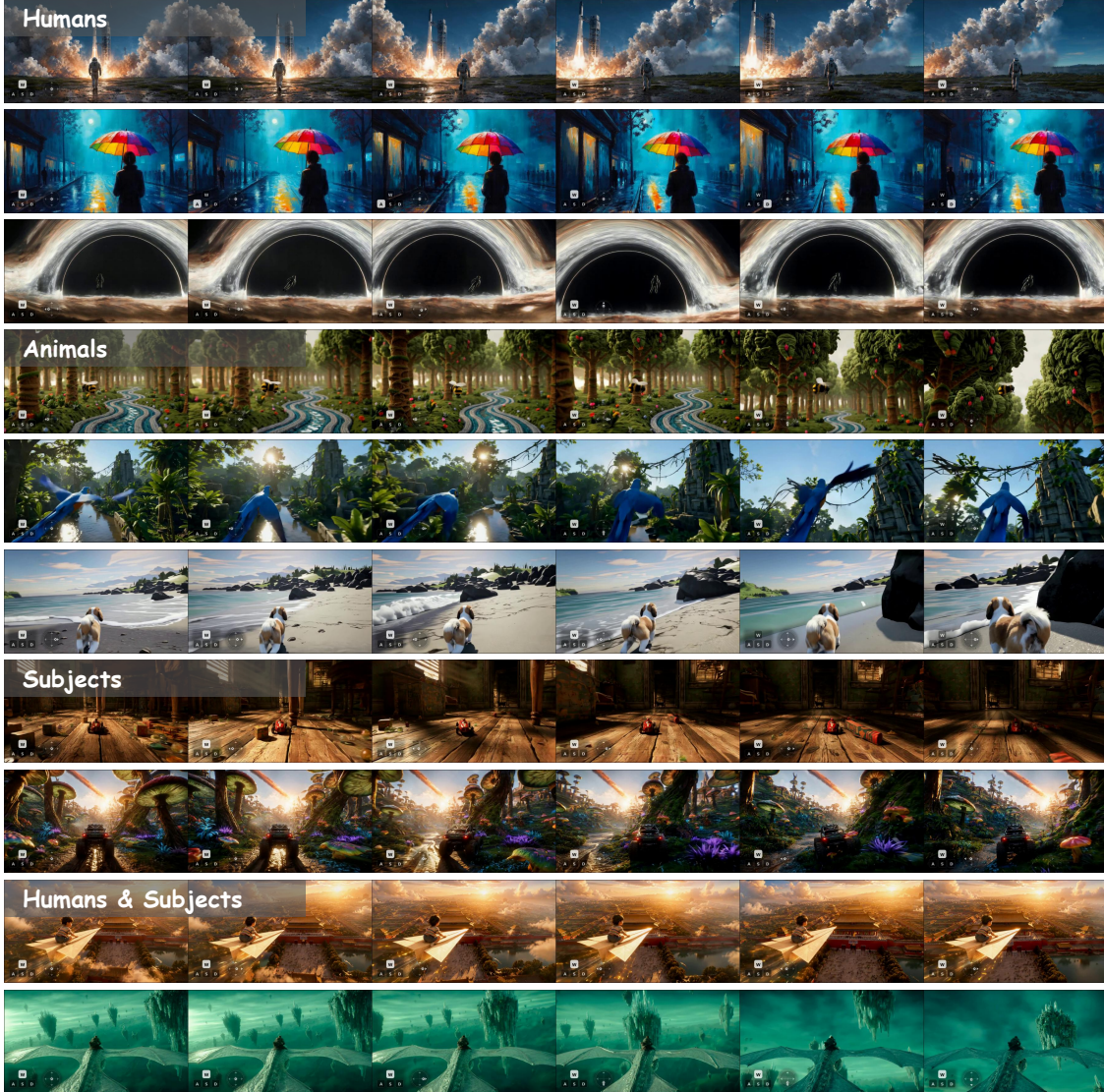


Figure 18 Third-person generalization across visible subjects. Representative rollouts span humans, animals, other foreground subjects, and scenes containing both humans and additional subjects. The shared camera-intent interface supports learned camera-character evolution across these subject configurations without a subject-specific controller.

specifies the desired evolution of observation; the learned world prior determines the associated subject motion and camera-character relationship. This analysis therefore focuses on whether one shared interface produces coherent camera-character evolution across subject types without viewpoint- or embodiment-specific controllers.

Spatial Consistency through 3D Reconstruction. Frame-wise visual quality alone does not establish that generated views describe a coherent 3D scene. We therefore visualize 3D reconstructions obtained from generated sequences alongside the temporally ordered frames used for reconstruction in Figure 19. Across diverse indoor and outdoor settings, the reconstructed views retain recognizable scene-level layouts and major surfaces as the viewpoint changes. Because the result also depends on the downstream reconstruction procedure, we treat it as qualitative supporting evidence of multi-view spatial coherence rather than a claim of metric geometric accuracy.

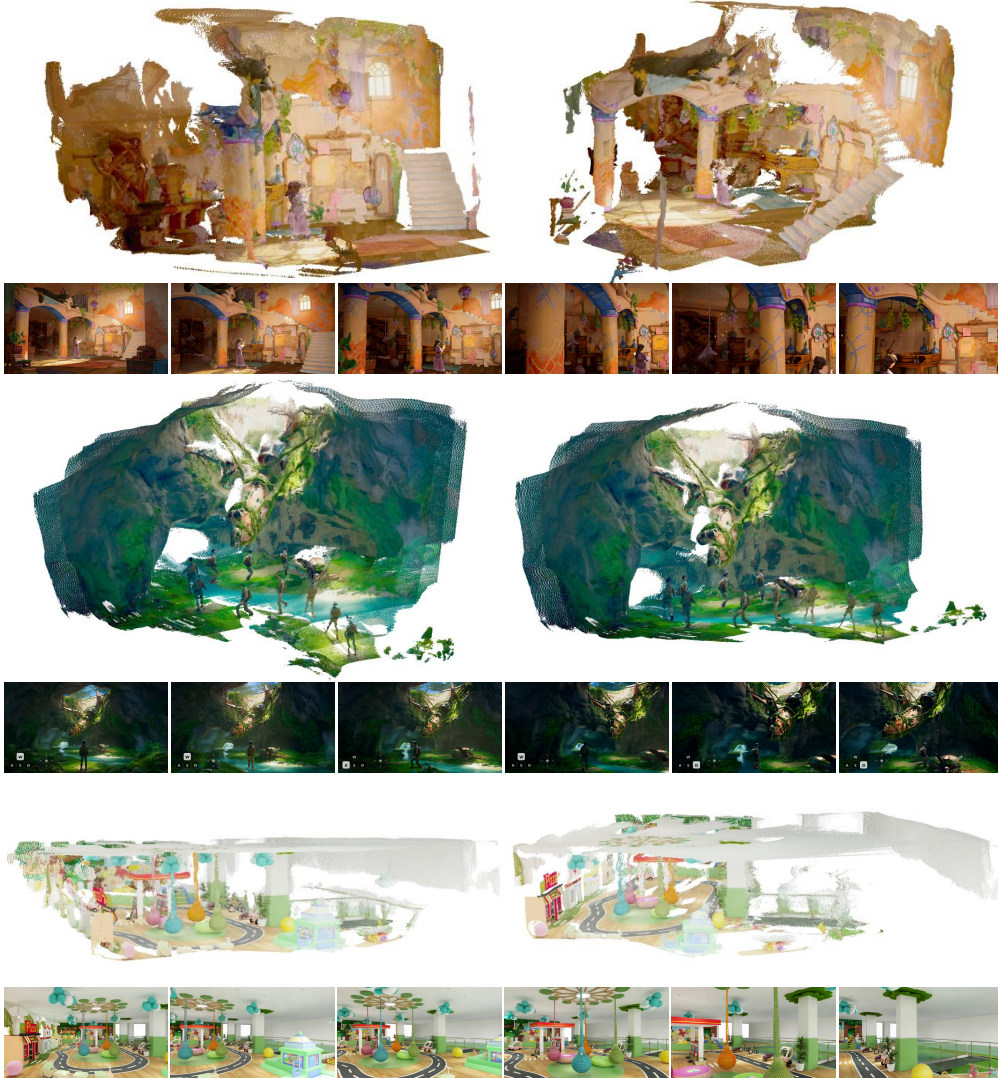


Figure 19 Qualitative 3D reconstruction from generated trajectories. Each example shows two views of the reconstructed scene together with six temporally ordered frames from the corresponding generated sequence. The recognizable scene-level layouts across viewpoint changes provide qualitative evidence of multi-view spatial coherence.

Generalization beyond the training domains. Our training mixture contains no 2D or 2.5D game data and no robot data. Figure 17 evaluates the same camera-intent interface on these unseen domains: the top rows show 2D/2.5D game scenes, while the bottom rows show robot environments. Across the displayed rollouts, ECHOWM maintains coherent viewpoint evolution and preserves major scene or subject structure under navigation controls. This provides qualitative evidence of cross-domain generalization rather than evidence of domain-specific training coverage.

Native Synchronized Audio-Visual Generation. We then examine the model’s native joint generation of video and audio across environmental sound, background music, and character speech. Figure 20 pairs each structured world and audio description with time-stamped output frames, the generated waveform, and a log-frequency spectrogram. The four examples include scene ambience and foreground effects, musical accompaniment, and spoken dialogue within the same generated sequences. These examples assess synchronization between generated sound and visible content; they do not posit a separately supervised action-to-audio controller.



Figure 20 Native synchronized audio-visual generation. Each example pairs its world and audio description with time-stamped video frames, the generated waveform, and a log-frequency spectrogram. The cases contain environmental sound, background music, and character speech within the same generated audio-visual sequence, rather than adding audio as post-processing.

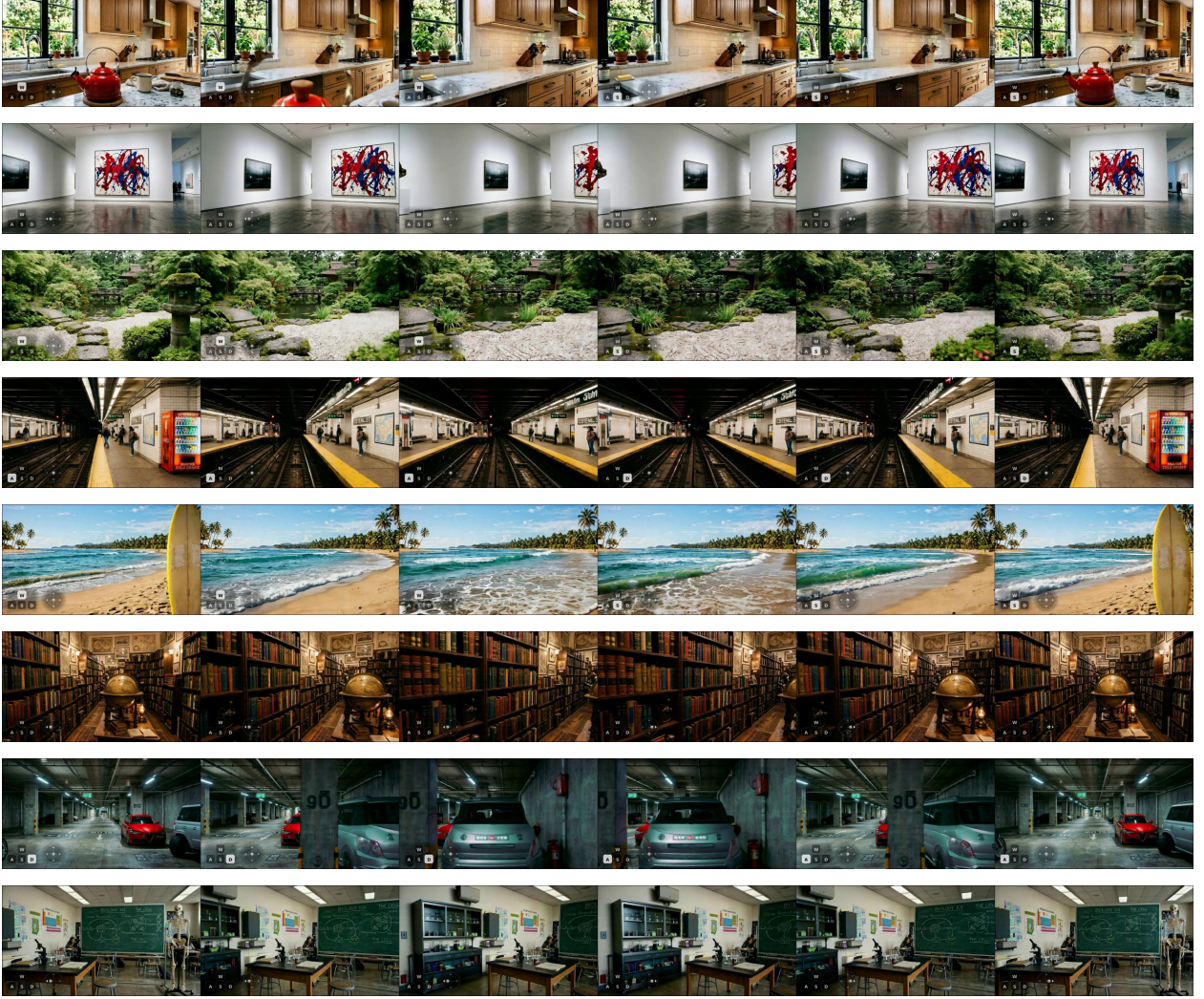


Figure 21 Visual memory during long-horizon continuation. Each row samples a temporally ordered trajectory that moves through and revisits a scene. Returning to previously observed views tests whether layout, appearance, and scene identity remain coherent over continued generation.

Multi-Turn Audio-Visual Continuation. Finally, we examine whether continued generation preserves previously observed visual context. Figure 21 shows long-horizon trajectories that move through and revisit eight different scenes. Each row samples the trajectory in temporal order, so the return to earlier viewpoints reveals whether layout, appearance, and scene identity remain coherent after intermediate motion. This figure focuses on the visual-memory component of multi-turn continuation.

Visual Memory. We assess whether the model preserves previously observed world states during long-horizon continuation. Figure 21 shows trajectories that leave and later revisit earlier views. The returned views retain major scene layout, appearance, and identity cues after intermediate motion, providing qualitative evidence that the model can reuse prior visual context across multiple turns.

5.5 Ablation Studies

We organize the ablations around trajectory scale, camera-intent semantics, inference-time speed control, and the audio-visual pretraining task mixture introduced in section 4.

5.5.1 Trajectory Representation and Scale Calibration

The trajectory pathway must be both numerically stable during training and sensitive to displacement magnitude at inference. We evaluate these properties separately because WBench adapts trajectory scale for robust open-domain scoring, which can hide errors in absolute motion magnitude.

Qualitative scale control. We compare matched navigation rollouts to examine whether the fixed global translation scale preserves the requested displacement amplitude as the camera approaches a foreground character or salient scene structure.

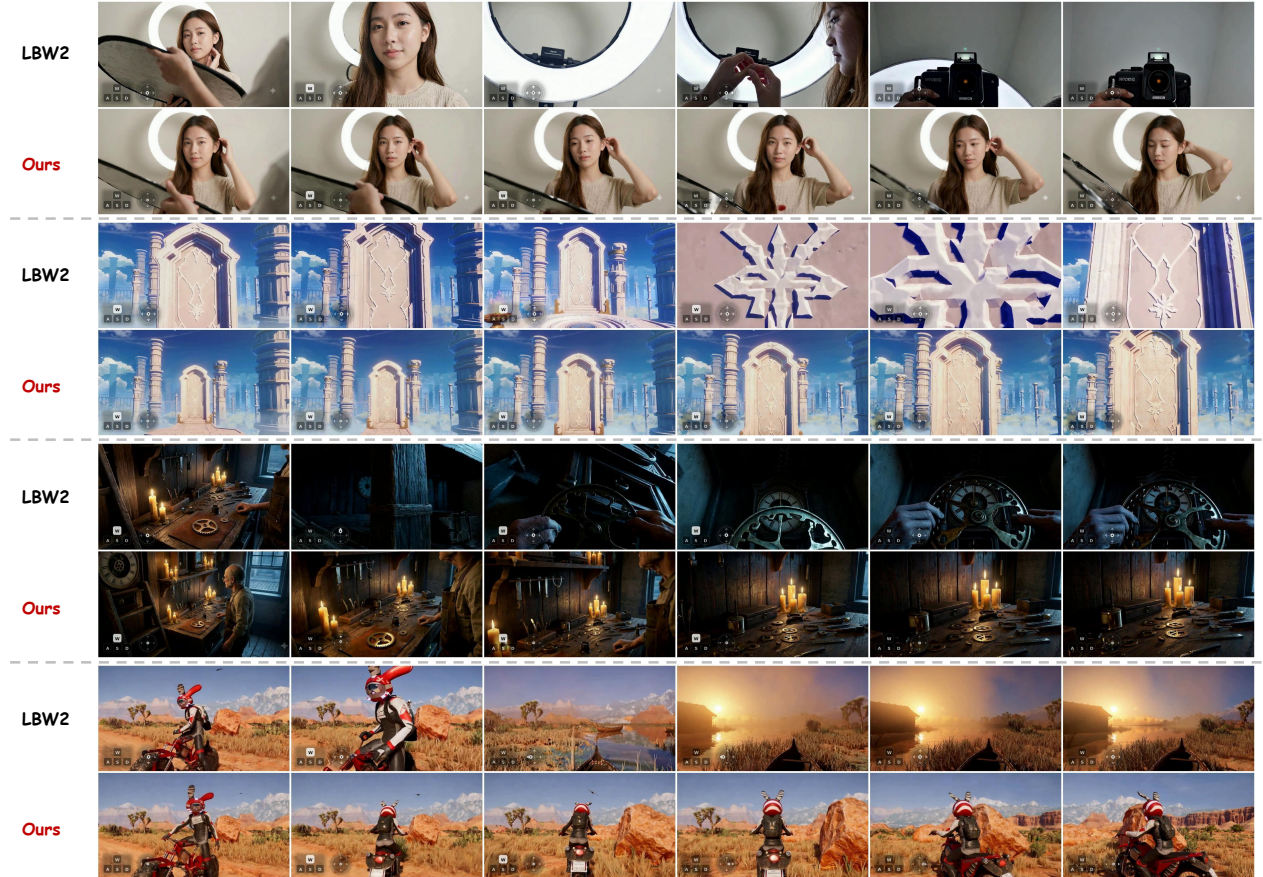


Figure 22 Qualitative scale-control comparison. LingBot-World-v2 [27] and ECHO WM receive matched reference observations and navigation conditions. When the camera approaches a foreground character or scene structure, LingBot-World-v2 exhibits larger changes in apparent subject scale and less predictable displacement response, whereas ECHO WM preserves a more gradual scale change across the trajectory. This qualitative comparison complements the metric-scale and speed-response ablations: a fixed global translation scale keeps requested displacement amplitudes comparable across clips and makes approach behavior more controllable.

Figure 22 provides a qualitative complement to this metric-scale ablation. Under matched navigation conditions, the LingBot-World-v2 baseline can change apparent subject scale abruptly when the camera approaches a character or a salient scene structure, making the final distance and approach rate difficult to control. In contrast, the fixed global translation scale used by ECHO WM preserves the relative amplitude of the requested displacement across examples, yielding a more gradual and predictable scale response. The figure is qualitative evidence for this scale-control behavior; the inference-time speed analysis below is retained for the corresponding quantitative evaluation.

Inference-time speed control. We construct trajectory groups that share the same reference observation, direction, and rotation but vary translation magnitude. For each group, we measure the realized displacement

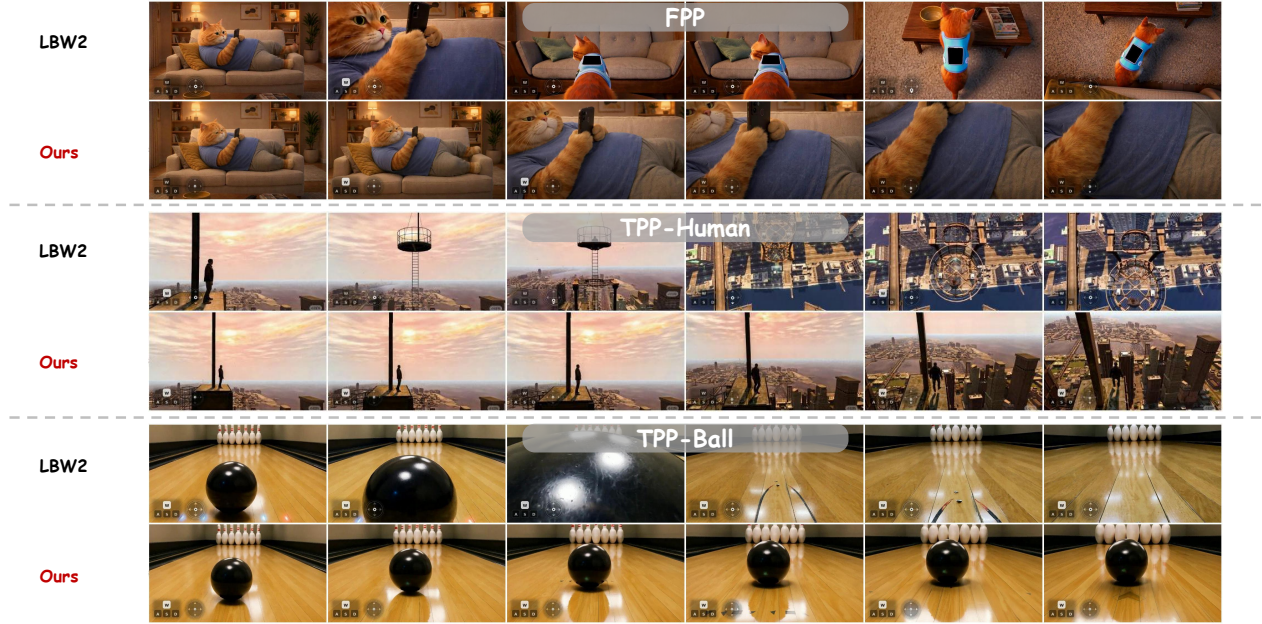


Figure 23 Qualitative viewpoint semantics and third-person tracking. LingBot-World-v2 (LBW2) [27] and ECHOWM receive matched reference observations and navigation conditions in first-person (FPP), third-person human (TPP-Human), and third-person object (TPP-Ball) cases. In the FPP example, LBW2 changes to an exocentric observation as the rollout proceeds, whereas ECHOWM retains the requested first-person semantics. In both TPP examples, LBW2 loses the controlled subject from view, while ECHOWM keeps the human or ball visible and maintains a more stable camera–subject relationship across the sequence.

in the generated video and report its correlation with the requested displacement, response slope, and absolute scale error. The corresponding response curves test the central distinction between per-clip scaling, which removes cross-clip magnitude, and the fixed global scale in section 4.2.2, which retains trajectory amplitude as a user control. For multi-chunk rollouts, we also report the change in realized speed across chunk boundaries to test whether a normalization strategy introduces artificial transitions.

5.5.2 Camera-Intent Modeling across Viewpoints

We qualitatively examine whether one camera-intent interface supports direct observer motion in first person and coordinated camera–character evolution in third person without viewpoint-specific controllers.

Figure 23 provides a qualitative comparison with LingBot-World-v2 under matched first- and third-person controls. In the first-person case, ECHOWM preserves the requested egocentric observation and avoids the transition to an exocentric view observed in the baseline rollout. In the two third-person cases, ECHOWM also keeps the human or ball visible throughout the sequence and maintains a more stable camera–subject relationship, whereas the baseline loses the subject as the camera evolves. These examples provide qualitative evidence for the role of explicit viewpoint semantics and a shared camera-intent condition.

6 Discussion

Echo-World connects two directions that have largely developed separately: generative audio-visual media and interactive world models. Its design suggests five principles that extend beyond the particular backbone.

Center interaction on observation intent. Rather than defining a separate control semantics for each viewpoint, we describe how the user intends to observe and traverse the world. The same camera intent can be realized as observer motion in first-person scenes or as coordinated camera and character evolution in third-person scenes. This lets the model learn plausible camera–character coupling from data, while cinematic

video and gameplay contribute complementary evidence about observation, composition, and world response.

Represent interaction at the world level. Input devices and benchmarks express navigation through different action vocabularies, but their effect can often be represented as a common camera-space change. Converting discrete and continuous inputs to a shared relative trajectory avoids binding the generator to a particular input layout or action encoder. This observation applies to navigation and viewpoint-related actions; it does not imply that arbitrary actor behavior can be reduced to camera motion.

Make scale and viewpoint semantics explicit. Per-example normalization can make optimization easier while removing the relative displacement that an interaction condition is intended to represent. A robust dataset-level scale preserves motion relationships across sources and keeps adjacent rollout chunks on the same velocity scale, avoiding normalization-induced speed jumps. At the same time, geometry alone does not specify whether a rotation is egocentric, character-following, or orbital. Conditioning on viewpoint separates camera intent from its scene-dependent realization.

Keep audio inside the generated world. In passive media, audio enriches presentation; in an enterable world, it also provides ambient, off-screen, and social context while the visual rollout is being navigated. Our trajectory condition enters only the video stream, so we do not impose an explicit action-to-sound controller. Instead, environmental sound and speech remain part of the native joint audio-visual process. Joint-FT optimizes this process on trajectory-conditioned examples, but the trajectory still reaches audio only through the model’s existing cross-modal coupling. We therefore evaluate acoustic coherence under interaction without claiming a separately supervised action-to-sound capability.

Match each training stage to its cleanest signal. Audio-visual generation and interaction rely on different data distributions and objectives. AV-CPT uses audio-rich and speech-rich clips to establish the world prior, while Action-SFT freezes that prior and uses control-clean motion to isolate navigation acquisition. A smaller balanced set then supports low-learning-rate Joint-FT of the full model. The curriculum thus separates capability acquisition before allowing the two pathways to co-adapt.

Together, these principles explain why Echo-World is more than a controlled video generator. The data engine supplies complementary world experience, the shared interaction condition makes it enterable, the joint backbone makes it perceptually responsive, and continuation turns individual generations into an evolving audio-visual experience.

7 Limitations

Echo-World currently focuses on navigation and viewpoint-related actions that can be represented by a relative 6-DoF trajectory. Discrete keyboard states are supported through this mapping, but the model does not explicitly represent arbitrary actor intent such as jumping, attacking, manipulation, or robot commands. It also does not implement the rules, collision state, or deterministic transitions of a conventional game engine.

The current model lacks an explicit persistent 3D memory. Geometry, subject identity, world state, and audio can therefore drift over repeated continuation turns. Its interaction range is also restricted to trajectories retained by the global filter; generalization to substantially larger translations, higher speeds, or unusual rotations is not established. Pose estimates from Internet and gameplay video remain imperfect even after filtering and may correlate with scene content or motion blur.

8 Conclusion

We presented ECHOWM, an omnimodal world model for enterable generative media that bridges generative media and interactive world modeling across general and game scenes. Echo-World organizes interaction around camera intent, allowing first-person observer motion and third-person camera-character evolution to be learned as different realizations of how a world is observed. It combines this perspective with a complementary gameplay-simulation-Internet data engine, a scale-consistent trajectory condition, and a progressive audio-visual-to-action curriculum. Continued audio-visual pretraining first learns from AV-rich data; action fine-tuning freezes this world prior and learns visual navigation from control-clean motion;

low-learning-rate joint fine-tuning then consolidates both on a smaller, jointly high-quality set. Sound remains part of the jointly generated world and provides perceptual context during interaction. The resulting model supports first- and third-person interaction, native joint audio-visual output, and multi-turn continuation in one system. More broadly, our results suggest a path from generative content that is only watched toward omnimodal worlds that can be continuously entered and influenced.

References

- [1] Jake Bruce, Michael D. Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *ICML*, 2024.
- [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] Haoxin Cheng et al. MMAudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *CVPR*, 2025.
- [4] DreamX Team, Yancheng Bai, Rui Chen, Xiangxiang Chu, Rujing 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.
- [5] Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, et al. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025.
- [6] 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/>.
- [7] Google DeepMind. Veo 3, 2025. URL <https://deepmind.google/technologies/veo/veo-3/>.
- [8] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.
- [9] 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.
- [10] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *ICML*, 2019.
- [11] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *ICLR*, 2020.
- [12] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. CameraCtrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024.
- [13] Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, et al. Matrix-Game 2.0: An open-source, real-time, and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025.
- [14] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. GAIA-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.
- [15] 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.
- [16] 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.
- [17] 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.
- [18] Jiaqi Li, Junshu Tang, Zhiyong Xu, Longhuang Wu, Yuan Zhou, Shuai Shao, Tianbao Yu, Zhiguo Cao, and Qinglin Lu. Hunyuan-GameCraft: High-dynamic interactive game video generation with hybrid history condition. *arXiv preprint arXiv:2506.17201*, 2025.
- [19] Haohe Liu, Gael Le Lan, Xinhao Mei, Zhaoheng Ni, Anurag Kumar, Varun Nagaraja, Wenwu Wang, Mark D. Plumbley, Yangyang Shi, and Vikas Chandra. SyncFlow: Toward temporally aligned joint audio-video generation from text. *arXiv preprint arXiv:2412.15220*, 2024.

- [20] Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-Foley: Synchronized video-to-audio synthesis with latent diffusion models. In NeurIPS, 2023.
- [21] Xiaofeng Mao, Shaoheng Lin, Zhen Li, Chuanhao Li, Wenshuo Peng, Tong He, Jiangmiao Pang, Mingmin Chi, Yu Qiao, and Kaipeng Zhang. YUME: An interactive world generation model. arXiv preprint arXiv:2507.17744, 2025.
- [22] NVIDIA. Cosmos world foundation model platform for physical AI. arXiv preprint arXiv:2501.03575, 2025.
- [23] NVIDIA. Cosmos 3: Omnimodal world models for physical AI. Technical report, NVIDIA, 2026.
- [24] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie Gen: A cast of media foundation models. arXiv preprint arXiv:2410.13720, 2024.
- [25] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. GEN3C: 3D-informed world-consistent video generation with precise camera control. In CVPR, 2025.
- [26] Riemann Dynamics. Matrix-Game 3.5: Enhancing real-time streaming interactive world models with patch memory. Technical report, Riemann Dynamics, 2026. URL <https://matrix-game-v3-5.github.io/>.
- [27] 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.
- [28] Georgy Savva, Oscar Michel, Daohan Lu, Suppakit Waiwitlikhit, Timothy Meehan, Dhairya Mishra, Srivats Poddar, Jack Lu, and Saining Xie. Solaris: Building a multiplayer video world model in Minecraft. arXiv preprint arXiv:2602.22208, 2026.
- [29] 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.
- [30] 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.
- [31] 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.
- [32] Wan Team, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Fei Wu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314, 2025.
- [33] Wan Team, Alibaba Group. Video = world + event stream. arXiv preprint arXiv:2607.15038, 2026.
- [34] Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT- ω . arXiv preprint arXiv:2605.15195, 2026.
- [35] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. DriveDreamer: Towards real-world-driven world models for autonomous driving. In ECCV, 2024.
- [36] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. MotionCtrl: A unified and flexible motion controller for video generation. In ACM SIGGRAPH, 2024.
- [37] Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixin 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.
- [38] Ronald J Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. Neural computation, 1(2):270–280, 1989.
- [39] 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.

- [40] 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.
- [41] Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. WorldMem: Long-term consistent world simulation with memory. In [Advances in Neural Information Processing Systems](#), 2025.
- [42] Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. CamCo: Camera-controllable 3D-consistent image-to-video generation. [arXiv preprint arXiv:2406.02509](#), 2024.
- [43] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-Omni technical report. [arXiv preprint arXiv:2509.17765](#), 2025.
- [44] Sherry Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Leslie Kaelbling, Dale Schuurmans, and Pieter Abbeel. UniSim: Learning interactive real-world simulators. [arXiv preprint arXiv:2310.06114](#), 2023.
- [45] Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, et al. Longlive: Real-time interactive long video generation. [arXiv preprint arXiv:2509.22622](#), 2025.
- [46] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In [Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition](#), 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] Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval. In [SIGGRAPH Asia 2025 Conference Papers](#), 2025.
- [50] 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.
- [51] Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. In [Advances in Neural Information Processing Systems](#), 2025.
- [52] Yiming Zhang, Yicheng Gu, Yanhong Zeng, Zhening Xing, Yuancheng Wang, Zhizheng Wu, and Kai Chen. FoleyCrafter: Bring silent videos to life with lifelike and synchronized sounds. [IJCV](#), 2026.
- [53] Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. RoboDreamer: Learning compositional world models for robot imagination. [arXiv preprint arXiv:2404.12377](#), 2024.
- [54] 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.
- [55] 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.