HELIXWORLD: A REAL-TIME INTERACTIVE AUDIO-VISUAL WORLD MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

World simulation is inherently multisensory, demanding synchronized visual and acoustic dynamics in real time. Yet prevailing interactive world models remain strictly silent, focusing exclusively on visual rendering and control while overlooking the acoustic dimension. We present HelixWorld, a real-time interactive audio-visual world model where visual scenes and camera-grounded spatial stereo sound co-evolve natively under user interaction. We curate a high-fidelity spatial audio-visual dataset with true stereo acoustics and metric camera poses, upon which we pre-train a bidirectional teacher conditioned on 6-DoF camera trajectories and user actions. To enable low-latency causal interaction, we distill the teacher into a few-step streaming student via an online trajectory distillation loss, sustaining drift-free joint audio-visual rollouts at 24 FPS on a single GPU. Furthermore, we formalize spatial-acoustic consistency and introduce HelixBench to evaluate whether synthesized sound fields faithfully track dynamic viewpoint motion. Extensive experiments demonstrate that HelixWorld matches state-of-the-art silent world models in visual fidelity and responsiveness, while significantly surpassing existing baselines in camera-aligned spatial-acoustic immersion.

1 INTRODUCTION

Interactive world models (Ha & Schmidhuber, 2018; Bruce et al., 2024; Valevski et al., 2025; Decart et al., 2024; Che et al., 2025; Zhang et al., 2025; Li et al., 2025a; Sun et al., 2025) provide an effective paradigm for physical simulation and virtual environments, allowing users to control generated rollouts in real time via keyboard actions or camera trajectories. Beyond visual synthesis, physical interactions naturally produce sound that conveys critical cues about surface materials, contact forces, and off-screen events. To provide immersive simulation, an interactive world model should synthesize visual scenes and spatially grounded stereo audio responsive to user inputs. However, prevailing world models focus almost exclusively on visual streams, leaving the acoustic dimension largely unexplored. Despite recent efforts to incorporate audio into world models (Zhang et al., 2026a), achieving fine-grained spatial-acoustic alignment under dynamic camera movement remains an open challenge due to the absence of high-quality spatial audio paired with camera trajectories, severe error accumulation during long-horizon streaming, and instability of causal model training.

Natively coupling interactive visual scenes with spatial audio introduces fundamental challenges that straightforward adaptations of existing models fail to resolve. Cascaded video-to-audio models (Cheng et al., 2025; Tian et al., 2026) lack access to camera trajectories and user actions, leaving sound oblivious to ego-motion, while their sequential latency precludes real-time responsiveness. Meanwhile, bidirectional audio-visual foundation models (HaCohen et al., 2026; Google DeepMind, 2025) lack external control interfaces. Moreover, naively converting them into causal autoregressive rollouts leads to catastrophic multimodal drift, where historical error accumulation rapidly degrades both visual stability and acoustic fidelity over extended sequences.

To address these challenges, we present **HelixWorld**, a real-time interactive audio-visual world model with camera-aligned spatial sound, trained in two stages (see Sec. 4 for details). First, *bidirectional teacher training* establishes action controllability and spatial-acoustic fidelity. We inject continuous 6-DoF camera trajectories into visual attention via PRoPE (Li et al., 2025b) and discrete user actions into diffusion timesteps via AdaLN-Zero (Peebles & Xie, 2023). Cross-modal attention subsequently binds visual geometry with acoustic features, naturally transferring camera ego-motion

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

into realistic stereo panning. Second, *causal distillation* converts this bidirectional teacher into a few-step causal student. To mitigate compounding drift under self-forcing (Huang et al., 2025b), we introduce an *online trajectory distillation loss* that supervises student velocities with teacher Probability Flow ODE targets directly from intermediate rollout states. Combined with long-horizon streaming fine-tuning, HelixWorld substantially reduces exposure bias and sustains synchronized audio-visual generation at 24 FPS on a single NVIDIA H800 GPU.

Beyond model architecture, the successful training of such an interactive system requires large-scale data coupling visual dynamics with high-quality spatial acoustics. However, existing world model datasets are largely silent; even when audio is present in existing video corpora, it is heavily contaminated by post-production background music, voiceover narration, or pseudo-stereo tracks that provide no authentic spatial cues. To bridge this training data gap, we construct a scalable, coarse-to-fine curation pipeline structured by computational cost (Sec. 3). The pipeline deploys inter-channel energy probes to eliminate pseudo-stereo audio at ingestion, utilizes multimodal LLMs to purge non-world footage and non-diegetic audio, and automatically recovers frame-accurate metric camera poses paired with decoupled three-track captions (visual scene, foreground acoustic events, and background ambiance), establishing a high-fidelity spatial audio-visual training corpus.

Finally, evaluating an interactive audio-visual world model requires assessing whether synthesized audio physically aligns with dynamic camera motion. Existing world model benchmarks (Ying et al., 2026) focus exclusively on visual dynamics while neglecting the acoustic stream, whereas conventional audio-visual benchmarks (Chen et al., 2020) evaluate only short, static clips without camera movement. To close this gap, we introduce **HelixBench**, the first benchmark dedicated to evaluating interactive audio-visual world models (Sec. 5). Spanning $1{,}015$ human-verified cases across diverse environments and interaction modes, HelixBench evaluates models across five key dimensions: temporal synchronization, audio-visual semantics, text-audio alignment, acoustic dynamics, and *spatial-acoustic consistency*. In particular, it formalizes spatial-acoustic consistency by correlating on-screen sound emitter positions with stereo audio panning, directly quantifying whether the synthesized sound field physically rotates with camera ego-motion.

In summary, our main contributions are:

- **HelixWorld Model:** We propose an interactive audio-visual world model that natively unifies continuous 6-DoF camera trajectories and discrete actions, achieving synchronized and camera-responsive rollouts in real time.
- **Causal Distillation Recipe:** We develop a principled distillation framework coupling causal initialization, online trajectory distillation, and streaming fine-tuning, which eliminates compounding rollout drift and sustains real-time 24 FPS streaming.
- **Multimodal Data Pipeline:** A systematic multi-stage curation workflow that purges pseudo-stereo and non-diegetic tracks while annotating metric camera extrinsics and decoupled three-track captions, establishing a high-quality spatial corpus.
- **HelixBench:** A benchmark of $1{,}015$ human-verified interactive cases formalizing spatial-acoustic consistency alongside temporal synchrony and multimodal semantics for interactive audio-visual evaluation.

2 Related Work

Interactive world models. Interactive world models (Ha & Schmidhuber, 2018) simulate future observations under diverse controls, including latent codes (Bruce et al., 2024; Parker-Holder et al., 2024), keyboard actions (Valevski et al., 2025; Decart et al., 2024; Che et al., 2025; Zhang et al., 2025; Li et al., 2025a), 6-DoF camera trajectories (He et al., 2025; Li et al., 2025b), and hybrid inputs (Sun et al., 2025). Yet, prevailing models remain silent; while EchoWM (Zhang et al., 2026a) incorporates audio, it lacks camera-conditioned spatial acoustics. HelixWorld conditions a joint audio-visual backbone on continuous camera trajectories and discrete actions, steering viewpoints alongside responsive spatial acoustics.

Audio-visual generation. Cascaded systems dub pre-rendered video via video-to-audio synthesis (Cheng et al., 2025; The Movie Gen team, 2024), but post-hoc dubbing lacks camera extrinsics or control intents, and compounding latency prevents real-time interaction. In contrast, joint models (HaCohen et al., 2026; Google DeepMind, 2025) co-denoise video and audio tokens in a shared

transformer for cross-modal synchrony. Yet, they remain passive and bidirectional, operating on fixed-length clips without control interfaces or streaming inference. We bring interactive control and causal streaming to joint models.

Causal distillation. Deploying diffusion models interactively requires converting bidirectional generation into causal streaming. Recent video distillation schemes mitigate exposure bias through distribution matching (DMD) (Yin et al., 2025; 2024), on-policy KV rollouts (Self-Forcing) (Huang et al., 2025b), ODE regression (Zhu et al., 2026b), and long-horizon streaming tuning (Yang et al., 2025). However, these techniques target video alone. When extended to joint audio-visual diffusion, distribution matching suffers from multimodal mode collapse, causing color degradation and acoustic dropouts over extended rollouts. We stabilize self-forcing with an online trajectory distillation loss, anchoring the student's joint velocity field against teacher probability flows.

Audio-visual evaluation. World-model benchmarks such as WBench (Ying et al., 2026) assess action-conditioned video dynamics but ignore audio entirely. Conversely, audio-visual benchmarks (Chen et al., 2020) measure semantic correspondence (ImageBind (Girdhar et al., 2023), CLAP (Wu et al., 2023)) and onset synchrony (Synchformer (Iashin et al., 2024), AV-Align (Yariv et al., 2024)) on short, fixed-viewpoint clips. While these metrics evaluate *what* sounds occur and *when*, they cannot test *where* sounds originate relative to a moving listener under viewpoint motion. HELIXBENCH addresses this gap, formalizing *spatial-acoustic consistency* alongside temporal, semantic, and dynamic metrics for interactive audio-visual evaluation.

3 DATASET CONSTRUCTION

Training interactive audio-visual world models requires continuous dynamics, authentic stereo acoustics, and synchronized control. Yet existing datasets largely lack authentic spatial audio, while in-the-wild video suffers from shot cuts, overlays, and non-diegetic sound. We address this with an automated pipeline (Fig. 1) that filters multimodal artifacts while annotating metric camera poses, discrete actions, and decoupled captions to power bidirectional pre-training and causal streaming distillation. More details are provided in Appendix C.

3.1 DATA SOURCES

We assemble footage across three complementary sources (Table 1). *Real-world video* captures first-person exploration across diverse indoor and outdoor environments. *Game footage* records first- and third-person gameplay with native audio, logging ground-truth camera trajectories and control telemetry. *Open-source corpora* incorporate audio-equipped video subsets (Li et al., 2025c; Che et al., 2025), subjected to identical stereo and semantic curation.

Table 1: **Corpus by source.**

Source	Hours	Share
Real-world	1.8k	43.9%
Game	1.4k	34.1%
Open-source	0.9k	22.0%
Total	4.1k	100%

3.2 CURATION PIPELINE

Raw footage cannot be fed directly into training due to pervasive noise and artifacts across both visual and acoustic streams. We design a four-stage pipeline (Fig. 1) ordered by computational cost, pruning corrupted data before invoking heavy multimodal models.

Stage 1: Heuristic filtering. Raw videos are partitioned into single continuous shots via scene transition detection (Castellano, n.d.), discarding fragments under five seconds. Acoustically, lightweight probes measure inter-channel correlation and energy differences to eliminate mono downmixes and channel-duplicated tracks. Visually, optical character recognition (OCR) (Zhang et al., 2026b) and heuristic detectors prune overlays, banners, and watermarks. Perceptual quality scoring (Ke et al., 2021) and coarse motion screening then filter severely degraded or entirely frozen clips, while preserving stationary viewpoints capturing physical motion.

Stage 2: Clip standardization. Continuous shots are partitioned into fixed-length clips matching the backbone training window. Video and audio are sliced synchronously without crossing shot boundaries, ensuring temporal alignment by construction.

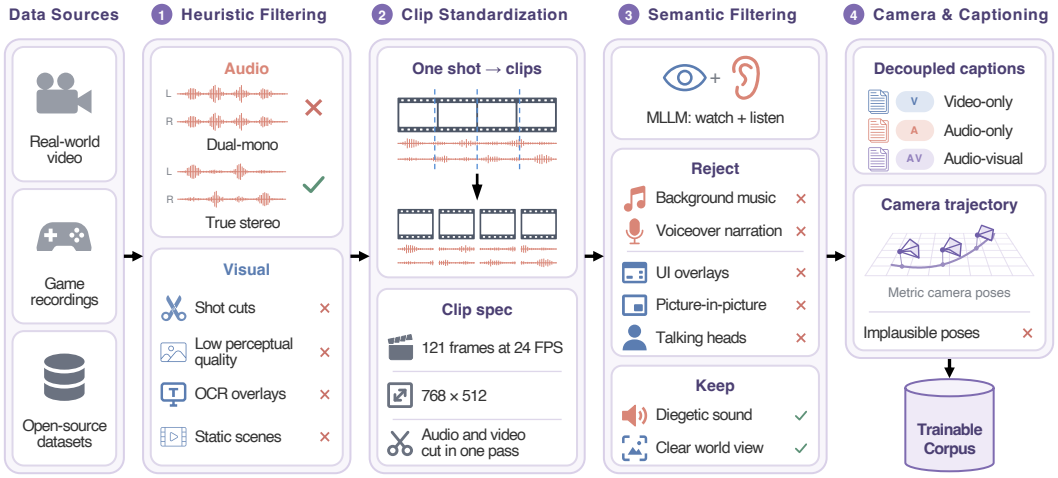


Figure 1: **Data pipeline.** Raw video and audio pass four cost-ordered stages; surviving clips fork into parallel camera recovery and decoupled captioning branches to yield the trainable corpus.

Stage 3: Semantic filtering. To resolve defects requiring high-level scene understanding, an audio-visual multimodal large language model (MLLM) (Qwen Team, 2025) evaluates clips under a structured perception schema. Visually, it purges non-embodied content (e.g., picture-in-picture, talking heads, occluding overlays) and resolves text ambiguous to OCR. Acoustically, it strips non-diegetic post-production tracks (background music, voiceover, synthetic effects). Crucially, sounds are judged by causal origin rather than on-screen visibility, ensuring off-screen environmental acoustics and natural speech are preserved.

Stage 4: Camera and captioning. Accepted clips enter two parallel annotation branches. For web and real-world footage, the camera branch estimates per-frame poses via VGGT-Ω (Wang et al., 2026a) and recovers metric scale with Depth Anything V3 (Lin et al., 2025), falling back to Metric3D v2 (Hu et al., 2024) when necessary; game footage uses engine-logged ground truth. After discarding tracking failures and physically implausible trajectories, surviving motions are discretized into an 81-class action vocabulary (Sec. 4.1; benchmarks in Appendix C). Concurrently, the captioning branch synthesizes three decoupled text tracks via independent forward passes: video-only, audio-only, and joint audio-visual descriptions that bind visible sound emitters to their acoustics. A clip is retained only when both branches succeed.

Table 2: **Curation yield.**

Stage	Hours	Pass Rate
Raw footage	4.1k	100%
Heuristic filtering	3.4k	81.9%
Semantic filtering	3.2k	95.5%
Camera annotation	3.0k	93.5%
Trainable corpus	3.0k	73.1%

In total, this multi-stage pipeline systematically purges visual artifacts and acoustic contamination across the 4.1k raw hours. Through successive screening and pose verification, 3.0k hours (amounting to 2.1M synchronized clips) survive to form the final trainable corpus, achieving an overall yield of 73.1% (Table 2). The resulting corpus provides a camera-conditioned audio-visual foundation with authentic stereo sound fields and per-frame action annotations, directly powering bidirectional pre-training and causal streaming distillation (Sec. 4).

4 Method

An interactive audio-visual world model requires responsive *action following* (steering visual dynamics and spatial sound) and causal *interactivity* (real-time generation). To this end, HelixWorld follows a two-stage training pipeline (Figs. 2 and 3): Sec. 4.1 trains a bidirectional teacher with progressive camera and action conditioning, while Sec. 4.2 distills it into a few-step causal student.

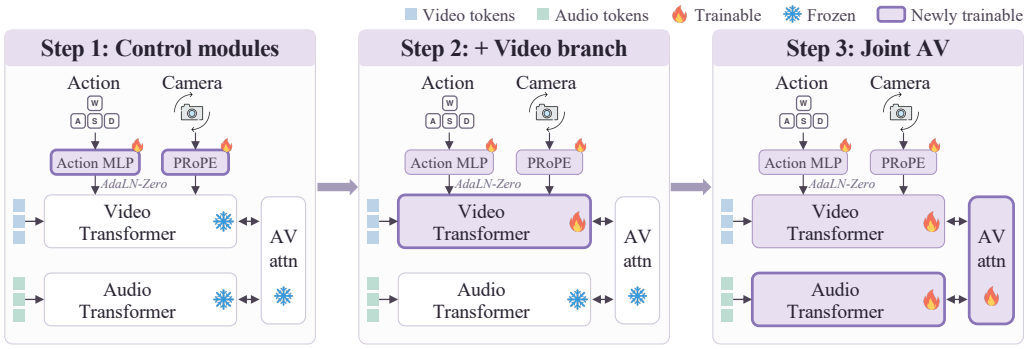


Figure 2: **Progressive condition injection.** Three-stage training: optimize control modules with the backbone frozen, unfreeze the video branch, and jointly train the full audio-visual backbone.

4.1 Bidirectional Teacher Training

We train the bidirectional model on our corpus (Sec. 3) by progressively injecting control signals. We employ both continuous camera trajectories and discrete user actions as control inputs. Both controls enter the video branch and propagate to the audio branch through cross-modal attention.

Condition Injection. Following Sun et al. (2025), we condition the video branch on both continuous camera trajectories and discrete actions. Continuous 6-DoF camera poses are injected via Projective Relative Positional Encoding (PRoPE) (Li et al., 2025b), transforming visual attention queries and keys by their respective camera projection matrices $\boldsymbol{P}_t = \mathrm{diag}(\boldsymbol{K}_{f,t}, 1)\boldsymbol{W}_t \in \mathbb{R}^{4\times4}$. Discrete actions are represented by an 81-class vocabulary corresponding to a 9×9 grid of translation and view-rotation velocity bins (Li et al., 2025a; Sun et al., 2025). For each frame t, the action index a_t is embedded and added to the diffusion noise level embedding to modulate the transformer backbone via adaptive layer normalization (AdaLN-Zero) (Peebles & Xie, 2023):

$$\boldsymbol{h}_{\mathrm{cond},t} = \mathrm{Emb}_\sigma(\sigma) + \mathrm{MLP}\big(\phi(a_t)\big). \tag{1}$$

Both controls enter the video branch and propagate to audio via cross-modal attention.

Progressive training. Directly tuning the entire backbone on unaligned controls destabilizes pre-trained representations. We therefore introduce conditioning in three progressive steps. First, we freeze the backbone and optimize only the control modules (the action MLP and PRoPE projections). Second, we unfreeze the video branch to adapt visual features to control inputs. Finally, we unfreeze the full audio-visual backbone for joint flow-matching optimization:

$$\mathcal{L}_{\mathrm{FM}} = \mathcal{L}^V_{\mathrm{FM}} + \lambda_A \mathcal{L}^A_{\mathrm{FM}}. \tag{2}$$

4.2 Causal Distillation

To enable real-time generation, we distill the bidirectional teacher into a few-step causal student. Addressing the visual drift of standard self-forcing (DMD2) (Huang et al., 2025b; Yin et al., 2024), we propose an *online trajectory distillation loss* that anchors student rollouts to the teacher's probability flow, complemented by long-horizon streaming tuning.

Causal initialization. We first synchronize video latents and co-temporal audio tokens into temporal blocks of length Δt. Attention within each block is fully bidirectional to maintain audio-visual interactions, while cross-block attention is strictly causal via a sliding Key-Value (KV) cache. Based on this structure, we adapt the teacher using the joint flow-matching objective $\mathcal{L}_{\mathrm{FM}}$ (Eq. 2) conditioned on ground-truth history firstly. To provide a stable starting checkpoint for few-step sampling, we then initialize the student by regressing its single-step predictions onto the Probability Flow ODE endpoints of the causal teacher (Zhu et al., 2026b):

$$\mathcal{L}_{\mathrm{init}}(\theta) = \mathbb{E}_{\boldsymbol{x}_\sigma,\sigma,\boldsymbol{c}} \left\| \hat{\boldsymbol{x}}^\theta_0(\boldsymbol{x}_\sigma, \sigma, \boldsymbol{c}) - \hat{\boldsymbol{x}}^{\mathrm{teacher}}_0(\boldsymbol{x}_\sigma, \boldsymbol{c}) \right\|^2_2, \tag{3}$$

where $\boldsymbol{x}_\sigma$ is the joint audio-visual latent, and $\boldsymbol{c}$ denotes historical context and conditions.

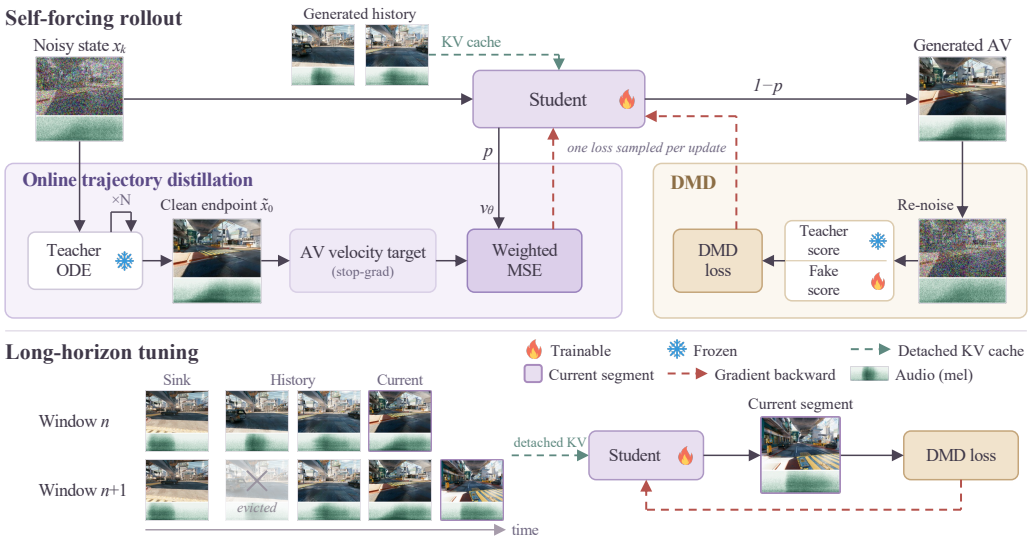


Figure 3: **Causal distillation pipeline.** *Top:* Self-forcing updates stochastically alternate between online trajectory distillation and distribution matching. *Bottom:* Long-horizon tuning extends rollouts segment by segment via a sliding KV cache with initial sink frames and detached history.

Self-forcing with online trajectory distillation. Following causal initialization, the student unrolls blocks autoregressively via self-forcing (Huang et al., 2025b) with its own KV cache. Standard self-forcing employs distribution matching ($\mathcal{L}_{\text{DMD}}$) (Yin et al., 2024), updating the student using the score difference between a frozen teacher and a learned fake-score model that tracks rollout distributions. However, distribution matching alone reduces sample diversity (Wu et al., 2026) and causes long-rollout visual drift (Fig. 6; Table 8). To maintain diversity and stabilize rollouts, we introduce an *online trajectory distillation loss* ($\mathcal{L}_{\text{traj}}$) alongside $\mathcal{L}_{\text{DMD}}$.

Specifically, at step k with noise σ_k and state $\boldsymbol{x}_k = (\boldsymbol{x}_k^V, \boldsymbol{x}_k^A)$, we integrate the frozen teacher's PF-ODE to $\sigma = 0$, obtaining target endpoint $\tilde{\boldsymbol{x}}_0 = (\tilde{\boldsymbol{x}}_0^V, \tilde{\boldsymbol{x}}_0^A)$. The student velocity $\boldsymbol{v}_\theta = (\boldsymbol{v}_\theta^V, \boldsymbol{v}_\theta^A)$ is then supervised to match the implied teacher flow:

$$\mathcal{L}_{\text{traj}}(\theta) = \sum_{m \in \{V,A\}} \lambda_m \left\| \boldsymbol{v}_\theta^m(\boldsymbol{x}_k, \sigma_k) - \text{sg}\left(\frac{\boldsymbol{x}_k^m - \tilde{\boldsymbol{x}}_0^m}{\sigma_k} \right) \right\|_2^2, \tag{4}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operator. Gradients are evaluated strictly at step k, with historical KV cache states and prior rollout steps detached. During training, each update stochastically samples either $\mathcal{L}_{\text{traj}}$ with probability p or $\mathcal{L}_{\text{DMD}}$ with probability $1-p$. This stochastic routing avoids gradient conflict between the two objectives, combining distribution matching for sample sharpness with trajectory guidance for rollout stability (hyperparameters in Appendix B).

Long-horizon streaming tuning. Extending rollouts to long horizons can still suffer from visual degradation and acoustic fading. To mitigate this drift, we incorporate streaming long tuning (Yang et al., 2025). As shown in Fig. 3, the student unrolls sequences segment by segment with a sliding KV cache, preserving initial sink frames as anchors while evicting distant history. Gradients from $\mathcal{L}_{\text{DMD}}$ are computed only on the active segment with history detached, keeping memory bounded.

5 HelixBench: A Benchmark for Audio-Visual World Models

Existing world-model benchmarks evaluate only silent video generation (Ying et al., 2026; Sun et al., 2025), while audio-visual benchmarks are restricted to static, non-interactive clips (Chen et al., 2020). In an interactive world model, the generated audio must sound realistic, synchronize with visual content, match text prompts, and adjust its stereo balance as the camera moves. To evaluate these properties during interactive rollouts, we introduce HelixBench.

Table 3: **Visual quality and action following on the WBench navigation split** (Ying et al., 2026). Audio indicates native audio synthesis; RTF is measured on a single NVIDIA H800.

Method	Audio	RTF ↓	Average ↑	Quality ↑	Setting ↑	Interaction ↑	Consistency ↑	Physical ↑
Alaya-EVOKE-Turbo	✗	4.29	82.0	81.9	82.1	83.9	88.1	74.0
EchoWM	✓	1.98	81.0	81.1	77.5	87.9	88.3	70.1
Zing-0.5	✗	2.81	81.0	80.6	77.8	84.2	88.5	73.8
LingBot-World v2	✗	5.21	79.4	81.8	76.8	82.8	86.5	69.1
HY-World 1.5	✗	12.33	78.1	78.1	72.2	86.8	86.9	66.3
Lyra 2.0	✗	30.87	76.4	77.1	73.2	85.6	79.3	66.7
SANA-WM	✗	0.79	76.0	79.3	76.1	82.2	80.7	61.9
DreamX-World	✗	1.04	75.0	77.5	80.8	78.6	74.9	63.3
Matrix-Game 3	✗	0.58	71.3	75.5	63.6	83.6	74.5	59.3
LTX-2.3 (base)	✓	14.61	74.2	77.1	85.2	66.4	77.2	64.9
HelixWorld (ours)	✓	0.77	79.9	79.2	76.5	86.4	86.4	70.9

Benchmark composition and metrics. HelixBench consists of 1,015 human-verified test clips across six scene domains and four interaction subsets: Navigation, Perspective, Event, and Subject. Each case provides a stereo reference clip, an initial frame, an audio-visual text prompt, and continuous camera poses. Across all clips, audio quality is measured by KL divergence (Koutini et al., 2022) and Fréchet Audio Distance (FAD) (Kilgour et al., 2019) on VGGish embeddings (Hershey et al., 2017), while multimodal alignment is evaluated via ImageBind (Girdhar et al., 2023) and CLAP (Wu et al., 2023). Finally, the Event subset measures temporal synchronization via DeSync (Iashin et al., 2024), while the Perspective subset evaluates whether stereo audio pans in coordination with on-screen sound sources (Spatial, Appendix D).

6 Experiments

6.1 Experimental Setup

Implementation and training. HelixWorld builds on open-source LTX-2.3 and follows the two-stage training procedure in Sec. 4. The full implementation details are provided in Appendix B.

Baselines. To evaluate visual quality, audio fidelity, and audio-visual alignment, we compare HelixWorld against strong baselines (Yin et al., 2026; Zhang et al., 2026a; Chen et al., 2026; Gao et al., 2026; Sun et al., 2025; Shen et al., 2026; Zhu et al., 2026a; DreamX Team et al., 2026; Wang et al., 2026b; Zhang et al., 2026a). Because EchoWM is the only world model supporting audio-visual generation, we also construct cascaded dubbing baselines using video-to-audio (V2A) models (Liu et al., 2025a;b; Tian et al., 2026). Specifically, we mute HelixWorld outputs and re-dub them with these models, keeping visual content fixed to isolate audio performance (details in Appendix D).

6.2 Main Results

Benchmark performance. To evaluate video quality, we benchmark on the WBench navigation split (Ying et al., 2026) (Table 3). HelixWorld outperforms most open-source silent baselines across visual quality and control responsiveness, demonstrating faithful action following and physical plausibility. To assess audio quality and multimodal coherence, we evaluate on HelixBench (Sec. 5) across audio fidelity, temporal sync, semantic alignment, and spatial consistency (Table 4; details in Appendix D). HelixWorld surpasses both the EchoWM and cascaded dubbing pipelines. While video-to-audio models achieve reasonable text alignment, they fail to capture emitter locations, resulting in negative spatial scores. EchoWM retains coarse spatial cues but falls behind in acoustic quality and semantic binding. These results demonstrate that HelixWorld effectively grounds spatial acoustics in interactive visual dynamics, achieving physically coherent audio-visual generation.

Inference efficiency. On a single NVIDIA H800 GPU, HelixWorld achieves a steady-state RTF of 0.77 at 768×512 and 24 fps, including full video and audio decoding. This enables real-time responsive rollouts to user control (details in Appendix D.7).

6.3 Ablation Studies

Progressive condition injection. To test whether staged conditioning prevents representation collapse (Sec. 4.1), we compare progressive against full fine-tuning. Evaluating ViPE-recovered trajectories (Huang et al., 2025a), progressive fine-tuning reduces both rotation and translation errors

Table 4: **Audio-visual quality on HelixBench**. HelixWorld+X re-dubs our video using V2A model X; joint generation achieves superior spatial acoustic alignment.

Model	KL ↓	FAD ↓	DeSync (s) ↓	IB ↑	CLAP ↑	Spatial ↑
LTX-2.3 (base)	1.8040	7.6354	**0.4042**	0.2777	0.2577	0.9826
EchoWM	1.9530	7.7472	0.6745	0.2036	0.2352	12.6476
HelixWorld + AudioX	2.0190	3.9166	1.1764	0.2569	**0.3094**	−5.7392
HelixWorld + ThinkSound	2.2335	6.1876	0.4236	0.2079	0.2888	14.8467
HelixWorld + PrismAudio	2.2427	6.3873	0.4582	0.2152	0.2716	−9.5072
HelixWorld (bidirectional teacher)	1.5414	3.0775	0.4988	0.2700	0.2676	33.2428
HelixWorld (causal student)	**1.3934**	**2.3872**	0.5867	**0.2987**	0.3016	**41.7583**

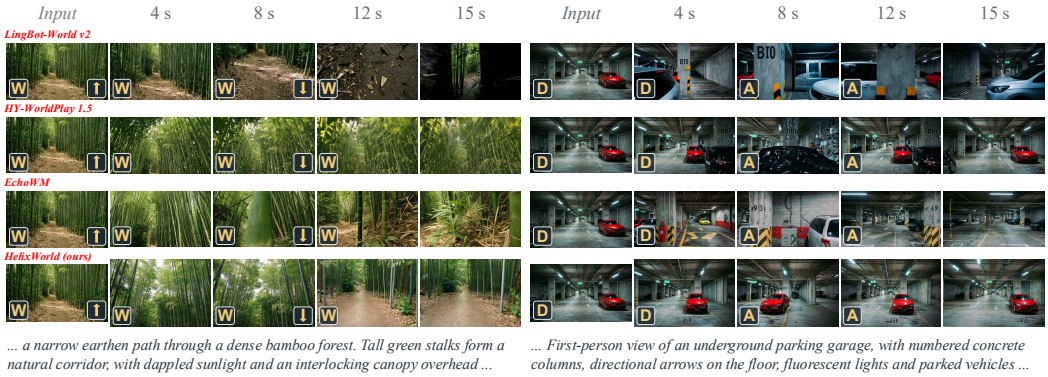


Figure 4: **Visual comparison under interactive control.** HelixWorld retains scene structure during navigation (left) and motion reversal (right). Key overlays indicate controls.

(Table 5). This confirms that gradually unfreezing branches is essential for learning geometric control while preserving pretrained visual representations.

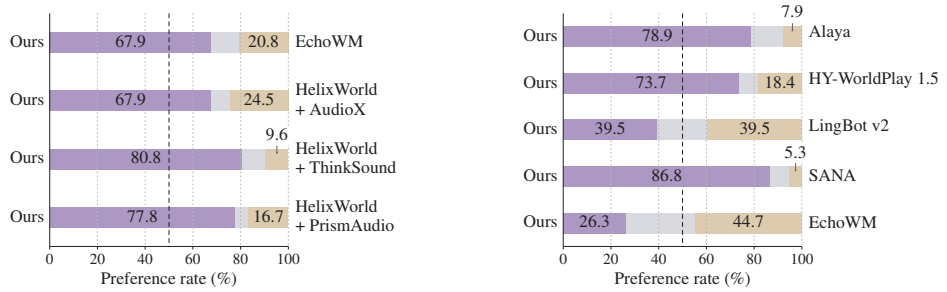
Multimodal caption composition. To test whether explicit sound-source binding improves joint generation, we compare three caption formats (Sec. 3): video-only (V), video-plus-audio (V+A), and three-track (V+A+AV). Cross-evaluating across prompt formats, three-track captions consistently yield the highest ImageBind similarity (Table 6). This confirms that richer audio-visual descriptions provide direct correspondence cues, strengthening cross-modal consistency.

Self-forcing with online trajectory distillation. To test whether online trajectory distillation ($\mathcal{L}_{\text{traj}}$) curbs rollout drift while preserving sample diversity (Sec. 4.2), we compare self-forcing training with and without $\mathcal{L}_{\text{traj}}$ on top of the DMD objective. As shown in Table 7, incorporating $\mathcal{L}_{\text{traj}}$ improves the overall WBench performance, with marked gains in consistency and physical plausibility. In extended 30-s rollouts, Figure 6 demonstrates that trajectory guidance visibly mitigates visual drift, maintaining cloud structures and scene landmarks with far fewer artifacts. Simultaneously, it boosts feature diversity on DINOv3 and CLIP over DMD alone (Table 8; setups in Appendix E). These findings validate that anchoring student rollouts to the teacher's probability flow successfully stabilizes autoregressive generation without causing mode collapse.

Long-horizon streaming tuning. To evaluate whether long-horizon tuning sustains generation quality over extended sequences (Sec. 4.2), we apply it to both self-forcing variants. As shown in Table 7, long-horizon tuning delivers consistent, orthogonal gains across both setups, improving overall performance over baselines. When combined with trajectory distillation, it achieves best overall score. This confirms that truncated streaming supervision effectively complements block-level trajectory guidance, enabling stable and interactive rollouts over extended horizons.

6.4 Qualitative Analysis and User Study

Qualitative results. Figure 4 compares HelixWorld with LingBot-World v2, HY-WorldPlay 1.5, and EchoWM on same scenes with shared first-frame and control inputs. In the bamboo scene, our model retains the forest corridor as the view rises into the canopy and returns to the path. In



(a) With audio (b) Muted video

Figure 5: **User-study preferences.** Purple indicates preference for HelixWorld, gray a tie, and beige preference for the baseline. Dashed lines mark 50%.

Table 5: **Ablation on camera control.**

Strategy	Rotation ↓	Translation ↓
Full fine-tuning	0.1237	0.0957
Progressive fine-tuning	**0.1144**	**0.0823**

Table 6: **Ablation on caption format.**

Train \ Infer	V	V+A	V+A+AV
V	0.2372	0.2346	0.2314
V+A	0.2366	0.2315	0.2419
V+A+AV	0.2316	0.2323	**0.2462**

Table 7: **Ablation on self-forcing and long-horizon tuning.** Trajectory distillation ($\mathcal{L}_{\text{traj}}$) and long-horizon tuning offer complementary gains, jointly reaching the best overall WBench score.

$\mathcal{L}_{\text{traj}}$	Long-Horizon	Average ↑	Quality ↑	Setting ↑	Interaction ↑	Consistency ↑	Physical ↑
✗	✗	77.9	77.0	75.3	85.0	85.3	66.9
✓	✗	78.7	77.8	74.1	86.1	**86.6**	**68.9**
✗	✓	78.4	79.0	**77.5**	86.5	83.7	65.1
✓	✓	**79.1**	**79.3**	77.1	**86.8**	84.6	67.7

Figure 6: **Visual quality in 30-s rollouts.** Self-forcing with $\mathcal{L}_{\text{traj}}$ (bottom) preserves cloud structure better than DMD only (top).

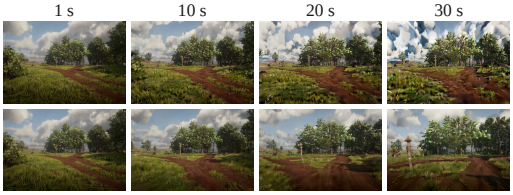


Table 8: **Visual diversity comparison.** Online trajectory distillation in self-forcing improves DINOv3 and CLIP diversity over DMD only.

Self-forcing	DINOv3 ↑	CLIP ↑
Without $\mathcal{L}_{\text{traj}}$	0.1089	0.0718
With $\mathcal{L}_{\text{traj}}$	**0.1326**	**0.0815**
Gain	+21.7%	+13.5%

the garage, the red car and nearby pillars remain recognizable through the right-to-left movement reversal. Additional audio–visual comparisons are provided in Appendix G.

User study. We conduct a blind pairwise user study across audio-visual and muted settings (Figure 5; Appendix F). With audio, HelixWorld is preferred over all baselines by a clear margin, winning the majority of votes in every matchup. In muted evaluations, it outperforms Alaya, HY-WorldPlay, and SANA, ties with LingBot v2, and trails only EchoWM.

7 Conclusion

In this work, we present HelixWorld, a real-time interactive audio-visual world model where visual scenes and spatial stereo soundscapes co-evolve natively under dynamic user control. Coupled with a compute-prioritized data curation pipeline, our causal distillation framework suppresses compounding rollout drift to sustain synchronized 24 FPS streaming on a single GPU. Finally, we formalize spatial-acoustic consistency as a foundational criterion separating interactive world models from passive generation, and introduce HelixBench to systematically evaluate it.

AI USE STATEMENT

In this work, generative AI tools were used for dataset curation and synthetic annotation (specifically, utilizing multimodal LLMs to filter out non-diegetic audio, visual occlusions, and corrupted clips as described in Sec. 3.2, and generating three-track decoupled captions). We have not used generative AI tools for formulating research hypotheses, developing theoretical models, designing experimental methodology, or interpreting experimental results. Additionally, generative AI tools were used for manuscript language polishing to improve readability and for formatting LaTeX code. All AI-assisted work was rigorously reviewed and verified by the authors: dataset curation decisions and generated captions were audited via clustered human spot-checks, all code and scripts were independently inspected and executed, and the final manuscript text was carefully examined to ensure technical and factual accuracy. The authors take full responsibility for the contents of this paper.

ETHICS STATEMENT

HelixWorld generates controllable synthetic audio-visual content, which carries potential societal risks associated with deceptive multimedia or fabricated footage. To address data privacy and ethical concerns, our training data are curated from publicly available web repositories and 3D virtual environments. As detailed in Sec. 3.2 and Appendix C.1, our automated curation pipeline systematically filters out talking-head footage, identifiable personal portraits, and private indoor spaces, thereby eliminating personally identifiable information (PII). To mitigate potential misuse, generated media will be distributed with provenance metadata and can be integrated with synthetic media detection and watermarking protocols. Furthermore, code and model checkpoints will be released exclusively for research and educational purposes under a responsible research license, accompanied by a code of conduct prohibiting deceptive or malicious deployment.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our work, comprehensive implementation and experimental details are provided throughout the manuscript and its appendix: the backbone architecture and conditioning modules are elaborated in Sec. 4 and Appendix A; the dataset curation criteria, filtering thresholds, and camera pose recovery procedures are detailed in Sec. 3 and Appendix C; the progressive pre-training protocol, causal distillation objectives, and hyperparameter configurations are described in Sec. 4.2 and Appendix B; and the benchmark protocol alongside full evaluation metrics are formalized in Sec. 5 and Appendix D. Furthermore, to facilitate community verification and foster future research in interactive audio-visual world modeling, all training and inference code, model weights, and the complete HelixBench evaluation suite will be made publicly available upon publication.

REFERENCES

Jake Bruce, Michael D. Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal M. P. Behbahani, Stephanie C. Y. Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott E. Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4603–4623. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/bruce24a.html.

Brandon Castellano. PySceneDetect, n.d. URL https://github.com/Breakthrough/PySceneDetect.

Haoxuan Che, Xuanhua He, Quande Liu, Cheng Jin, and Hao Chen. Gamegen-x: Interactive open-world game video generation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=8VG8tpPZhe.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pp. 721–725. IEEE, 2020. doi: 10.1109/ICASSP40776.2020.9053174. URL https://doi.org/10.1109/ICASSP40776.2020.9053174.

Mingyang Chen, Shengdong Chen, Xiaoxiao Fu, Bosheng Gong, Haoyuan Guo, Bowen Li, Jiawen Li, Kejun Li, Tianpeng Li, Yin Liu, et al. Zing-0.5: Toward playable worlds with real-time joint action and text control. *arXiv preprint arXiv:2609.17909*, 2026.

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander G. Schwing, and Yuki Mitsufuji. Mmaudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 28901–28911. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02691. URL https://openaccess.thecvf.com/content/CVPR2025/html/Cheng_MMAudio_Taming_Multimodal_Joint_Training_for_High-Quality_Video-to-Audio_Synthesis_CVPR_2025_paper.html.

Decart, Julian Quevedo, Quinn McIntyre, Spruce Campbell, Xinlei Chen, and Robert Wachen. Oasis: A universe in a transformer, 2024. URL https://oasis-model.github.io/.

DreamX Team, Rui Chen, Xiangxiang Chu, Geng Li, Jifan Li, Qingfeng Shi, Datao Tang, Jing Tang, Jun Wang, and Pengfei Zhang. Dreamx-phi 1.0: Action-conditioned video world model for robotic manipulation. *CoRR*, abs/2608.13489, 2026. doi: 10.48550/ARXIV.2608.13489. URL https://doi.org/10.48550/arXiv.2608.13489.

Zelin Gao, Qiuyu Wang, Jiapeng Zhu, Jingye Chen, Zichen Liu, Qingyan Bai, Jiahao Wang, Yufeng Yuan, Hanlin Wang, Yichong Lu, Ka Leong Cheng, Haojie Zhang, Jian Gao, Tianrui Feng, Yuzheng Liu, Yao Yao, Yinghao Xu, Xing Zhu, Yujun Shen, and Hao Ouyang. Infinite worlds with versatile interactions. *CoRR*, abs/2607.07534, 2026. doi: 10.48550/ARXIV.2607.07534. URL https://doi.org/10.48550/arXiv.2607.07534.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Google DeepMind. Veo 3: State-of-the-art video generation with audio. https://deepmind.google/models/veo/, 2025.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3):440, 2018.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, Eitan Richardson, Guy Shiran, Itay Chachy, Jonathan Chetboun, Michael Finkelson, Michael Kupchick, Nir Zabari, Nitzan Guetta, Noa Kotler, Ofir Bibi, Ori Gordon, Poriya Panet, Roi Benita, Shahar Armon, Victor Kulikov, Yaron Inger, Yonatan Shiftan, Zeev Melumian, and Zeev Farbman. LTX-2: efficient joint audio-visual foundation model. *CoRR*, abs/2601.03233, 2026. doi: 10.48550/ARXIV.2601.03233. URL https://doi.org/10.48550/arXiv.2601.03233.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for video diffusion models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=Z4evOUYrk7.

Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin W. Wilson. CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pp. 131–135. IEEE, 2017. doi: 10.1109/ICASSP.2017.7952132. URL https://doi.org/10.1109/ICASSP.2017.7952132.

Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10579–10596, 2024.

Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, Jiawei Ren, Kevin Xie, Joydeep Biswas, Laura Leal-Taixé, and Sanja Fidler. Vipe: Video pose engine for 3d geometric perception. *CoRR*, abs/2508.10934, 2025a. doi: 10.48550/ARXIV.2508.10934. URL https://doi.org/10.48550/arXiv.2508.10934.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025b. URL http://papers.nips.cc/paper_files/paper/2025/hash/f4823f831af67a3ef15e41a85434422a-Abstract-Conference.html.

Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pp. 5325–5329. IEEE, 2024. doi: 10.1109/ICASSP48485.2024.10448489. URL https://doi.org/10.1109/ICASSP48485.2024.10448489.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: multi-scale image quality transformer. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 5128–5137. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00510. URL https://doi.org/10.1109/ICCV48922.2021.00510.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In Gernot Kubin and Zdravko Kacic (eds.), *20th Annual Conference of the International Speech Communication Association, Interspeech 2019, Graz, Austria, September 15-19, 2019*, pp. 2350–2354. ISCA, 2019. doi: 10.21437/INTERSPEECH.2019-2219. URL https://doi.org/10.21437/Interspeech.2019-2219.

Khaled Koutini, Jan Schlüter, Hamid Eghbal-zadeh, and Gerhard Widmer. Efficient training of audio transformers with patchout. In Hanseok Ko and John H. L. Hansen (eds.), *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pp. 2753–2757. ISCA, 2022. doi: 10.21437/INTERSPEECH.2022-227. URL https://doi.org/10.21437/Interspeech.2022-227.

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. *CoRR*, abs/2506.17201, 2025a. doi: 10.48550/ARXIV.2506.17201. URL https://doi.org/10.48550/arXiv.2506.17201.

Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. Cameras as relative positional encoding. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025b. URL http://papers.nips.cc/paper_files/paper/2025/hash/17a7075094632c88cccdd86270ad715b-Abstract-Conference.html.

Zhen Li, Chuanhao Li, Xiaofeng Mao, Shaoheng Lin, Ming Li, Shitian Zhao, Zhaopan Xu, Xinyue Li, Yukang Feng, Jianwen Sun, Zizhen Li, Fanrui Zhang, Jiaxin Ai, Zhixiang Wang,

Yuwei Wu, Tong He, Yunde Jia, and Kaipeng Zhang. Sekai: A video dataset towards world exploration. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025c. URL http://papers.nips.cc/paper_files/paper/2025/hash/2368d4a3f24122b9f3779aaffa261719-Abstract-Datasets_and_Benchmarks_Track.html.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *CoRR*, abs/2511.10647, 2025. doi: 10.48550/ARXIV.2511.10647. URL https://doi.org/10.48550/arXiv.2511.10647.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22160–22169. IEEE, 2024.

Huadai Liu, Kaicheng Luo, Wen Wang, Qian Chen, Peiwen Sun, Rongjie Huang, Xiangang Li, Jieping Ye, and Wei Xue. Prismaudio: Decomposed chain-of-thoughts and multi-dimensional rewards for video-to-audio generation. *CoRR*, abs/2511.18833, 2025a. doi: 10.48550/ARXIV.2511.18833. URL https://doi.org/10.48550/arXiv.2511.18833.

Huadai Liu, Jialei Wang, Kaicheng Luo, Wen Wang, Qian Chen, Zhou Zhao, and Wei Xue. Thinksound: Chain-of-thought reasoning in multimodal large language models for audio generation and editing. *CoRR*, abs/2506.21448, 2025b. doi: 10.48550/ARXIV.2506.21448. URL https://doi.org/10.48550/arXiv.2506.21448.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLVII*, volume 15105 of *Lecture Notes in Computer Science*, pp. 38–55. Springer, 2024. doi: 10.1007/978-3-031-72970-6_3. URL https://doi.org/10.1007/978-3-031-72970-6_3.

Jack Parker-Holder, Philip Ball, Jake Bruce, Vibhavari Dasagi, Kristian Holsheimer, Christos Kaplanis, Alexandre Moufarek, Guy Scully, Jeremy Shar, Jimmy Shi, Stephen Spencer, Jessica Yung, Michael Dennis, Sultan Kenjeyev, Shangbang Long, Vlad Mnih, Harris Chan, Maxime Gazeau, Bonnie Li, Fabio Pardo, Luyu Wang, Lei Zhang, Frederic Besse, Tim Harley, Anna Mitenkova, Jane Wang, Jeff Clune, Demis Hassabis, Raia Hadsell, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 2: A large-scale foundation world model. https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model/, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 4172–4182. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00387. URL https://doi.org/10.1109/ICCV51070.2023.00387.

Qwen Team. Qwen3-omni technical report. *CoRR*, abs/2509.17765, 2025. doi: 10.48550/ARXIV.2509.17765. URL https://doi.org/10.48550/arXiv.2509.17765.

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

Tianchang Shen, Sherwin Bahmani, Kai He, Sangeetha Grama Srinivasan, Tianshi Cao, Jiawei Ren, Ruilong Li, Zian Wang, Nicholas Sharp, Zan Gojcic, Sanja Fidler, Jiahui Huang, Huan Ling, Jun

Gao, and Xuanchi Ren. Lyra 2.0: Explorable generative 3d worlds. *CoRR*, abs/2604.13036, 2026. doi: 10.48550/ARXIV.2604.13036. URL https://doi.org/10.48550/arXiv.2604.13036.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

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. *CoRR*, abs/2512.14614, 2025. doi: 10.48550/ARXIV.2512.14614. URL https://doi.org/10.48550/arXiv.2512.14614.

The Movie Gen team. Movie gen: A cast of media foundation models. *CoRR*, abs/2410.13720, 2024. doi: 10.48550/ARXIV.2410.13720. URL https://doi.org/10.48550/arXiv.2410.13720.

Zeyue Tian, Zhaoyang Liu, Yizhu Jin, Ruibin Yuan, Liumeng Xue, Xu Tan, Qifeng Chen, Wei Xue, and Yike Guo. Audiox: a unified framework for anything-to-audio generation. In *International Conference on Learning Representations*, volume 2026, pp. 110625–110651, 2026.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=P8pqeEkn1H.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotný. VGGT: visual geometry grounded transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 5294–5306. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.00499. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_VGGT_Visual_Geometry_Grounded_Transformer_CVPR_2025_paper.html.

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*, 2026a.

Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, Yidan Xietian, Jiangbo Pei, Liang Hu, Boyi Jiang, Hua Xue, Zidong Wang, Haofeng Sun, Wei Li, Wanli Ouyang, Xianglong He, Yang Liu, Yangguang Li, and Yahui Zhou. Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory. *CoRR*, abs/2604.08995, 2026b. doi: 10.48550/ARXIV.2604.08995. URL https://doi.org/10.48550/arXiv.2604.08995.

Tianhe Wu, Ruibin Li, Lei Zhang, and Kede Ma. Diversity-preserved distribution matching distillation for fast visual synthesis. *CoRR*, abs/2602.03139, 2026. doi: 10.48550/ARXIV.2602.03139. URL https://doi.org/10.48550/arXiv.2602.03139.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pp. 1–5. IEEE, 2023. doi: 10.1109/ICASSP49357.2023.10095969. URL https://doi.org/10.1109/ICASSP49357.2023.10095969.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, Song Han, and Yukang Chen. Longlive: Real-time interactive long video generation. *CoRR*, abs/2509.22622, 2025. doi: 10.48550/ARXIV.2509.22622. URL https://doi.org/10.48550/arXiv.2509.22622.

Guy Yariv, Itai Gat, Sagie Benaim, Lior Wolf, Idan Schwartz, and Yossi Adi. Diverse and aligned audio-to-video generation via text-to-video model adaptation. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (eds.), *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*, pp. 6639–6647. AAAI Press, 2024. doi: 10.1609/AAAI.V38I7.28486. URL https://doi.org/10.1609/aaai.v38i7.28486.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Frédo Durand, and Bill Freeman. Improved distribution matching distillation for fast image synthesis. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/54dcf25318f9de5a7a01f0a4125c541e-Abstract-Conference.html.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Frédo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 22963–22974. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02138. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yin_From_Slow_Bidirectional_to_Fast_Autoregressive_Video_Diffusion_Models_CVPR_2025_paper.html.

Yuanyang Yin, Gongxuan Wang, Yifan Zhang, Chuanhao Li, Kaipeng Zhang, and Feng Zhao. Alaya-evoke: From linear-scaling supervision to endless world. *CoRR*, abs/2608.13546, 2026. doi: 10.48550/ARXIV.2608.13546. URL https://doi.org/10.48550/arXiv.2608.13546.

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. *CoRR*, abs/2605.25874, 2026. doi: 10.48550/ARXIV.2605.25874. URL https://doi.org/10.48550/arXiv.2605.25874.

Songchun Zhang, Yaowei Li, Junhao Zhuang, Weiyang Jin, Haoyu Wang, Xin Lu, Yilang Sun, Shiyi Zhang, Haoran Li, Xiaoxiao Ma, Yuming Li, Yijun Liu, Yaofeng Su, Yanwen Ma, Haoyu Wu, Zihan Su, Yue Ma, Lvmin Zhang, Haoyang Huang, Zeyue Xue, Anyi Rao, and Nan Duan. Echowm: Open and enterable omnimodal world models. *CoRR*, abs/2608.23189, 2026a. doi: 10.48550/ARXIV.2608.23189. URL https://doi.org/10.48550/arXiv.2608.23189.

Yifan Zhang, Chunli Peng, Boyang Wang, Puyi Wang, Qingcheng Zhu, Fei Kang, Biao Jiang, Zedong Gao, Eric Li, Yang Liu, and Yahui Zhou. Matrix-game: Interactive world foundation model. *CoRR*, abs/2506.18701, 2025. doi: 10.48550/ARXIV.2506.18701. URL https://doi.org/10.48550/arXiv.2506.18701.

Yubo Zhang, Xueqing Wang, Manhui Lin, Yue Zhang, Penglongyi Deng, Ting Sun, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Changda Zhou, et al. Pp-ocrv6: From 1.5 m to 34.5 m parameters, surpassing billion-scale vlms on ocr tasks. *arXiv preprint arXiv:2606.13108*, 2026b.

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. *CoRR*, abs/2605.15178, 2026a. doi: 10.48550/ARXIV.2605.15178. URL https://doi.org/10.48550/arXiv.2605.15178.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. *CoRR*, abs/2602.02214, 2026b. doi: 10.48550/ARXIV.2602.02214. URL https://doi.org/10.48550/arXiv.2602.02214.

A Architecture and Conditioning Details

Camera conditioning. The camera branch supplements the video self-attention in Sec. 4.1. Intrinsics are normalized to the latent grid, and extrinsics are expressed relative to the first frame, whose relative pose is the identity. The focal-only lift retains focal lengths but omits the principal point. Invalid or missing camera tokens, as well as audio tokens, receive identity transforms. The camera branch shares the base attention's Q/K/V projections and contributes through a zero-initialized output projection.

Action conditioning. The 81 action classes factor as $\mathrm{id} = 9\,\mathrm{translation} + \mathrm{rotation}$. Translation and rotation each enumerate nine bins in $\{-, 0, +\}^2$, for local (x, z) translation and (yaw, pitch), respectively. Each video latent frame has its own action index. The index is encoded with 256-dimensional sinusoidal features and a Linear–SiLU–Linear MLP with a zero-initialized output, then added to the noise-level embedding. Camera and action controls enter the video branch and reach audio through cross-modal attention.

B Training and Streaming Inference Details

B.1 Temporal Blocks and First-Frame Conditioning

A short self-forcing window contains 16 video latent frames and 127 audio tokens. The video is partitioned into four blocks of four latent frames; the corresponding audio blocks contain 27, 33, 33, and 34 tokens, assigned by their timestamps. The conditioning image occupies the first latent of the first block. It participates in joint audio–visual attention at noise level zero, is restored after each denoising and re-noising operation, and is excluded from the generated-token loss. The remaining 15 video latents and all audio tokens are generated.

Attention is bidirectional within a block and causal across blocks. We detach historical latents and committed KV tensors and retain gradients within the current block. Camera and action sequences are aligned on the source timeline before temporal blocks are selected.

B.2 Causal Initialization

Teacher forcing. We adapt the bidirectional model to block-causal attention using clean ground-truth history and the joint audio–visual flow-matching objective. Short clips supply four consecutive blocks. For long sources, the initialization recipe packs the first block with a later contiguous three-block window and supervises the final target block. This exposes the model to later source positions within a bounded context.

ODE initialization. The student regresses onto precomputed endpoints of the teacher-forced causal model (Eq. 3). The teacher uses 50 dense solver steps, CFG 3, and no STG. Input states are sampled from the four nonzero student knots $\{1.0, 0.9, 0.7, 0.4\}$. The image-conditioned initialization uses the first frame on every example and optimizes the joint video/audio endpoint MSE with a constant learning rate of 2×10^{-6}.

B.3 Self-Forcing Optimization

The generator starts from the ODE-initialized causal model. The frozen real-score teacher and the trainable fake-score model both start from the bidirectional model after progressive condition injection. Generator and fake-score adaptation use LoRA with the shared settings in Table 9.

Rollout and gradient path. The student denoises each block with the schedule $[1.0, 0.9, 0.7, 0.4, 0]$, drawing independent video and audio noise for each re-noising transition. A clean, detached KV commit follows each completed block, and later blocks append to this history. A short training window contains four blocks, all retained in the history context. The loss covers generated tokens across the window, with gradients retained at one selected denoising step per block.

Table 9: **Self-forcing hyperparameters.** Settings shared by DMD-only self-forcing and self-forcing with online trajectory distillation loss.

Setting	Value
Generator / fake-score adaptation	LoRA rank 256, alpha 256, dropout 0
Optimizer	AdamW, $\beta_1 = 0.9$, $\beta_2 = 0.999$
Generator / fake learning rate	10^{-5}, constant; no warmup or decay
Gradient norm clipping	1.0
Precision	BF16 with gradient checkpointing
Global batch size	48; one sample per GPU, no accumulation
Training hardware	Six nodes, each with eight H200 GPUs
Fake updates per DMD generator event	5
Teacher video / audio CFG	4 / 2
Teacher STG	Scale 1, transformer block 29
Teacher guidance rescale / modality scale	0 / 1
Student inference CFG / STG	1 / 0

DMD score evaluation. The bidirectional real and fake models evaluate the same noised student sample at the same physical noise level. For the self-forcing recipe, a sampled raw level u is shifted once as

$$\sigma = \frac{\gamma u}{1 + (\gamma - 1)u}, \qquad \gamma = 3. \tag{5}$$

Generator-score queries use $u \sim \mathcal{U}(0.02, 0.98)$; fake-score training uses $u \sim \mathcal{U}(0, 1)$. The resulting σ is shared by video and audio and used consistently for noising, model queries, and clean-sample conversion. Student rollouts use the fixed four-knot schedule above.

Online trajectory distillation. With probability 0.1, a generator event uses Eq. 4 in place of a DMD update; otherwise it follows five fake-score updates and one DMD generator update. For a trajectory event, one student denoising index is shared across the four blocks. The frozen bidirectional teacher integrates from the corresponding noisy state to a detached endpoint on a dense schedule with eight steps per student interval (32 over the full schedule). Gradients update the selected student predictions while conditioning-image tokens and historical KV states remain fixed.

B.4 Long-Horizon Tuning and Streaming Inference

Long-horizon tuning. The student rolls out approximately one minute of synchronized video and audio, advancing through the source in temporal order. The DMD loss supervises the newest window, with earlier latents and KV states detached. The bidirectional score models retain the short-window geometry of 16 video latents and 127 audio tokens. Each continuation starts from its self-forcing generator and the corresponding fake-score checkpoint. Table 7 compares a DMD-only self-forcing checkpoint with its DMD-only long-horizon continuation; neither stage uses online trajectory distillation in this experiment. The separate self-forcing comparison in Table 7 evaluates online trajectory distillation at matched generator-update counts, before long-horizon tuning.

Bounded inference history. The inference context contains the fixed first block, the two most recent completed blocks, and the current target block: at most four blocks in total. The first block retains the conditioning-image latent. For example, the fifth target block reads blocks 1, 3, and 4. Temporal keys are cached before RoPE; video and audio positions share a compact local timeline for the retained context.

Cache reconstruction. The reported rebuild evaluations recompute history KV from the selected clean, generated latents before denoising the next target. Each history block attends to itself and earlier retained blocks under the block-causal mask. Reconstruction updates cached keys and values while keeping the generated history latents fixed. The visual quality and diversity comparisons use the same rebuild policy in both arms. Inference jointly generates both modalities with four stochastic denoising steps, CFG 1, and STG 0.

C Dataset Construction and Curation Details

This section provides comprehensive details on the data curation and annotation pipeline described in Sec. 3, including domain distribution and ethical privacy considerations, signal-level thresholds, MLLM perception schema, camera pose and metric scale recovery, and multimodal caption contracts.

C.1 Data Sources, Domain Distribution, and Ethical Considerations

We assemble raw footage totaling $4.1\,\mathrm{k}$ hours across three complementary domains (Table 10), systematically balancing real-world dynamics, interactive gaming environments, and diverse open-source distributions.

Real-world web video. We collect $1.8\,\mathrm{k}$ hours of high-resolution first-person exploration, urban driving, and environmental walkthroughs from publicly available video platforms. To safeguard personal privacy and prevent unauthorized disclosure, we apply automated face de-identification and license-plate obfuscation during pre-processing; any footage centered on identifiable individuals or private personal spaces is systematically purged in Stage 1 and Stage 3.

Game screen recordings. We record $1.4\,\mathrm{k}$ hours of high-dynamic first- and third-person navigation across diverse 3D virtual environments and modern game engines, capturing native multi-channel engine audio and synchronized camera trajectories. Direct access to engine telemetry provides noise-free ground-truth camera extrinsics and user control inputs, serving as a reliable geometric anchor for interactive control learning.

Open-source corpora. We incorporate $0.9\,\mathrm{k}$ hours from open-domain video corpora with native audio tracks, including Sekai (Li et al., 2025c) and GameGen-X (Che et al., 2025), covering diverse open-world scenes and interactive scenarios. These subsets are subjected to identical stereo and semantic curation pipelines to ensure corpus-wide distribution uniformity.

Table 10: **Data sources, domain distribution, and corpus statistics.** Pose origin indicates whether camera trajectories are logged directly from engine states or reconstructed offline via geometric estimation. All statistics reflect the frozen corpus.

Source Family	Pose Origin	Raw Hours	Share	Trainable Clips
Real-world web video	Estimated (VGGT-Ω + DA3)	$1.8\,\mathrm{k}$	43.9%	$\sim 0.91\,\mathrm{M}$
Game screen recordings	Engine-logged (ground truth)	$1.4\,\mathrm{k}$	34.1%	$\sim 0.73\,\mathrm{M}$
Open-source subsets	Mixed (logged / estimated)	$0.9\,\mathrm{k}$	22.0%	$\sim 0.46\,\mathrm{M}$
Total Raw Footage	—	$4.1\,\mathrm{k}$	100%	$\sim 2.93\,\mathrm{M}$
Trainable Corpus	Verified metric poses	$\mathbf{3.0}\,\mathrm{k}$	$\mathbf{73.1}\%$	$\mathbf{2.10}\,\mathrm{M}$

C.2 Signal-Level Filtering and Clip Standardization (Stages 1 & 2)

Raw footage exhibits pervasive distribution corruption, including dual-mono downmixes, rapid montage cuts, static title cards, and interface clutter. Stage 1 executes lightweight signal probes to eliminate defective footage prior to compute-intensive neural processing.

Container and format validation. The ingestion probe enforces strict container constraints: video streams must have a spatial resolution of at least 768×512 with horizontal landscape aspect ratio, and frame rates bounded within $[23.9, 121]$ fps (subsequently normalized to 24 fps). The audio stream must contain a valid two-channel stereo track, immediately pruning mono downmixes and missing audio streams. Clips failing container integrity or exhibiting corrupt stream headers are pruned immediately.

True stereo verification. A significant fraction of online videos duplicate a single mono recording across two channels, creating "dual-mono" tracks devoid of spatial acoustic information. We measure inter-channel stereo energy over sliding 20-second analysis windows via normalized root-mean-square difference:

$$\delta = \frac{\|L - R\|_2^2}{\|L\|_2^2 + \|R\|_2^2}, \tag{6}$$

where L and R denote the discrete time-domain signals of the left and right channels, respectively. For identical channels, $\delta \equiv 0$, whereas authentic spatial acoustics yield $\delta \in [0.05, 1.0]$. We discard any upload with $\delta < 0.01$. Surviving streams are subsequently verified for phase correlation to eliminate artificial out-of-phase stereo widening artifacts.

Shot transition and temporal continuity. We employ PySceneDetect (Castellano, n.d.) to segment continuous shots, tracking HSV color histogram differences across frames subsampled at 4 fps. A cut boundary is placed whenever the histogram difference exceeds 0.35. Fragmented shots shorter than 5 seconds are dropped to prevent abrupt scene cuts from contaminating the temporal attention window. Furthermore, any video whose mean continuous shot duration falls below 2 seconds (e.g., promotional trailers or rapid montages) is rejected in its entirety.

Static scene discrimination. To eliminate static still-image slideshows and frozen streams while retaining stationary cameras that observe active physical events (e.g., vehicles driving past a fixed camera), we compute the Median Absolute Difference (MAD) between consecutive grayscale frames sampled at 1 fps:

$$\mathrm{MAD} = \mathrm{median}_{(x,y)} \left| I_{t+1}(x, y) - I_t(x, y) \right|. \tag{7}$$

Shots with $\mathrm{MAD} < 1.5$ (on an 8-bit scale $[0, 255]$) are rejected. Employing the median rather than the mean prevents localized high-frequency perturbations (such as blinking watermarks, subtitle updates, or compression noise) from falsely validating an otherwise frozen scene.

Clip standardization (Stage 2). Surviving continuous shots are segmented into standardized clips matching the backbone training context: exactly 121 video frames at 24 fps (5.04 s) at 768×512 resolution, coupled with synchronized two-channel 48 kHz stereo audio (241,920 audio samples per clip). Shots are tiled front-to-back without overlap. To maximize data efficiency, whenever a trailing segment exceeds 2.5 s, an additional clip is carved backwards from the shot terminus, raising temporal coverage from 98.4% to 100%. Audio and video streams are sliced synchronously in a single ffmpeg pass, ensuring microsecond-level temporal synchronization by construction.

C.3 Semantic Filtering with Audio-Visual MLLM (Stage 3)

While low-level heuristic probes eliminate signal defects, they cannot reason about high-level multimodal semantics: pixel differences fail to separate 3D physical environments from 2D screen recordings or talking heads, and audio format checks cannot detect added background music. Stage 3 deploys an audio-visual foundation model (Qwen3-Omni-30B (Qwen Team, 2025)) to evaluate video and audio synchronously under constrained grammar decoding. A clip is retained if and only if it satisfies two physical criteria:

- **Visual scene integrity**: The clip must depict an authentic 3D physical environment undergoing dynamic physical evolution, strictly excluding non-world content (e.g., talking-head interviews, desktop screencasts, picture-in-picture feeds) and intrusive graphic or subtitle overlays.
- **Acoustic authenticity**: The clip must feature natural diegetic sound, systematically pruning post-production background music, voiceover narration, and artificial sound effects. Crucially, acoustic filtering evaluates sounds by their *physical genesis* rather than source visibility: off-screen environmental sounds (e.g., thunder or passing vehicles) are preserved as authentic diegetic acoustics.

Because soundtrack contamination typically spans an entire video, we apply an early termination heuristic: if 10 consecutive clips within an upload are flagged for non-diegetic audio, the remainder of that upload is terminated immediately, saving substantial inference compute.

Table 11: **Mean camera estimation errors and runtime across benchmark scenes.** Translation L1 and RMSE are normalized relative to reference trajectory RMS radius after $\mathrm{Sim}(3)$ alignment; rotation error reflects global orientation alignment. Runtime includes initialization and excludes depth recovery and window stitching. VGGT and VGGT-Omega process 32 evaluation frames; ViPE processes 163–399 frames before subsampling.

Method	Translation (%) ↓		Rotation (°) ↓	Runtime (s) ↓
	L1	RMSE		
VGGT (Wang et al., 2025)	1.010	0.760	0.315	**14.23**
VGGT-Omega (Wang et al., 2026a)	1.001	0.768	**0.234**	16.09
ViPE (Huang et al., 2025a)	**0.763**	**0.582**	0.263	187.32

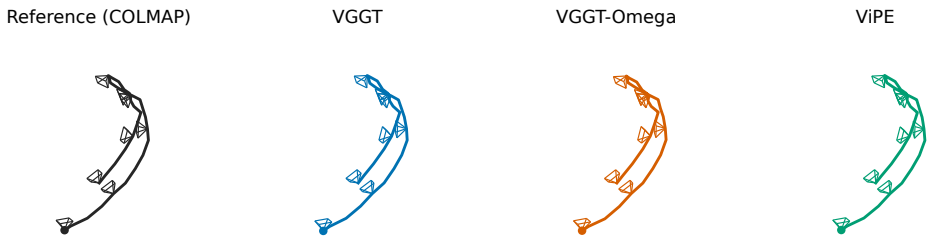


Figure 7: **Single-sample camera trajectory comparison.** From left to right: COLMAP reference, VGGT, VGGT-Omega, and ViPE. All trajectories are aligned to the reference coordinate frame via Sim(3). Solid lines denote camera translation paths; schematic frustums indicate camera orientations.

C.4 Camera Pose Estimation and Metric Scale Recovery (Stage 4)

High-fidelity interactive world modeling requires precise, metric-scaled 6-DoF camera trajectories. For game footage, camera extrinsics and focal parameters are logged directly from engine telemetry. For real-world and open-source footage, we develop a multi-view geometric estimation and metric scale recovery pipeline.

Camera pose estimator benchmark. We evaluate video geometry foundation models on representative scenes from the DL3DV-10K benchmark (Ling et al., 2024), comparing VGGT-Omega (Wang et al., 2026a) against VGGT (Wang et al., 2025) and ViPE (Huang et al., 2025a) using COLMAP reference poses across 32 matched evaluation frames. As reported in Table 11, VGGT-Omega achieves the lowest rotation error ($0.234°$) while maintaining inference efficiency comparable to VGGT (16.09 s vs 187.32 s for ViPE). Figure 7 illustrates a representative qualitative trajectory comparison.

Metric scale recovery and temporal stitching. For each video clip, VGGT-Omega outputs up-to-scale relative camera extrinsics $\{\boldsymbol{R}_t, \tilde{\boldsymbol{t}}_t\}_{t=1}^{T}$ and dense relative depth maps $\tilde{D}_t \in \mathbb{R}^{H\times W}$. Because interactive world models require metric physical dimensions to ground linear velocities and spatial audio inverse-square attenuation, we recover physical metric scale using monocular metric depth estimation via Depth Anything 3 (DA3) (Lin et al., 2025), with Metric3D-v2 (Hu et al., 2024) serving as an architectural fallback for scenes with extreme perspective distortion. Specifically, for keyframes sampled across the clip (at 2 fps), we compute the pixel-wise metric scale ratio over a confidence mask $\mathcal{M}_t$ excluding sky regions and specular highlights:

$$r_t(u,v) = \frac{D_t^{\mathrm{metric}}(u,v)}{\tilde{D}_t(u,v)}, \quad (u,v) \in \mathcal{M}_t. \tag{8}$$

The global metric scale factor s for the clip is computed via the robust median across all valid keyframe pixels: $s = \mathrm{median}_{(t,u,v)\in\mathcal{M}}\, r_t(u,v)$, from which metric translations are recovered via

$\boldsymbol{t}_t = s \cdot \tilde{\boldsymbol{t}}_t$. When estimating extended shots across multiple consecutive 5.04-second windows, the estimator operates with a temporal sliding window overlapping by 12 frames (0.5 s). We solve an $\mathrm{SE}(3)$ registration problem across the overlapping frames to align adjacent coordinate systems, measuring trajectory continuity by the residual camera center RMS displacement:

$$\mathrm{RMS}_{\mathrm{stitch}} = \sqrt{\frac{1}{K}\sum_{k=1}^{K} \|\hat{\boldsymbol{t}}_k^{(w)} - \boldsymbol{t}_k^{(w-1)}\|_2^2}, \tag{9}$$

where $\hat{\boldsymbol{t}}_k^{(w)}$ denotes the camera position in window w mapped into the reference frame of window $w-1$.

Kinematic plausibility gating and action derivation. To eliminate tracking failure artifacts and unphysical camera jumps, all estimated trajectories are validated against kinematic physical bounds: (i) mean velocity $\bar{v} \in [0, 40)$ m/s, (ii) maximum instantaneous velocity $v_{\max} \leq 50$ m/s, (iii) maximum linear acceleration $a_{\max} \leq 20\,\mathrm{m/s^2}$, and (iv) cross-window stitching error $\mathrm{RMS}_{\mathrm{stitch}} \leq 0.05$ m. No minimum-motion threshold is enforced, preserving stationary viewpoints and pure in-place rotations. Finally, continuous relative camera motions between adjacent frames are quantized into the 81-class discrete action vocabulary (9×9 bins of linear velocity and yaw rotation rate) detailed in Sec. 4.1.

C.5 Decoupled Multimodal Captioning (Stage 4)

Conditioning an interactive audio-visual world model on a single entangled text prompt induces cross-modal hallucination during training (e.g., visual models hallucinating sound effects that are absent, or audio models generating voices when text mentions visible actors). To prevent cross-modal contamination, we formulate a decoupled caption contract comprising three independent caption streams per clip, generated using Qwen3-Omni-30B (Qwen Team, 2025) via isolated forward passes:

- **Visual caption** (V): Describes setting and salient subject motion while strictly ignoring audio. Viewpoint changes and camera-induced parallax are forbidden in the text, ensuring camera dynamics remain exclusively governed by camera conditioning.
- **Audio caption** (A): Describes acoustic layers, event order, and temporal dynamics while ignoring visual content. Because captioning uses a mono downmix, directional terms (e.g., left/right) are strictly prohibited, preserving stereo panning for geometric control. Speech is treated as ambient sound without transcription.
- **Joint audio-visual caption** (AV): Explicitly grounds audible events to visible on-screen emitters, tracking cross-modal synchronization and volume changes as sound sources enter or leave the camera frustum.

This tripartite representation eliminates textual cross-talk and provides the structured textual supervision for the caption ablation in Sec. 6.2.

D HelixBench and Efficiency Evaluation

D.1 Case Selection

Case organization. Each case carries a scene label and an interaction label, assigned by human annotators as content descriptions. Per-metric applicability is decided separately from the content labels, so label counts are not metric denominators; each metric reports on its applicable subset together with the number of scored and non-scorable cases. The benchmark contains 1,015 cases; DeSync applies to the 165 Event cases, and the spatial score to the Perspective subset.

Selection pipeline. Sources are isolated from the training corpus at the whole-video level; a source that matches training data cannot re-enter through a different time window. Candidate windows pass technical gates: frame continuity over the full clip (scene cuts, black or frozen frames, montage effects), audio continuity and audibility (missing audio, long or edge silences, suspected fades via

RMS-envelope checks), and a true-stereo gate (side/mid energy ratio ≤ -45 dB rejected; weak-stereo cases routed to review), since duplicated mono channels carry no spatial information.

Spatial subset pipeline. The spatial subset is built by scanning full source videos: stereo energy and direction statistics are computed in 0.1-second windows, candidate anchors are proposed from audio events and visual changes under a per-interval budget, and windows of about 5 seconds are cut around anchors. A case is admitted only if a single dominant sound source is visible and localizable; the source is bound with open-vocabulary detection using visual evidence alone, so audio direction never influences which object is selected. Background music, rain, waves, and multi-source scenes are excluded; a reference alignment threshold selects candidates for human review, and overlapping windows are removed.

D.2 Audio Quality

KL divergence. Generated audio is compared with the reference audio of the corresponding case, using the first 5 seconds of each. PaSST (Koutini et al., 2022) produces 527 audio-event logits per clip; applying softmax gives reference and generated distributions p_i and q_i.

$$\mathrm{KL} = \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{527} p_{i,c}\log\frac{p_{i,c}}{q_{i,c}}, \qquad N = 1{,}015. \tag{10}$$

The direction is reference-to-generated, $D_{\mathrm{KL}}(p_i\|q_i)$, with natural logarithms and equal averaging over clips. Lower values indicate closer agreement with the reference audio-event distributions.

FAD. Fréchet Audio Distance (Kilgour et al., 2019) compares generated and reference audio distributions in the VGGish (Hershey et al., 2017) embedding space. From the first 5 seconds of each clip, five 128-dimensional window embeddings are extracted, pooling all 5,075 embeddings from the 1,015 clips in each condition. Let (μ_r, Σ_r) and (μ_g, Σ_g) denote the empirical means and covariances of the reference and generated embeddings:

$$\mathrm{FAD} = \|\mu_r - \mu_g\|_2^2 + \mathrm{tr}\Big(\Sigma_r + \Sigma_g - 2\big(\Sigma_r^{1/2}\Sigma_g\Sigma_r^{1/2}\big)^{1/2}\Big). \tag{11}$$

All models use the same reference audio set; FAD is computed once per condition from the pooled window embeddings, and lower values indicate closer feature distributions.

D.3 Cross-Modal Alignment

DeSync. The Synchformer-based evaluator (Iashin et al., 2024) predicts the audio–video offset for the first and last 4.8-second windows of each generated clip. For each window, the offset class with the highest predicted probability is taken; the absolute values of the two offsets are averaged within a clip, then equally over clips. The result is an estimated temporal misalignment in seconds; lower is better.

ImageBind. ImageBind (Girdhar et al., 2023) embeds generated video and audio in a shared semantic space. For unit-normalized video and audio embeddings v_i and a_i from the same generated clip, the score is the mean matched cosine similarity $\frac{1}{N}\sum_{i=1}^{N} v_i^\top a_i$ with $N = 1{,}015$; higher values indicate stronger semantic agreement.

CLAP. CLAP (Wu et al., 2023) measures cosine similarity between the generated audio embedding and the embedding of its audio-visual input caption. The standard text encoder is used with its native 77-token truncation, and clip scores are averaged equally over all 1,015 cases. Higher scores indicate stronger audio-caption semantic agreement.

D.4 Spatial Audio–Visual Agreement

Spatial three-region score. The score measures agreement between the horizontal position of a detected sound-source candidate in the generated video and the stereo direction of its paired generated audio.

Visual localization. The generated video is sampled at 10 Hz and Grounding DINO (Liu et al., 2024) is applied with object-category queries derived from the case's video and joint audio-visual captions. Candidate detections are associated across frames, and a target track is selected using detection confidence, temporal coverage, and visual motion; when track selection fails, the pipeline falls back to a dominant caption-matched detection in each frame. Target selection uses visual evidence alone. For the selected bounding box, the horizontal center x_t is normalized as $d_t = 2x_t/W - 1$, where W is the frame width.

Audio direction. Left- and right-channel energies are computed in 100-ms windows with a 50-ms hop, giving the stereo energy pan $p_t = (E_{R,t} - E_{L,t})/(E_{R,t} + E_{L,t})$ with numerical stabilization; the pan sequence is interpolated to the visual sampling timestamps. Energies are measured from the mixed audio track without source separation. For joint audio-visual models this is the model's native generated audio; for V2A baselines it is the replacement audio synthesized for the same video, so the fixed video provides identical visual detections across audio conditions.

Three-region rule. Visual direction is left for $d_t < -\tau_v$, right for $d_t > \tau_v$, and center otherwise, with $\tau_v = 0.2$; audio direction uses the same rule on p_t with $\tau_a = 0.1$, equality belonging to the center region. The visual center region therefore spans 40–60% of image width, and the audio boundary corresponds to a channel energy ratio of $(1 + \tau_a)/(1 - \tau_a) \approx 1.222$, or approximately 0.87 dB. These tolerances stabilize left/right labels against small visual displacements and channel-energy fluctuations while retaining sensitivity to lateral deviations.

Scoring and eligibility. For visually lateral samples, same-side audio receives $+1$, centered audio 0, and opposite-side audio -1; visual-center samples are excluded. Within each eligible clip, valid visual-side samples are averaged and multiplied by 100; clip scores are then averaged equally over clips with at least one valid visual-side sample. The score ranges from -100 to 100, and always-centered audio scores zero. A clip qualifies if it contains at least four consecutive valid 10 Hz samples; an invalid sample or timestamp gap breaks the run, and eligible clips contribute all valid samples. The score quantifies lateral audio–visual agreement over each model's detected targets and eligible clips.

D.5 Post-hoc Video-to-Audio Comparison

Matched-video protocol. The dubbing comparison uses PrismAudio (Liu et al., 2025a), ThinkSound (Liu et al., 2025b), and AudioX (Tian et al., 2026). All three are conditioned on the same generated videos and their original joint audio-visual captions. Their synthesized audio replaces the original track while preserving the video stream and frame timestamps; the native audio provides the matched reference condition. The protocol covers all 1,015 test cases, with the same video and caption used across the four audio conditions of each case; evaluation follows the metric definitions and applicable subsets above.

D.6 Spatial Threshold Sensitivity

We evaluate sensitivity to the visual and audio thresholds defining the three spatial regions. Tables 12 and 13 report all 20 combinations of four visual thresholds $\tau_v \in \{0.1, 0.2, 1/3, 0.4\}$ and five audio thresholds $\tau_a \in \{0.01, 0.05, 0.1, 0.2, 0.3\}$, reusing the same generated clips, visual detections, and audio-pan trajectories under the eligibility rule above. The world-model sweep uses 145 Perspective candidates, with each model contributing its eligible clips; widening the visual threshold reduces the contributing cohorts from at most 143 clips to 113. The fixed-video comparison shares visual detections and the eligibility mask across all four audio conditions, with 147 of the 150 candidates satisfying the consecutive-sample rule and identical contributing counts for every condition. HelixWorld ranks first among all four world models across the 20 settings, and its native audio outperforms all three V2A replacements throughout the grid.

D.7 Inference Efficiency

Hardware and measurement boundaries. All timing runs use a single NVIDIA H800 for generation and decoding. After two complete warm-up rollouts, we report the median of two timed runs. Timing excludes model loading, pre-rollout text/image preparation, and MP4 encoding, but includes in-loop geometry processing, CPU offload, decoding, and frame transfer to CPU.

Table 12: **World-model sensitivity to spatial thresholds.** Results use the 145-candidate Perspective set. Pre-trained denotes the bidirectional model after progressive condition injection. Bold scores mark the best model per setting; bold thresholds mark the main-table setting. Higher is better.

τ_v	τ_a	LTX-2.3	EchoWM	HelixWorld	Pre-trained
0.10	0.01	2.16	22.58	**48.69**	30.14
0.10	0.05	1.96	21.85	**46.52**	30.34
0.10	0.10	1.55	20.87	**42.15**	29.30
0.10	0.20	1.16	17.57	**31.65**	25.91
0.10	0.30	0.90	14.91	**22.94**	22.80
0.20	0.01	2.41	25.92	**51.53**	36.92
0.20	0.05	2.20	25.30	**49.88**	36.27
0.20	**0.10**	1.82	24.05	**46.41**	33.61
0.20	0.20	1.03	20.69	**35.07**	28.00
0.20	0.30	0.61	18.41	**25.73**	24.90
1/3	0.01	4.78	31.07	**59.85**	46.47
1/3	0.05	4.40	31.34	**58.09**	46.22
1/3	0.10	3.50	30.38	**54.98**	42.61
1/3	0.20	1.07	24.70	**42.94**	35.68
1/3	0.30	0.68	21.75	**31.52**	31.45
0.40	0.01	5.06	32.62	**62.53**	46.35
0.40	0.05	4.54	32.60	**61.56**	46.90
0.40	0.10	4.24	31.79	**58.46**	43.88
0.40	0.20	1.80	27.12	**46.35**	37.29
0.40	0.30	1.60	24.44	**34.53**	32.63

Table 13: **Native versus post-hoc audio across spatial thresholds.** All conditions use fixed HelixWorld videos from the 150-candidate set, sharing detections and eligibility. Bold scores mark the best audio condition; bold thresholds mark the main-table setting. Higher is better.

τ_v	τ_a	HelixWorld (native)	+ AudioX	+ ThinkSound	+ PrismAudio
0.10	0.01	**49.32**	−3.94	21.72	−9.94
0.10	0.05	**47.11**	−5.33	21.36	−8.81
0.10	0.10	**42.79**	−4.84	19.04	−6.69
0.10	0.20	**32.06**	−4.69	14.30	−4.12
0.10	0.30	**23.07**	−4.82	10.47	−2.23
0.20	0.01	**51.97**	−3.73	19.79	−11.07
0.20	0.05	**50.26**	−4.36	19.75	−9.99
0.20	**0.10**	**46.88**	−4.01	18.05	−8.04
0.20	0.20	**35.30**	−4.87	13.48	−4.82
0.20	0.30	**25.70**	−5.22	10.07	−2.44
1/3	0.01	**60.18**	−3.24	21.40	−13.81
1/3	0.05	**58.34**	−1.98	20.42	−11.89
1/3	0.10	**55.45**	−2.21	18.67	−10.15
1/3	0.20	**43.01**	−4.46	11.26	−6.28
1/3	0.30	**31.37**	−5.97	7.95	−4.01
0.40	0.01	**63.28**	−2.50	22.17	−13.41
0.40	0.05	**62.35**	−1.38	21.56	−11.42
0.40	0.10	**59.32**	−2.75	19.37	−10.39
0.40	0.20	**46.45**	−4.52	12.22	−7.32
0.40	0.30	**34.46**	−4.59	8.55	−5.09

Real-time factor. For streaming models, steady-state RTF measures the wall-clock time after the first decoded block divided by the duration of the remaining newly generated frames. Let T be total rollout wall time, t_1 the first-block delivery time, F the output frame count, b the exclusive frame index of the first delivered block, and f the output frame rate. Then

$$\mathrm{RTF}_{\mathrm{steady}} = \frac{T - t_1}{(F - b)/f}. \tag{12}$$

The denominator counts newly generated frames after the first delivered block over the completed rollout. Lower RTF is faster; RTF below one indicates faster-than-real-time throughput after the first block. For the non-streaming bidirectional reference, RTF uses the complete generation-and-decoding time divided by the full generated duration.

Output settings. All models are evaluated at 768×512 resolution, using the output frame rates reported in their official repositories.

E Additional Trajectory Distillation Results

E.1 Conditional Visual Diversity

Inputs and features. Our diversity calculation follows *Diversity-Preserved Distribution Matching Distillation for Fast Visual Synthesis* (Wu et al., 2026), adapting its per-prompt image comparison to matched video frames. We extract DINOv3 ViT-L/16 pooled features (Siméoni et al., 2025) and CLIP ViT-L/14 CLS features (Radford et al., 2021) at matched times across generations of the same input, then normalize each feature vector.

Aggregation. Let $\boldsymbol{f}_{i,t,s}$ be a unit-normalized feature for input i, time $t \in \mathcal{T}$, and generation $s \in \{1, \ldots, S\}$. For each input we average cosine distance over generation pairs and times:

$$d_i = \frac{1}{|\mathcal{T}|\binom{S}{2}} \sum_{t\in\mathcal{T}} \sum_{1\le s<s'\le S} \left(1 - \boldsymbol{f}_{i,t,s}^{\top} \boldsymbol{f}_{i,t,s'}\right). \tag{13}$$

The fixed first frame is excluded. If $\mathcal{G}$ is the set of sources and $\mathcal{I}_g$ contains the inputs from source g, the reported score is

$$D = \frac{1}{|\mathcal{G}|} \sum_{g\in\mathcal{G}} \frac{1}{|\mathcal{I}_g|} \sum_{i\in\mathcal{I}_g} d_i. \tag{14}$$

Each source receives equal weight, and higher scores indicate greater conditional visual diversity.

Results. Self-forcing with online trajectory distillation loss increases DINOv3 diversity by 21.7% and CLIP diversity by 13.5% over DMD-only self-forcing (Table 8).

E.2 Additional Qualitative Comparisons

Figure 8 shows further examples of reduced visual drift with trajectory distillation.

F User Study Protocol

Study design. Model identities are hidden, and A/B placement is randomized and balanced. The study presents 20 audio-enabled comparisons followed by 20 muted-video comparisons. With audio, we compare HelixWorld with EchoWM and three post-hoc dubbing baselines that apply AudioX, ThinkSound, or PrismAudio to the same HelixWorld video. The muted-video comparison includes Alaya, HY-WorldPlay 1.5, LingBot v2, SANA, and EchoWM. We analyze 212 audio-enabled and 190 muted-video responses, covering 20 scenes in each setting.

Evaluation criteria. The audio-enabled study asks about audio–visual synchrony, audio quality, spatial acoustic realism, and overall preference. The muted-video study asks about visual quality, action following, scene following, temporal consistency, and overall preference.

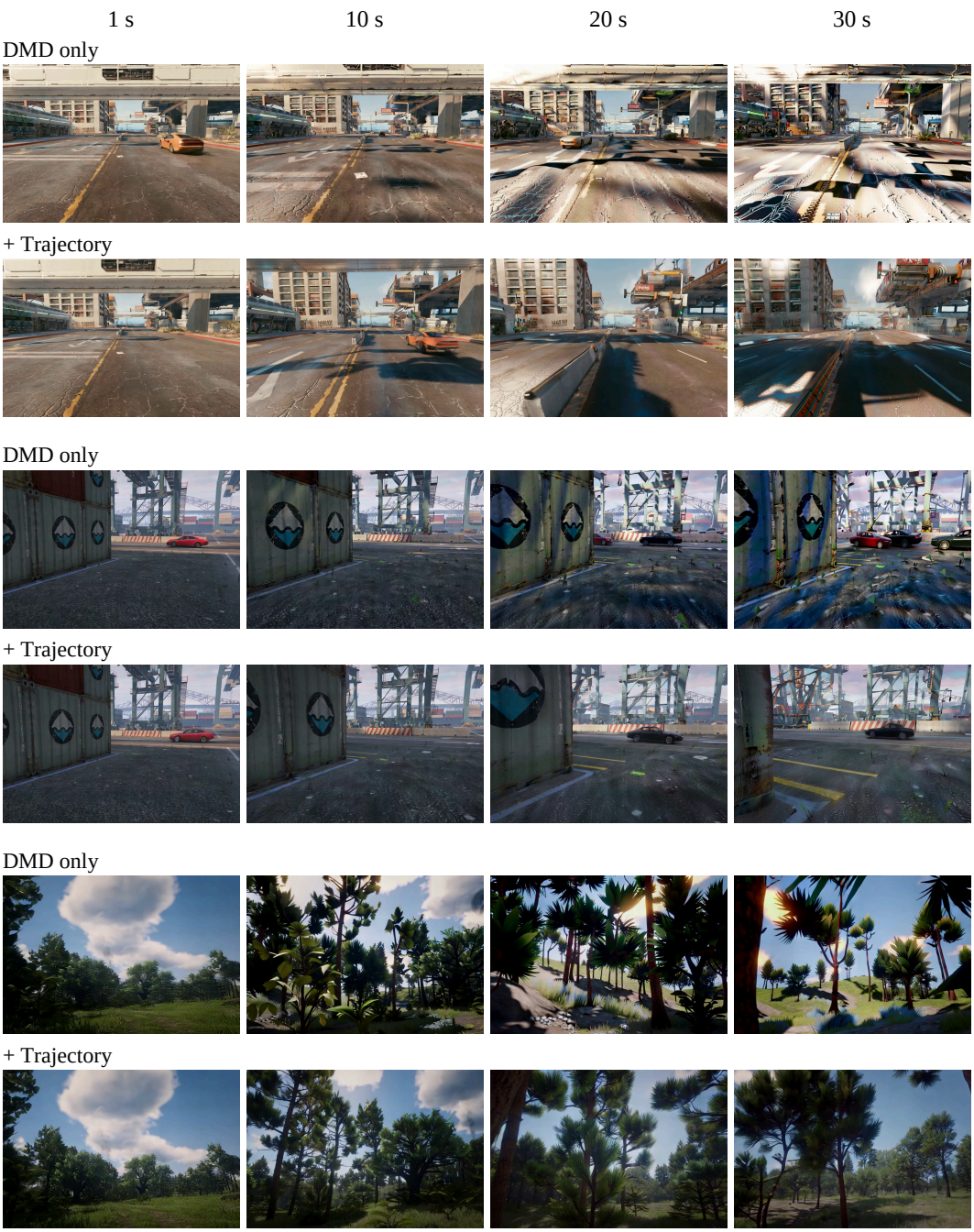


Figure 8: **Additional visual comparisons of trajectory distillation.** Each pair shows DMD only (top) and with trajectory distillation (bottom).

Preference aggregation. Responses are summarized as a preference for HelixWorld, a tie, or a preference for the baseline. For each model pair, a category's share is its response count divided by the total number of responses, with ties retained in the denominator. Figure 5 reports overall preference for each pair. Win rates reported in the main text exclude ties.

G Additional Audio-Visual Comparisons

Figures 9 and 10 illustrate audio–visual synchronization; Figures 11 and 12 show stereo correspondence. AudioX, ThinkSound, and PrismAudio use the same HelixWorld video, while EchoWM is shown with its own generated video and audio. Mel power is normalized per recording, using a shared reference for the left and right channels.

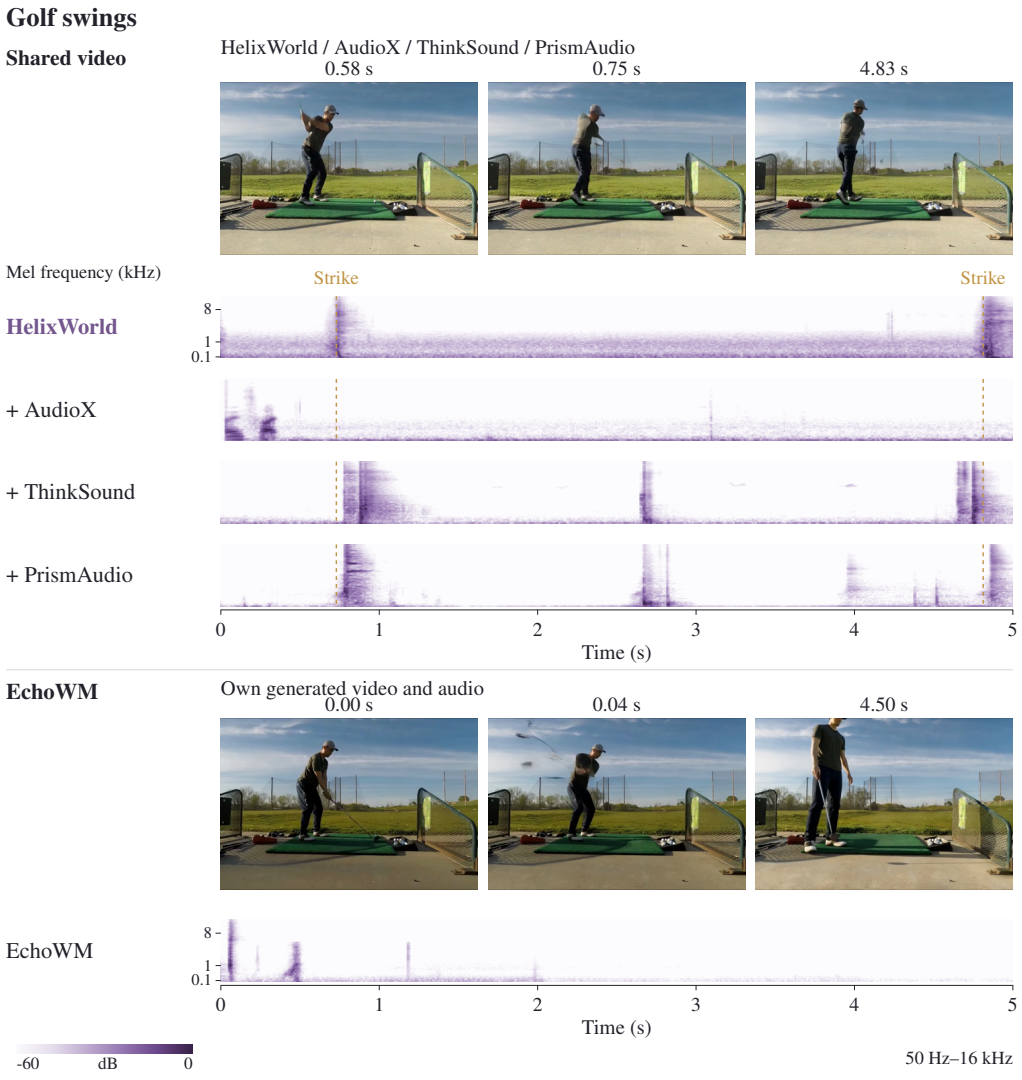


Figure 9: **Audio–visual synchronization during golf swings.** Dashed lines mark the visible strikes in the shared video. HelixWorld's acoustic transients coincide with both events.

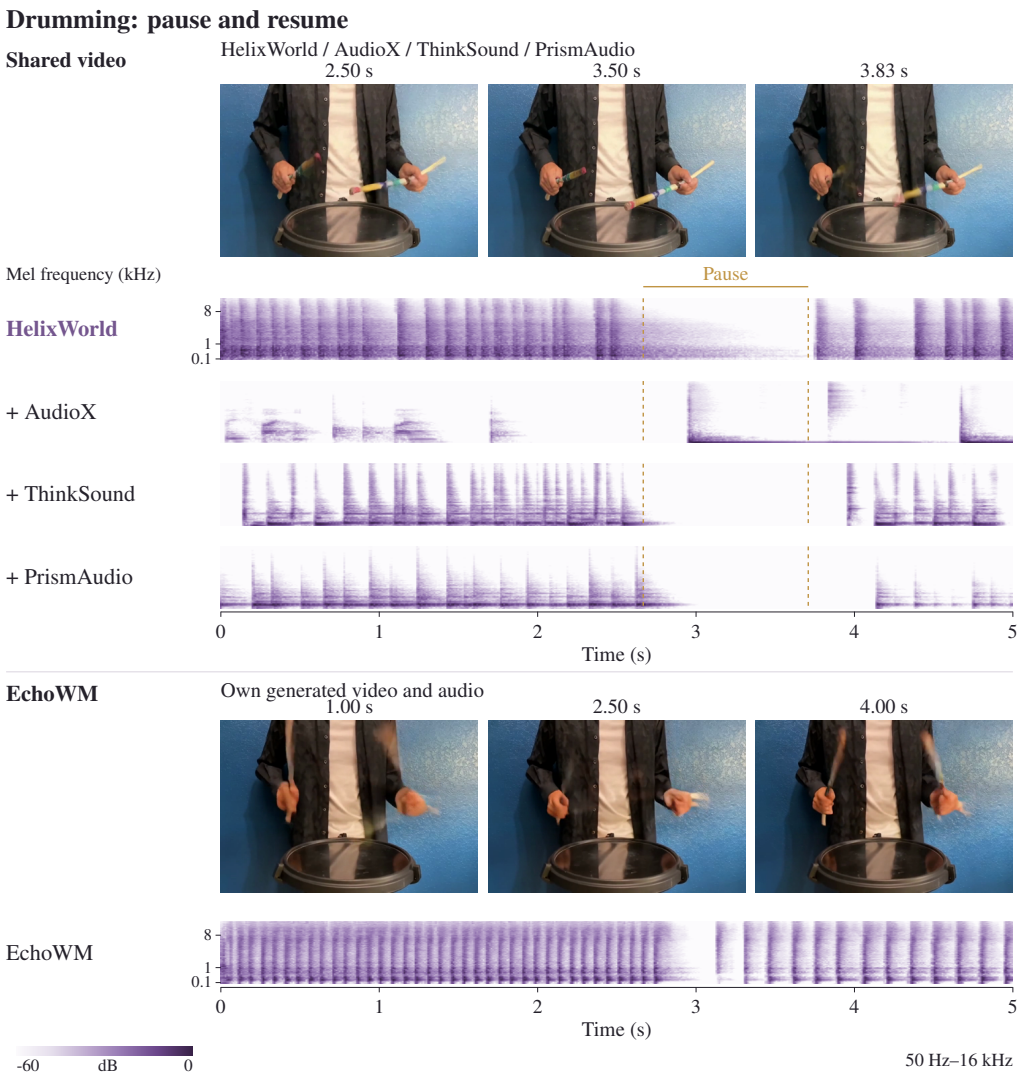


Figure 10: **Audio–visual synchronization during drumming.** The marked interval is a visible pause. HelixWorld's transients subside during the pause and resume with the motion; AudioX produces a transient during the pause.

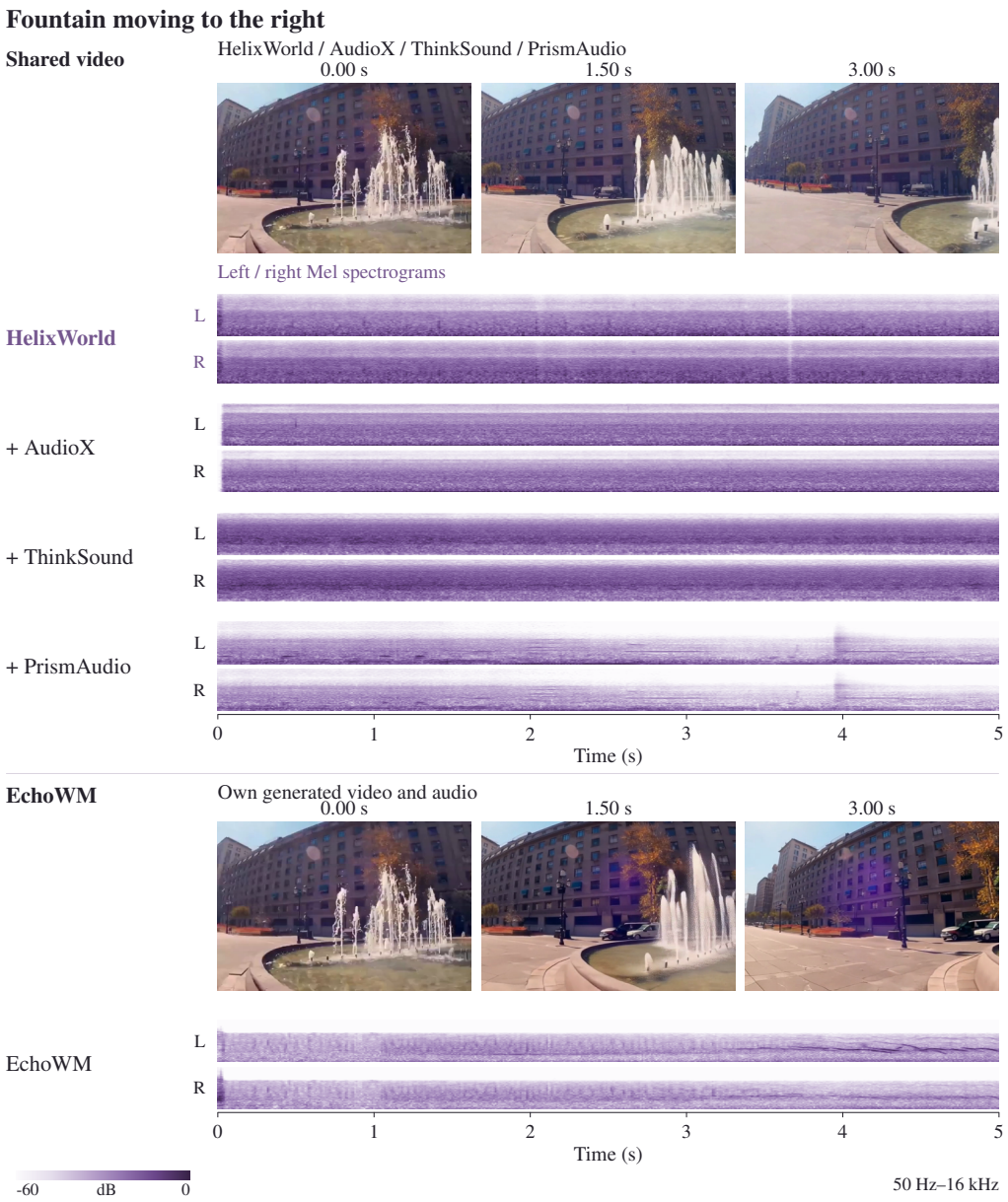


Figure 11: **Stereo audio as the fountain moves rightward.** L and R denote the left and right channels. HelixWorld maintains stronger right-channel energy, consistent with the fountain's position in the video.

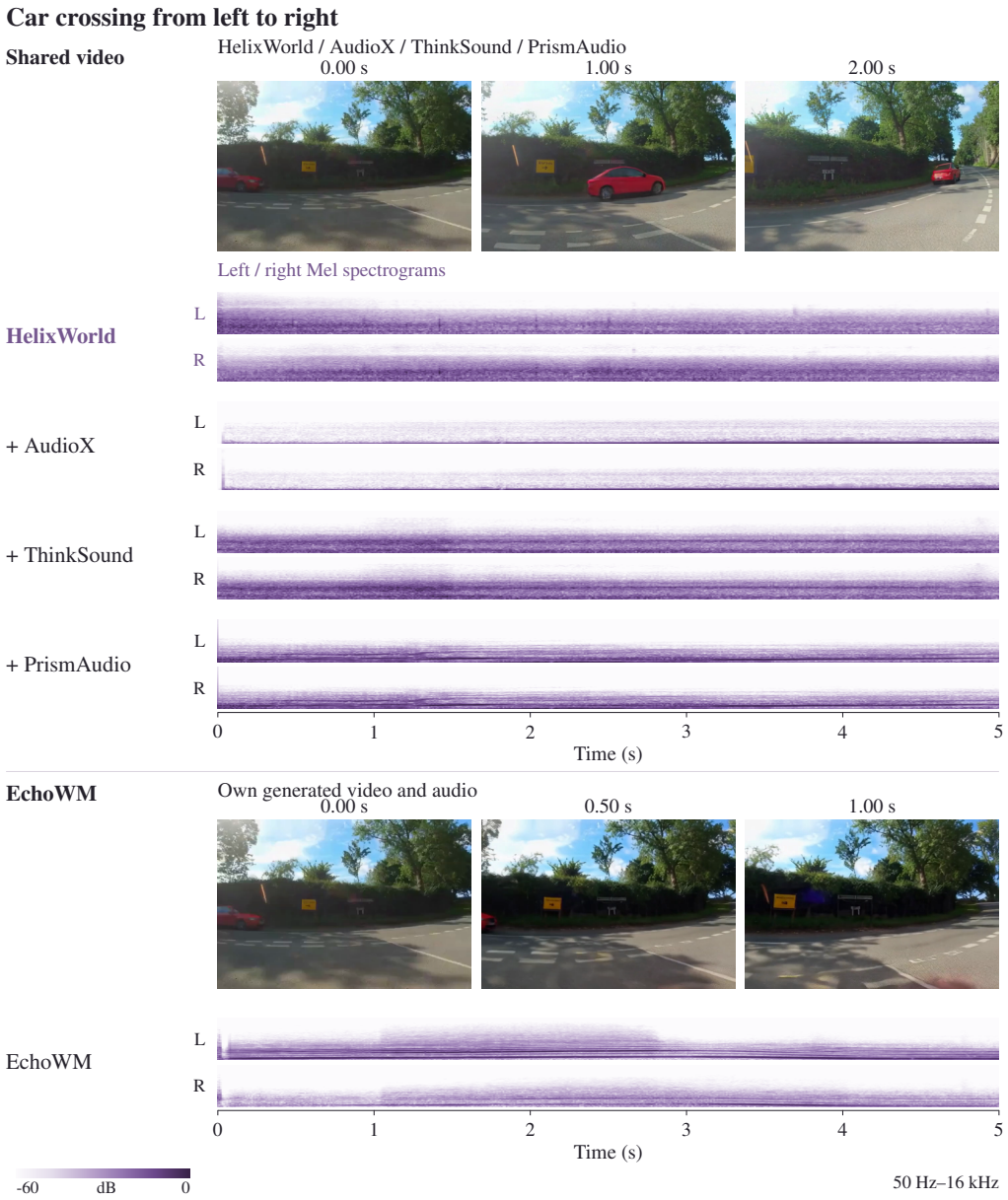


Figure 12: **Stereo audio during a left-to-right vehicle pass.** HelixWorld's channel balance shifts from left to right with the car's motion. L and R denote the two audio channels.