ISAUDIO: INTERACTIVE SPATIAL AUDIO GENERATION FOR WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial audio is essential for world models to simulate immersive interactions with action-responsive auditory feedback. Unlike video-to-audio synthesis from fully observed clips, this interaction loop requires streaming spatial audio that responds to listener actions and predicts upcoming sounds before the corresponding visual observations arrive. In this work, we introduce ISAUDIO, an action-conditioned framework for streaming binaural audio prediction. To address data scarcity, we construct spatial audiovisual data from panoramic recordings with controlled listener trajectories and first-person gameplay with aligned action signals. For efficient streaming, ISAUDIO combines autoregressive flow matching in a spatial latent space with causal consistency distillation, enabling few-step prediction conditioned only on observed visual prefixes, audio history, and current actions. To preserve semantic, temporal, and spatial consistency, we propose Contrastive Spatial Audio-Video Pretraining (CSAVP), which learns representations that capture changes in listener viewpoint through same-scene trajectory contrasts. We further refine the distilled generator with dual-granularity Group Relative Policy Optimization (GRPO), combining clip-level perceptual rewards with chunk-local event and spatial feedback. Experiments show that ISAUDIO achieves an S-CLAP score of 0.787, compared with 0.664 for the strongest baseline, and enables real-time streaming inference with a real-time factor of 0.53. Together, these results highlight interactive spatial audio generation as an important step toward richer multimodal world models. Demos and code are available at `https://isaudio.github.io`.

1 INTRODUCTION

Simulating authentic, dynamic interactions between humans and the environments lies at the core of world models (Liu et al., 2026b; Qian et al., 2026; Hu et al., 2026). Recent Vision-Language-Action models and embodied agents have achieved remarkable progress in simulating visual dynamics (Hafner et al., 2019; Bruce et al., 2024; Valevski et al., 2025; Chen et al., 2026). However, authentic real-world interaction is inherently multimodal, where accurate perception of auditory signals is equally crucial. Auditory cues, such as footsteps, background music, and off-screen collisions, reveal subtle events and spatial relationships both within and beyond the current field of view, providing indispensable guidance for navigation and tactical decision-making. Integrating binaural spatial audio signals into simulations to achieve rich, dynamic audiovisual feedback represents a crucial frontier for the next-generation foundation models.

In recent years, video-to-audio (V2A) synthesis frameworks (Luo et al., 2023; Cheng et al., 2025; Liu et al., 2025; Zhang et al., 2026) have significantly improved semantic realism and temporal alignment for silent video clips. Building upon these advances, spatial audio methods (Dagli et al., 2025; Zhao et al., 2025; Zhang et al., 2025) further model the acoustic properties of multi-channel sound, aligning auditory direction and perspective with the visual scene to deliver a more immersive experience. These methodologies provide the groundwork for interactive spatial audio generation, which predicts binaural sound grounded in first-person observations, listener actions, and available scene states. While standard video-conditioned models rely on fully observed target clips, interactive generation must predict upcoming audio intervals before corresponding visual observations arrive by dynamically responding to ongoing listener controls.

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

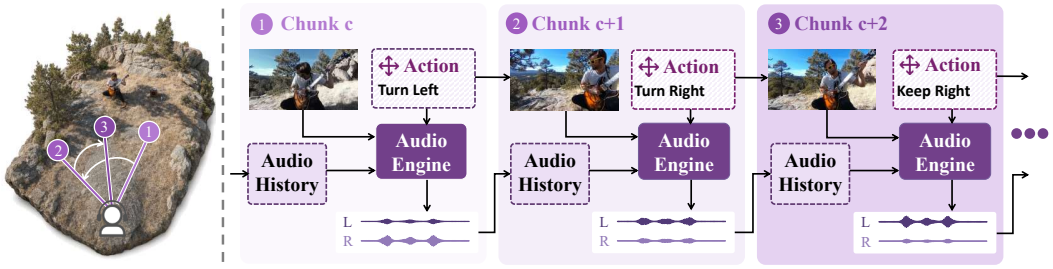

Figure 1: **Interactive spatial audio generation.** Left: a listener turns within a fixed acoustic scene. Right: at each chunk start, ISAUDIO uses the observed visual prefix, audio history, and current action to predict the next binaural chunk. Predictions become fixed history for subsequent steps, with new observations incorporated only at the next chunk boundary.

Achieving true interactive spatial audio generation presents three major challenges. First is the data bottleneck: precisely aligned spatial audiovisual recordings paired with detailed action controls, listener states, and acoustic annotations are exceptionally scarce and difficult to collect, severely restricting in-depth exploration in this domain. Second is interactive streaming efficiency: the open-ended and long-horizon nature of interactive environments demands that the generator continuously perform predictions in a streaming fashion with minimal latency, ensuring seamless synchronization with real-time user inputs. Third is causal prediction fidelity: causal generation must simultaneously preserve semantic, temporal, and spatial consistency. The synthesized binaural sound must plausibly reflect ongoing acoustic events and their temporal evolution, while remaining strictly aligned with the listener's dynamic orientation and spatial relationships within the scene.

To address these challenges, we introduce ISAudio, an action-conditioned framework for streaming binaural audio generation in world models. First, to overcome the data bottleneck, we establish the ISAudio dataset , which comprises ViewSpatial derived from panoramic videos and First-Order Ambisonic (FOA) recordings along controlled listener trajectories, alongside InteractSpatial constructed from first-person gameplay with aligned player controls and multi-actor states. To achieve real-time interactive generation, ISAudio employs an autoregressive flow-matching generator operating in a binaural latent space, and integrates causal consistency distillation to reduce sampling overhead for efficient, few-step streaming inference. To ensure causal prediction fidelity across spatial, temporal, and semantic dimensions, we propose Contrastive Spatial Audio-Video Pretraining (CSAVP), which learns viewpoint-dependent audiovisual representations through same-scene trajectory contrasts. Its causal encoder provides spatially grounded visual conditioning during streaming generation, while its bidirectional counterpart enables comprehensive audiovisual evaluation. Finally, dual-granularity GRPO refines the distilled generator by combining clip-level perceptual rewards with chunk-local event and spatial feedback. We evaluate video-conditioned spatial audio generation and action-conditioned interactive prediction in terms of acoustic quality, temporal alignment, spatial accuracy, and generation efficiency. Our main contributions are summarized as follows:

- We introduce ISAudio, an action-conditioned streaming binaural audio framework for world models that combines autoregressive flow matching with causal consistency distillation for low-latency, few-step generation.
- We propose CSAVP to learn viewpoint-dependent audiovisual representations through same-scene trajectory contrasts, utilizing its causal encoder for streaming spatial conditioning and its bidirectional counterpart for evaluation.
- We develop dual-granularity GRPO to refine the generator by coupling clip-level perceptual rewards with chunk-local event and spatial feedback.
- We construct the ViewSpatial and InteractSpatial datasets alongside dedicated spatial evaluation protocols, establishing a standardized benchmark for video-conditioned and interactive streaming spatial audio generation.

2 Related Work

Video-to-Audio Generation. Video-to-audio (V2A) synthesis learns sound content and event timing from silent video. Diff-Foley (Luo et al., 2023) and FoleyCrafter (Zhang et al., 2026) emphasize audiovisual alignment, while Seeing-and-Hearing (Xing et al., 2024), MMAudio (Cheng et al., 2025), and Movie Gen (Polyak et al., 2024) develop broader multimodal generation capabilities. ThinkSound (Liu et al., 2025) and PrismAudio (Liu et al., 2026a) further incorporate reasoning or reward-based optimization. Whereas these methods typically condition on the target video interval, ISAudio predicts its sound from past observations and current controls.

Spatial Audio Generation. SoundSpaces (Chen et al., 2020) establishes acoustic simulation as a tool for embodied perception. SEE-2-SOUND (Dagli et al., 2025), FoleySpace (Zhao et al., 2025), and SpatialV2A (Wang et al., 2026) explore visually grounded spatial synthesis. BinauralFlow Liang et al. (2025) uses causal flow matching for streaming geometry-conditioned speech rendering. Particularly close to our output space, ViSAudio (Zhang et al., 2025) generates binaural audio directly from video using conditional flow matching and models audiovisual spatial-temporal relationships. Sonic4D Xie et al. (2026) combines monocular 4D reconstruction, monaural audio generation, sound-source tracking, and physics-based spatial rendering for viewpoint-adaptive scene exploration. ISAudio learns listener-relative audiovisual correspondence through same-scene trajectory contrasts to guide binaural generation.

World Models and Interactive Generation. Dreamer (Hafner et al., 2019), Genie (Bruce et al., 2024), and GameNGen (Valevski et al., 2025) learn action-dependent environment dynamics. Li & Torralba (2025) extend action-conditioned video prediction to support fine-grained multimodal controls. AV-CDiT in Audio-Visual World Models (Wang et al., 2025) autoregressively predicts future visual frames, audio, and task rewards from audiovisual history and navigation actions, introducing AVW-4k and downstream navigation evaluation. ISAudio develops the auditory component of interactive world models, conditioning sound on listener controls and available multi-actor state.

3 Interactive Spatial Audio Data

Interactive spatial audio prediction requires synchronized visual observations, binaural sound, and available listener controls and scene states. We construct two complementary datasets, **ViewSpatial** and **InteractSpatial**, combining controlled viewpoint changes in real-world recordings with player-driven motion and sound-producing actions in gameplay (Table 1).

ViewSpatial. We derive ViewSpatial from synchronized panoramic video and first-order Ambisonic (FOA) recordings, varying listener orientation at the recording position. A shared listener trajectory controls both the field-of-view crop and sound-field rotation, with head-related transfer functions decoding the rotated audio into aligned binaural channels. Fixed directions yield static pairs, while smooth turns yield dynamic pairs with timestamped listener orientation. Ordinary field-of-view recordings contribute additional fixed-view pairs. Trajectories from the same scene and time interval share the underlying acoustic events while changing listener-relative spatial cues, providing controlled within-scene contrasts for CSAVP. Segment-level filtering removes silent, visually static, or poorly aligned content while preserving the retained segments' grouped views and trajectories.

InteractSpatial. Using Counter-Strike 2 demo replays, we pair first-person video and the original binaural game mix with timestamped listener controls and states.[1] The recordings capture translation, turning, and event-triggering actions, together with available actors' positions, actions, and event states. These synchronized streams describe changes in both the listener's perspective and scene activity, including sound sources outside the field of view. Structured game records additionally provide event times and listener-relative directional labels for supervision and evaluation.

Training and evaluation are split by original video for ViewSpatial and by match for InteractSpatial, keeping all derived windows, views, and trajectories from each source in the same partition. Further construction and split details are provided in Appendix B.

[1] https://github.com/vast7777/cs2-demo-to-dataset.

Table 1: Comparison of datasets for visual prediction and spatial audio generation. Counts: train/eval clips; traj.: trajectories.

Dataset	Samples	Video	Audio	Format	Action	State
RoboNet (TFDS)	162,417 traj.	✓	—	—	Robot control	Robot state
Sphere360	100,533 / 3,063	✓	✓	4-ch FOA	—	—
BiAudio	94,845 / 2,695	✓	✓	2-ch binaural	Rotation	Orientation
ViewSpatial	239,892 / 3,658	✓	✓	2-ch binaural	Rotation	Orientation
InteractSpatial	61,253 / 5,144	✓	✓	2-ch binaural	Player controls	Listener + actors

4 METHOD

As shown in Figure 2, ISAUDIO generates binaural audio from first-person observations, listener controls, and available actor states. Chunk-causal flow matching predicts upcoming sound before future observations arrive (Section 4.1), while CSAVP supplies listener-relative spatial features through same-scene trajectory contrasts (Section 4.2). We then reduce sampling cost through causal consistency distillation (Section 4.3). Finally, dual-granularity GRPO refines the few-step generator with clip-level perceptual feedback and chunk-local event and spatial rewards (Section 4.4). Architecture details for the generator and CSAVP are provided in Appendix A.

4.1 ACTION-CONDITIONED CHUNK-CAUSAL GENERATION

Task formulation. Let V denote timestamped video, A the listener's control/state stream, S the available multi-actor state, and $Y = (Y_L, Y_R)$ the binaural waveform. A frozen stereo autoencoder (Evans et al., 2026) maps Y to latent chunks $Z^{1:C}$, with chunk c covering the interval $[s_c, s_{c+1})$. Conditioning each prediction on information available at its start gives

$$p_\theta(Z^{1:C} \mid V, A, S) = \prod_{c=1}^{C} p_\theta\big(Z^c \mid Z^{<c}, V_{\le s_c}, A_{\le s_c}, S_{\le s_c}\big)\,. \tag{1}$$

The decoder reconstructs the binaural waveform as $\hat{Y} = D_{\text{vae}}(\hat{Z}^{1:C})$. All visual, control, and state inputs obey the cutoff s_c, with conditions updated only at the next chunk boundary. Bidirectional self-attention within a chunk jointly models its acoustic structure, while causal attention across chunks retains audio history without exposing future chunks.

Conditional flow matching. Flow matching learns a velocity field that transports Gaussian noise to the audio latent distribution (Lipman et al., 2022). To accommodate varying prediction horizons, we sample chunk boundaries at valid latent positions with probability p_b. Each chunk receives its own flow time t_c, allowing training to cover history and current chunks across different noise levels. For a clean chunk Z^c and independent noise $\epsilon^c \sim \mathcal{N}(0, I)$, we construct the linear interpolation $Z^c_{t_c} = (1 - t_c)\epsilon^c + t_c Z^c$ with target velocity $u^c = Z^c - \epsilon^c$. Here, the flow time $t_c \in [0, 1]$ indexes the noise level. Let $Z_{\boldsymbol{t}}$ denote the sequence of noised chunks and $\boldsymbol{\mathcal{C}} = (\mathcal{C}^1, \ldots, \mathcal{C}^C)$ represent their corresponding chunk-start conditions. We train the masked velocity field using the objective:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}\left[\frac{1}{C}\sum_{c=1}^{C} \|v_\theta(Z_{\boldsymbol{t}}, \boldsymbol{t}, \boldsymbol{\mathcal{C}})_c - u^c\|_2^2\right], \tag{2}$$

where c indexes the chunk ID, and temporal attention masks enforce the information boundary in Eq. 1 while enabling parallel training across all chunks. At inference, we integrate each chunk sequentially from noise to data, retaining previously generated latents as fixed history.

Conditioning mechanisms. Causal CSAVP complements CLIP (Radford et al., 2021) and DINO (Siméoni et al., 2025) with fine-grained audiovisual spatial correspondence. Discrete listener controls enter cross-attention as timestamped tokens, while continuous state is embedded and concatenated with the noisy audio latents. A shared encoder embeds available actors' listener-relative positions, actions, and event states, representing potential sound sources both within and outside the field of view. The actor branch is masked when state information is unavailable.

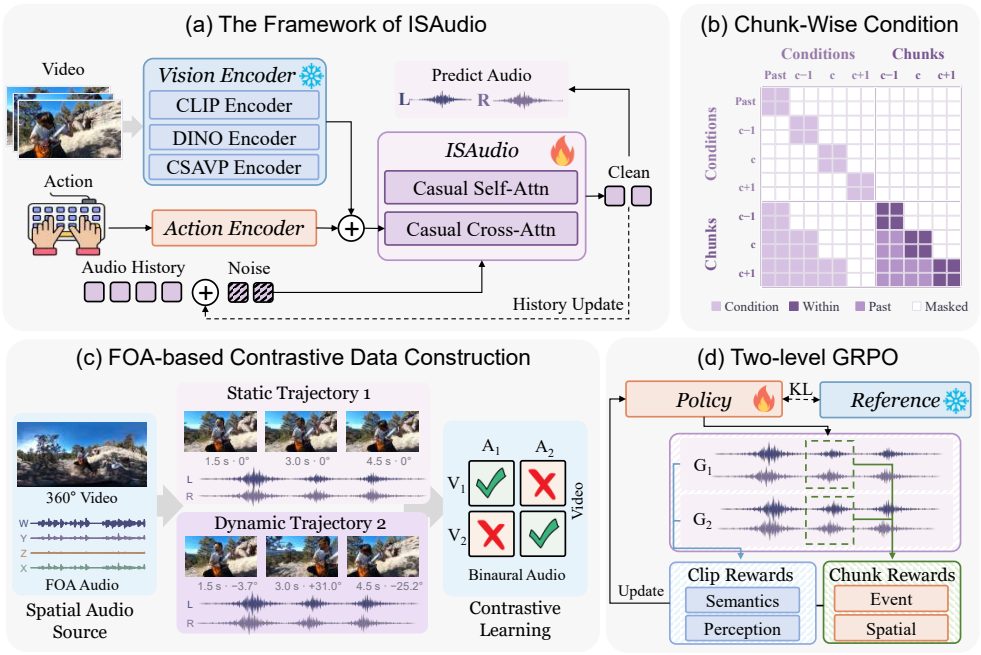

Figure 2: **ISAUDIO overview.** (a) The autoregressive flow generator predicts binaural audio chunks from available visual observations, listener controls, and generated audio history. (b) Chunk-wise attention restricts conditioning to the prefix ending at chunk start s_c, while allowing audio attention within the current chunk and to preceding chunks. (c) Coupled video and binaural rendering constructs same-scene trajectory contrasts for CSAVP. (d) Dual-granularity GRPO combines clip-level semantic and quality rewards with chunk-local event and spatial feedback.

4.2 SPATIAL AUDIO-VIDEO PRETRAINING

Contrastive spatial correspondence. Previous audio-visual contrastive learning (Luo et al., 2023) can capture the semantic and temporal correspondence between video and audio, but rarely learns their spatial correspondence. CSAVP instead contrasts listener trajectories within the same scene, making viewpoint-dependent audiovisual correspondence the discriminative cue. For each physical scene segment, we pair field-of-view video with its binaural rendering along multiple static or dynamic listener trajectories. These trajectories share the same underlying events and temporal window but alter the sources' listener-relative directions. Matching trajectories form positive audiovisual pairs, while alternative trajectories from the same segment provide spatially mismatched hard negatives. The visual encoder captures viewpoint-dependent scene structure, while the audio encoder combines channel-specific semantic features with binaural spatial features. For M trajectories, let $e_V^{(i)}$ and $e_A^{(i)}$ be their unit-normalized embeddings, and define the similarity matrix $\mathbf{S}_{ij} = \langle e_V^{(i)}, e_A^{(j)} \rangle / \tau$ with temperature τ. Symmetric trajectory matching minimizes

$$\ell_{\text{traj}}(\mathbf{S}) = -\frac{1}{2M} \sum_{i=1}^{M} \left[\log \frac{\exp \mathbf{S}_{ii}}{\sum_j \exp \mathbf{S}_{ij}} + \log \frac{\exp \mathbf{S}_{ii}}{\sum_j \exp \mathbf{S}_{ji}} \right]. \tag{3}$$

Long-context contrastive learning. Listener motion makes spatial correspondence a temporal property: a clip-level match alone does not specify when the acoustic perspective changes. We first learn short-window correspondence using trajectory matching, cross-content contrast, and binaural perturbations, then extend the visual encoder to longer temporal context while freezing the audio

representation. For a clip divided into K audiovisual intervals, the temporal objective is

$$\mathcal{L}_{\text{CSAVP}} = \frac{1}{K}\sum_{k=1}^{K} \ell_{\text{traj}}(\mathbf{S}^k) + \ell_{\text{traj}}(\mathbf{S}^{\text{clip}}) + \mathcal{L}_{\text{cross}}^{\text{clip}}. \tag{4}$$

Here $\mathbf{S}^k$ and $\mathbf{S}^{\text{clip}}$ compare trajectories at the interval and clip levels, respectively; clip-level audio embeddings are obtained by normalizing the mean of interval embeddings. The local term preserves correspondence as the listener moves, while the clip term captures the trajectory's overall spatial evolution. The cross-content term preserves semantic discrimination across scenes, excluding off-diagonal same-source pairs from its negatives.

Causal adaptation. Causality must hold within the visual encoder as well as the generator, since a cross-attention mask cannot remove future information already encoded in its inputs. We adapt the visual pathway with causal temporal attention and left-padded temporal convolutions. Training with the same contrastive objective makes each feature summarize only the observed prefix while retaining spatial correspondence. The causal encoder supplies generation conditions, whereas the frozen bidirectional model evaluates audiovisual alignment over completed outputs.

4.3 Causal Consistency Distillation

Causal prediction controls what the model can observe, but repeated flow evaluations still determine how quickly each chunk becomes available. We therefore distill the supervised flow model into a few-step student by aligning endpoint predictions along the same teacher trajectory. For chunk c, let h^c collect the fixed audio history and conditions available at s_c, and write t for its flow time t_c. The student predicts a clean endpoint as $G_\phi(z, t; h^c) = z + (1-t)v_\phi(z, t; h^c)$. A frozen teacher takes one Euler step to an adjacent time $t' > t$, giving $\widetilde{Z}_{t'}^c = Z_t^c + (t'-t)v_{\theta_T}(Z_t^c, t; h^c)$. We align the student's endpoint with that of an exponential-moving-average (EMA) student evaluated at the teacher-advanced state:

$$\mathcal{L}_{\text{CD}} = \mathbb{E}\left[\left\| G_\phi(Z_t^c, t; h^c) - \text{sg}\left[G_{\bar{\phi}}(\widetilde{Z}_{t'}^c, t'; h^c)\right]\right\|_2^2\right], \tag{5}$$

where $\bar{\phi}$ denotes the EMA parameters and sg stops gradients. All branches share the same audio history and chunk-start conditions, preserving the causal prediction boundary during distillation.

4.4 Dual-Granularity GRPO

Interactive generation requires coherent audio and locally accurate events under changing actions and viewpoints, which traditional clip-level rewards fail to evaluate. To address this, we optimize the distilled generator using GRPO, combining clip-level perceptual guidance with chunk-local event and spatial feedback across G sampled candidate audio streams per input.

Multi-granularity rewards. At the clip level, semantic reward r_{sem}^g measures LAION-CLAP (Wu* et al., 2023) similarity, while quality reward r_{qual}^g averages Audiobox-Aesthetics (Tjandra et al., 2025) scores across content enjoyment, usefulness, and production quality. At the chunk level, event fidelity uses PE-A-Frame (Bolya et al., 2025) to calculate temporal intersection-over-union for annotated events alongside an occupancy penalty for hallucinated sounds. Spatial fidelity uses the CSAVP audio encoder to compute local cosine similarity between generated and reference audio, with reference-active windows mapped to generation chunks by temporal overlap.

Local credit and group-relative advantage. Normalized local contributions $q_{\text{evt}}^{g,c}$ and $q_{\text{sp}}^{g,c}$ are combined as $q^{g,c} = w_{\text{evt}}q_{\text{evt}}^{g,c} + w_{\text{sp}}q_{\text{sp}}^{g,c}$ and reassigned across temporal overlaps to yield $\widetilde{q}^{g,c}$. Because each generated chunk serves as context for subsequent steps, the return $Q^{g,c} = R_{\text{clip}}^g + \sum_{j=c}^{C}\widetilde{q}^{g,j}$ accumulates future local rewards. The normalized advantage $\hat{A}^{g,c} = (Q^{g,c} - \bar{Q}^c)/(s_{\text{res}} + \delta)$ compares candidates at matched temporal positions, where $\bar{Q}^c$ is the candidate mean, s_{res} is the root-mean-square centered residual, and $\delta > 0$ stabilizes calculation.

Optimizing the distilled policy. We initialize both the policy and its frozen reference from the distilled EMA model to refine endpoint-based transitions. Given transition likelihood ratio $\rho_\theta = \pi_\theta/\pi_{\text{old}}$, post-training minimizes:

$$\begin{aligned}\mathcal{L}_{\text{post}} = &- \mathbb{E}\Big[\min\Big(\rho_\theta \hat{A}^{g,c},\ \text{clip}(\rho_\theta, 1-\varepsilon, 1+\varepsilon)\hat{A}^{g,c}\Big)\Big] \\ &+ \beta\, \mathbb{E}[D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})]\,,\end{aligned} \tag{6}$$

where threshold ε clips policy updates, and the β-weighted KL term prevents divergence from the distilled reference. Implementation details are provided in Appendix C.1.

5 Experiments

5.1 Experimental Setup

Dataset and Implementation Details. ViewSpatial pairs real-world panoramic video with FOA-derived binaural audio under static and rotating viewpoints, while InteractSpatial pairs gameplay with binaural audio, listener controls, and actor states. ViewSpatial contains 239,892 training and 3,658 evaluation samples, while InteractSpatial contains 61,253 training and 5,144 evaluation samples. Stage1 pretrains video-conditioned generation on both datasets for 100k updates; Stage2 introduces listener controls and actor states for 50k updates of interactive SFT. Using the Stage2 teacher, causal consistency distillation produces a four-step student in 10k updates. GRPO then refines the student on InteractSpatial for 1k updates. More details are provided in Appendices B and C.1.

Evaluation axes. We evaluate semantic fidelity using MS-CLAP (Elizalde et al., 2023), temporal fidelity (DeSync (Iashin et al., 2024), event IoU, and FPO), aesthetic quality (PQ, CE, CU), and distributional quality (FD, KL). Spatial correspondence is assessed using Local C-sim, which averages temporally aligned CSAVP audiovisual similarities, and Spatial-CLAP (Seki et al., 2025) (S-CLAP). Mean Opinion Scores for Audio Quality (MOS-Q) and Audiovisual Consistency (MOS-C) evaluate subjective overall quality and audiovisual consistency across event, temporal, and spatial dimensions. On the InteractSpatial dataset, IoU quantifies event overlap, whereas false-positive occupancy (FPO) penalizes predicted activity in non-event windows. Generation efficiency is measured by complete-segment RTF, defined as end-to-end generation time divided by audio duration. We additionally assess incremental delivery using service RTF and time to first audio (TTFA). See Appendix C.2 for further evaluation details.

Baseline Models. We compare MMAudio, AudioX, ThinkSound, and PrismAudio (Cheng et al., 2025; Liu et al., 2025; 2026a). MMAudio produces mono audio and is evaluated on the non-spatial axes; the other baselines produce binaural spatial audio. Both ISAudio modes generate audio autoregressively: **Video-guided** uses visual observations through the current target chunk, whereas **Interactive** uses only observations available at its start. We report sampling-function evaluations (NFE) per clip for offline baselines and per audio chunk for ISAudio.

5.2 Spatial Correspondence Learned by CSAVP

We test spatial correspondence by retrieving the matching listener trajectory among eight candidates from the same scene and time interval. The test covers 360 segments from unseen source videos and includes static and dynamic views. Full and Causal share the audio encoder and decoded inputs, isolating the effect of causal visual encoding. Table 2 reports mean bidirectional Top-1 accuracy at the chunk and clip levels. Causal achieves 34.50% clip accuracy against 12.50% chance, with differences from Full of 1.24 percentage points at the chunk level and 0.42 points at the clip level. Because candidates share scene content and event timing, this result supports sensitivity to viewpoint-dependent correspondence rather than scene identity alone. It also shows that prefix-only visual encoding retains spatial information for streaming conditioning. Directional retrieval scores and source-bootstrap intervals are reported in Appendix C.3.

Table 2: **CSAVP on FOA-T.** Mean bidirectional Top-1 accuracy (%).

Video Encoder	Chunk↑	Clip↑
Chance	12.50	12.50
Full	**32.66**	**34.91**
Causal	<u>31.42</u>	<u>34.50</u>

5.3 SPATIAL AUDIO GENERATION

Table 3: **Spatial audio generation on ViewSpatial.** Bold and underlined values indicate the best and second-best results.

	Efficiency	Semantic	Temporal	Aesthetic			Spatial Accuracy		Distribution		Subjective↑	
Model	NFE	CLAP↑	DeSync↓	PQ↑	CE↑	CU↑	C-sim↑	S-CLAP↑	FD↓	KL↓	MOS-Q	MOS-C
MMAudio	25	**0.515**	0.642	5.69	**3.80**	5.01	—	—	**6.34**	1.973	$\mathbf{3.62} \pm 0.12$	$\mathbf{3.42} \pm 0.23$
AudioX	250	0.397	1.141	**5.93**	3.78	**5.33**	0.0686	0.263	6.86	2.397	3.22 ± 0.16	2.81 ± 0.22
ThinkSound	24	0.437	0.740	5.63	3.78	5.02	0.0565	0.416	7.43	2.182	3.31 ± 0.15	3.02 ± 0.24
PrismAudio	24	0.415	**0.634**	5.37	3.31	4.52	0.0501	0.251	9.25	2.073	3.11 ± 0.16	3.18 ± 0.23
ISAUDIO – Video-guided												
SFT	25	0.487	1.143	5.18	3.42	4.47	**0.2488**	**0.621**	8.59	**1.436**	3.39 ± 0.18	3.30 ± 0.24
ISAUDIO – Interactive												
SFT	25	0.464	1.131	5.11	3.37	4.45	0.1697	0.582	8.94	1.520	3.46 ± 0.17	3.30 ± 0.22

Table 3 compares Video-guided and Interactive ISAUDIO at NFE$= 25$ against four V2A baselines, isolating observation regimes from few-step acceleration. Without future video frames, Interactive ISAUDIO exceeds the best binaural baselines in CLAP (0.464 vs. 0.437), C-sim (0.1697 vs. 0.0686), and S-CLAP (0.582 vs. 0.416). These results demonstrate semantic fidelity and stronger spatial correspondence under causal prediction. Four-step distillation results are provided in Appendix C.3.

5.4 INTERACTIVE SPATIAL AUDIO GENERATION

Table 4: **Interactive spatial audio generation on InteractSpatial.** GRPO refines the distilled model. IoU and FPO are reported as percentages.

	Efficiency		Semantic	Temporal			Aesthetic			Spatial Accuracy		Subjective↑	
Model	NFE	RTF↓	CLAP↑	DeSync↓	IoU↑	FPO↓	PQ↑	CE↑	CU↑	C-sim↑	S-CLAP↑	MOS-Q	MOS-C
MMAudio	25	0.45	0.579	0.510	9.14	17.75	6.90	**3.90**	6.44	—	—	3.54 ± 0.12	2.78 ± 0.23
AudioX	250	1.53	0.568	1.205	6.37	12.50	**7.01**	3.25	**6.85**	0.0259	0.664	3.64 ± 0.11	2.91 ± 0.19
ThinkSound	24	0.30	0.399	0.682	7.63	**5.96**	5.79	3.57	4.14	0.1493	0.227	2.92 ± 0.13	2.77 ± 0.13
PrismAudio	24	**0.28**	0.476	**0.421**	**10.39**	27.44	6.79	3.45	6.20	0.0145	0.469	3.71 ± 0.13	3.34 ± 0.22
ISAUDIO – Video-guided													
SFT	25	1.31	**0.591**	0.736	10.36	9.68	6.91	3.39	6.56	0.1890	**0.830**	$\mathbf{4.04} \pm 0.10$	$\mathbf{3.77} \pm 0.18$
ISAUDIO – Interactive													
SFT	25	1.31	0.549	1.098	9.67	11.26	6.89	3.36	6.48	0.1851	0.787	3.99 ± 0.10	3.61 ± 0.17
SFT	4	0.53	0.467	1.064	8.39	9.61	6.55	3.34	6.31	**0.2181**	0.783	3.84 ± 0.10	3.47 ± 0.20
+ Distillation	4	0.53	0.515	1.120	9.48	9.23	6.83	3.37	6.28	0.1703	0.731	3.87 ± 0.10	3.48 ± 0.18
+ GRPO	4	0.53	0.534	1.123	9.85	12.57	6.89	3.37	6.38	0.1707	0.773	3.96 ± 0.12	3.55 ± 0.20

Table 4 shows that, under chunk-start conditioning, Interactive ISAUDIO achieves competitive audio quality and stronger spatial correspondence than V2A baselines. At NFE$= 4$, distillation improves generation quality over direct SFT sampling, while GRPO further improves semantic and spatial alignment and event IoU. On one A100, the distilled model achieves a complete-segment RTF of 0.53. In eight-second incremental-delivery tests, four-step GRPO achieves a service RTF of 0.65 and a median TTFA of 1.18 seconds. Additional results are provided in Appendix C.3.

5.5 ABLATION STUDIES

Conditioning inputs. Table 5 shows that removing any visual stream lowers CLAP, with causal CSAVP removal also reducing C-sim from 0.1787 to 0.1688. Masking action/state conditioning at inference, while retaining video and generated audio history, reduces semantic and spatial correspondence and event IoU while increasing FPO, despite similar aesthetic scores.

Reward objectives. Table 5 highlights the complementary roles of the reward components. Single-reward training shows targeted gains from Semantic, Aesthetic, and Spatial feedback, supporting their complementary roles in perceptual optimization. Joint GRPO improves CLAP, PQ, S-CLAP, and IoU over the distilled baseline, supporting our dual-granularity design that combines clip-level perceptual guidance with chunk-local event and spatial feedback.

Table 5: **Ablation studies on conditioning inputs and GRPO rewards.** Visual ablations use ViewSpatial and InteractSpatial; action/state and reward studies use InteractSpatial. Conditioning studies use SFT (NFE= 25), while reward studies share a distilled initialization (NFE= 4).

	Semantic	Temporal			Aesthetic			Spatial Accuracy	
Variant	CLAP↑	DeSync↓	IoU↑	FPO↓	PQ↑	CE↑	CU↑	C-sim↑	S-CLAP↑
(a) Visual Conditioning (SFT, NFE= 25)									
Full SFT	**0.514**	**1.112**	**9.67**	11.26	6.15	3.36	5.64	0.1787	**0.702**
w/o DINO	0.507	1.122	9.63	11.60	6.17	**3.39**	5.64	**0.1807**	0.699
w/o CLIP	0.507	1.118	9.57	**10.89**	**6.19**	**3.39**	**5.66**	0.1771	**0.702**
w/o causal CSAVP	0.506	1.114	9.51	11.77	6.15	3.36	5.63	0.1688	0.693
(b) Action/State Conditioning (SFT, NFE= 25, InteractSpatial)									
Full SFT	**0.549**	**1.098**	**9.67**	**11.26**	**6.89**	**3.36**	**6.48**	**0.1851**	**0.787**
w/o action & state	0.509	1.197	8.76	15.35	6.86	3.35	6.43	0.1833	0.740
(c) Single-Reward Baselines (NFE= 4)									
No RL (Distilled)	0.515	1.120	9.48	**9.23**	6.83	3.37	6.28	0.1703	0.731
Semantic only	0.532	1.122	9.78	11.78	6.88	3.35	6.35	0.1846	0.763
Aesthetic only	0.519	1.142	9.33	9.45	**7.02**	**3.41**	**6.55**	0.1475	0.737
Spatial only	0.523	1.116	9.73	11.16	6.82	3.32	6.24	**0.1849**	0.772
(d) Event Reward Strategy Ablations (NFE= 4)									
Clip-level Event only	0.521	1.117	9.79	10.34	6.80	3.34	6.24	0.1747	0.752
No penalty Event only	0.521	1.127	**9.97**	14.09	6.77	3.35	6.26	0.1611	**0.780**
Chunk-level Event only	0.522	**1.115**	9.86	11.35	6.78	3.34	6.22	0.1666	0.760
(e) Joint Reward Optimization (NFE= 4)									
Full GRPO	**0.534**	1.123	9.85	12.57	6.89	3.37	6.38	0.1707	0.773

Event feedback. Chunk-level feedback yields higher IoU than clip-level feedback. The occupancy penalty reduces FPO from 14.09% to 11.35% while retaining similar IoU, supporting its role in suppressing excess event activity. Additional results are provided in Appendix C.3.5.

6 Limitations

While ISAudio enables action-conditioned streaming spatial audio generation, several limitations exist. First, using First-Order Ambisonics (FOA) data supports continuous view rotations but restricts fidelity under complex 3D translational motions with dynamic room impulse responses. Second, maintaining long-horizon acoustic consistency (e.g., avoiding feature drift for persistent sound sources) remains challenging for autoregressive modeling. Third, our interactive dataset relies primarily on a limited set of gaming environments, constraining scene and action diversity. Scaling to broader open-domain interactive scenarios remains an important future direction.

7 Conclusion

In this paper, we present ISAudio, a novel action-conditioned streaming spatial audio generation framework tailored for interactive world models. To address severe data scarcity, we construct spatial audiovisual datasets, ViewSpatial and InteractSpatial, with aligned camera trajectories, player controls, and multi-actor states. To enable low-latency interactive streaming, ISAudio integrates Causal-Consistency Distillation into an autoregressive flow-matching architecture within a binaural latent space, significantly accelerating sampling while strictly obeying causal observation boundaries. For fine-grained spatial grounding, we propose Contrastive Spatial Audio-Video Pretraining (CSAVP) to learn viewpoint-dependent representations across same-scene trajectory contrasts, serving both as a streaming visual conditioner and a bidirectional evaluator. Finally, we introduce a dual-granularity Group Relative Policy Optimization (GRPO) strategy to optimize the distilled generator by coupling chunk-local event and spatial rewards with clip-level perceptual feedback. Extensive evaluations on our benchmarks demonstrate that ISAudio achieves state-of-the-art performance in spatial consistency, temporal and event alignment, and real-time streaming efficiency.

AI USE STATEMENT

Generative AI tools assist with literature exploration, manuscript drafting and revision, and conceptual figure creation. The authors are responsible for the final text, references, figures, and scientific claims.

ETHICS STATEMENT

This work supports research on interactive world models and immersive audiovisual content creation. Our experiments use gameplay and real-world spatial audiovisual recordings, together with pretrained models, whose use and redistribution remain subject to the corresponding licenses and usage terms. Participants in our subjective listening evaluation provide informed consent. As realistic spatial audio could be used to create misleading audiovisual content, we encourage clear disclosure of synthetic outputs and responsible use that respects privacy and intellectual property.

REPRODUCIBILITY STATEMENT

Sections 4 and 3 describe our method and dataset construction. Appendices A, B, and C.1 provide model specifications, data processing and splits, and training, inference, and baseline settings. Evaluation protocols are detailed in Appendix C.2. We provide the anonymous demo page and an inference script for review.

REFERENCES

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Daniel Li, Piotr Dollár, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv:2504.13181*, 2025.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first international conference on machine learning*, 2024.

Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *European conference on computer vision*, pp. 17–36. Springer, 2020.

Taiye Chen, Xun Hu, Zihan Ding, and Chi Jin. Learning world models for interactive video generation. *Advances in Neural Information Processing Systems*, 38:154456–154483, 2026.

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. Mmaudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28901–28911. IEEE, 2025.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. *Advances in neural information processing systems*, 36:47704–47720, 2023.

Rishit Dagli, Shivesh Prakash, Robert Wu, and Houman Khosravani. See-2-sound: Zero-shot spatial environment-to-spatial sound. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Posters*, pp. 1–2, 2025.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Zach Evans, Julian D. Parker, Matthew Rice, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio 3, 2026. URL https://arxiv.org/abs/2605.17991.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

Anthony Hu, Václav Volhejn, Adrien Ramanana Rahary, Chris Mulder, Aditya Makkar, Alyx Liao, Amélie Royer, Manu Orsini, Adam Jelley, Eloi Alonso, et al. Multiplayer interactive world models with representation autoencoders. *arXiv preprint arXiv:2607.05352*, 2026.

Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5325–5329. IEEE, 2024.

Khaled Koutini, Jan Schlüter, Hamid Eghbal-Zadeh, and Gerhard Widmer. Efficient training of audio transformers with patchout. *arXiv preprint arXiv:2110.05069*, 2021.

Yichen Li and Antonio Torralba. Multimodal action conditioned video simulation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14173–14183. IEEE, 2025.

Susan Liang, Dejan Markovic, Israel D Gebru, Steven Krenn, Todd Keebler, Jacob Sandakly, Frank Yu, Samuel Hassel, Chenliang Xu, and Alexander Richard. Binauralflow: A causal and streamable approach for high-quality binaural speech synthesis with flow matching models. *arXiv preprint arXiv:2505.22865*, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2022.

Huadai Liu, Kaicheng Luo, Jialei Wang, Wen Wang, Qian Chen, Zhou Zhao, and Wei Xue. Thinksound: Chain-of-thought reasoning in multimodal large language models for audio generation and editing. *arXiv preprint arXiv:2506.21448*, 2025.

Huadai Liu, Kaicheng Luo, Wen Wang, Qian Chen, Peiwen Sun, Rongjie Huang, Xiangang Li, Jieping Ye, and Wei Xue. Prismaudio: Decomposed chain-of-thought and multi-dimensional rewards for video-to-audio generation. In *International Conference on Learning Representations*, volume 2026, pp. 25176–25204, 2026a.

Jiuming Liu, Chaojun Ni, Mengmeng Liu, Chensheng Peng, Fangjinhua Wang, Sitian Shen, Marc Pollefeys, Masayoshi Tomizuka, Ayush Tewari, and Per Ola Kristensson. Towards interactive video world modeling: Frontiers, challenges, benchmarks, and future trends. *arXiv preprint arXiv:2606.01164*, 2026b.

Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. *Advances in Neural Information Processing Systems*, 36: 48855–48876, 2023.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

Runjia Qian, Zile Wang, Jihai Zhang, Kai Zou, Wei Yu, Jiaxing Li, Zexiang Liu, Yaokun Li, Fei Kang, Kaichen Huang, et al. Matrix-game 3.5: Enhancing real-time streaming interactive world models with patch memory. *arXiv preprint arXiv:2608.29910*, 2026.

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.

Kentaro Seki, Yuki Okamoto, Kouei Yamaoka, Yuki Saito, Shinnosuke Takamichi, and Hiroshi Saruwatari. Spatial-clap: Learning spatially-aware audio–text embeddings for multi-source conditions. *arXiv preprint arXiv:2509.14785*, 2025.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

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

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. In *International Conference on Learning Representations*, volume 2025, pp. 73754–73776, 2025.

Jiahua Wang, Shannan Yan, Leqi Zheng, Jialong Wu, and Yaoxin Mao. Audio-visual world models: Towards multisensory imagination in sight and sound. *arXiv e-prints*, pp. arXiv–2512, 2025.

Yanan Wang, Linjie Ren, Zihao Li, Junyi Wang, and Tian Gan. Spatialv2a: Visual-guided high-fidelity spatial audio generation. *arXiv preprint arXiv:2601.15017*, 2026.

Kevin Wilkinghoff and Zheng-Hua Tan. DSpAST: disentangled representations for spatial audio reasoning with large language models. *arXiv:2509.13927*, 2025.

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.

Siyi Xie, Hanxin Zhu, Xinyi Chen, Tianyu He, Xin Li, and Zhibo Chen. Sonic4d: Spatial audio generation for immersive 4d scene exploration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11087–11095, 2026.

Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7151–7161. IEEE, 2024.

Mengchen Zhang, Qi Chen, Tong Wu, Zihan Liu, and Dahua Lin. Visaudio: End-to-end video-driven binaural spatial audio generation. *arXiv preprint arXiv:2512.03036*, 2025.

Yiming Zhang, Yicheng Gu, Yanhong Zeng, Zhening Xing, Yuancheng Wang, Zhizheng Wu, Bin Liu, and Kai Chen. Foleycrafter: Bring silent videos to life with lifelike and synchronized sounds. *International Journal of Computer Vision*, 134(1):46, 2026.

Lei Zhao, Rujin Chen, Chi Zhang, Xiao-Lei Zhang, and Xuelong Li. Foleyspace: Vision-aligned binaural spatial audio generation. *arXiv preprint arXiv:2508.12918*, 2025.

A Model Architecture

A.1 Audio Generator

The generator uses a frozen SAME-L stereo autoencoder and a diffusion transformer. CLIP supplies per-frame semantic features, DINO supplies spatial grids, and causal CSAVP supplies spatial-temporal features. Each stream is projected into the transformer context space. Table A.1 summarizes the dimensions.

Table A.1: **Generator architecture.** Continuous action width includes circular encoding of yaw.

Component	Configuration
Stereo autoencoder	44.1 kHz; downsampling factor 4096; latent width 256
Transformer	Width 768; 18 blocks; 16 heads
Cross-attention	6 blocks
Visual features	CLIP/DINO: 1024; CSAVP: 512
Spatial grid	3×16 cells per video frame
Actions	14 discrete controls; 12 continuous features
Actor state	10 slots; 42 features per slot; masked attention pooling
Interactive inference chunks	8 audio latent frames

A latent frame corresponds nominally to $4096/44100$ seconds. Interactive validation and inference use 8-frame prediction horizons, approximately 0.743 seconds. Training samples a new boundary independently at each valid latent frame with probability 0.2, always starting a chunk at the first frame. The untruncated mean length is 5 frames, approximately 0.464 seconds; finite clips truncate the last chunk. Every latent in a chunk uses the same start-time condition cutoff. Continuous listener state uses the latest valid sample no later than that start, rather than aggregating observations inside the predicted interval. Button tokens retain their own timestamps and are masked after the cutoff. DINO grid coordinates receive Fourier positional encodings and describe image layout.

Multi-actor inputs. The CS2 branch uses 10 stable round-local actor slots, with the listener in slot zero. Each 42-dimensional feature contains action/state attributes, listener-relative displacement and distance, an identity-of-listener flag, current/past shot-onset information, listener world position, and weapon-family features. Relative coordinates use front/left/up axes. Inputs include current game states for off-screen actors. A shared MLP and slot embeddings feed masked attention pooling; the pooled vector and listener vector are concatenated and projected to the transformer width. The zero-initialized output is added to diffusion-time block modulation. State is sampled at 16 Hz, with stale or unavailable snapshots masked. FOA inputs have no valid actor slots.

Condition timestamps. For a condition feature f, let $\tau(f)$ be the latest source-observation time used to construct it. Chunk c may access f only when $\tau(f) \leq s_c$.

A.2 CSAVP Encoders

The audio tower fuses separate left/right CLAP embeddings, their mean and difference, projected DSpAST (Wilkinghoff & Tan, 2025) features, and an audio-derived azimuth token. Cross-ear fusion uses two Transformer blocks with eight attention heads, hidden width 512, and feed-forward width 2,048. The video tower combines a PE-AV/EVA backbone with a temporal encoder, and the joint embedding width is 512. The PE-AV temporal encoder has four Transformer blocks with six attention heads, hidden width 768, and feed-forward width 2,048. The bidirectional encoder supports full-clip evaluation; its causal counterpart uses prefix-only temporal operations and the latest frame for clip readout. Its two kernel-three temporal convolutions use left-only padding in causal mode.

B DATA

B.1 DATA CONSTRUCTION

FOA rendering and synchronization. The renderer uses AmbiX first-order audio with ACN channel order (W, Y, Z, X) and SN3D normalization. Its world frame uses forward, left, and up axes, with positive yaw turning left. A world azimuth α therefore has listener-relative azimuth $\alpha - \psi$ under listener yaw ψ. The same yaw/pitch trajectory determines the panoramic field-of-view crop and the FOA rotation before SOFA-HRTF binaural decoding. Static trajectories use fixed directions, while dynamic trajectories include smooth randomized and source-guided rotations. The state stream and both media modalities share the segment's temporal origin. Ordinary field-of-view sources retain a static observed view rather than receiving unsupported panoramic rotations.

Trajectory construction and scope. Each physical segment is cut before expanding its viewing trajectories. Dynamic trajectories include smooth randomized turns and turns guided by an estimated dominant acoustic direction; acoustic guidance is used during data construction. Audio preserves the recorded mixture, including background and off-screen activity. Trajectories vary orientation at the recording position, with panoramic projection metadata determining the visual crop. Grouping trajectories from the same scene and time interval directs contrastive learning toward viewpoint-dependent correspondence.

Segment-level filtering. We assess opposing front and rear static views. Audio silence is measured using 20 ms windows and a -35 dBFS threshold; a segment is rejected if both views are silent for more than 90% of their duration. Visual filtering uses 20 sampled grayscale frames, rejecting a segment when both views have more than 85% near-static adjacent-frame pairs, with mean squared difference below 5. The audiovisual filter averages ImageBind scores over short aligned windows and rejects a segment only when both view means fall below 0.15. Retained segments preserve their associated views and trajectories.

CS2 alignment and annotation scope. Recorded player-state sequences map demo ticks to media time, including skipped ticks. Events use the recorded perspective's valid interval. Shot-counter increments and button/state transitions provide event candidates, running-distance rules approximate footsteps, and grenade-related damage records supply proxy supervision. Available positions are converted into listener-relative labels using event-specific conventions. The event reward compares detections with these engine-derived temporal annotations, while reference-based screening is used for spatial feedback.

B.2 DATASETS AND SPLITS

CSAVP uses 28,015 training, 299 validation, and 372 test FOA segments counted before trajectory expansion, with 360 test segments eligible for retrieval evaluation. Generation evaluation uses 3,658 ViewSpatial audiovisual pairs from held-out validation data and 5,144 InteractSpatial windows from four held-out matches, with identical evaluation lists across methods.

Splits are source-disjoint at the original-video level for ViewSpatial and the match level for InteractSpatial; all windows, views, and trajectories from the same source remain in one split. Generation evaluation sources are excluded from both CSAVP representation training and generator training.

C EXPERIMENTAL DETAILS

C.1 EXPERIMENTAL CONFIGURATION

C.1.1 GENERATOR TRAINING AND POST-TRAINING

The generator undergoes video-conditioned pretraining (Stage1), control/state-conditioned supervised fine-tuning (Stage2), causal consistency distillation, and GRPO post-training. The main Stage1 model includes all three visual streams; the three leave-one-stream-out variants supply the corresponding Stage2 visual ablations. Stage2 enables listener and actor branches and starts a fresh

optimizer. Distillation uses the frozen Stage2 teacher, and the resulting EMA student initializes both the trainable GRPO policy and its frozen reference. Table C.1 specifies the supervised settings and the distilled-policy GRPO configuration. Rollout and replay use identical chunk-start cutoffs and recorded history realizations, including any history-latent perturbations. CSAVP representations are trained separately from the audio generator.

Table C.1: **Generator training settings.** Update limits specify optimization budgets.

Setting	Stage1	Stage2	GRPO
Learning rate	10^{-4}	5×10^{-5}	2×10^{-6}
Warmup updates	5,000	1,000	*n/a*
Update limit	100,000	50,000	1,000
Batch per device	4 clips	8 clips	1 prompt
Gradient clipping	0.5	0.5	1.0
EMA decay	0.9999	0.9999	*n/a*
Candidates per prompt	*n/a*	*n/a*	8
Sampling steps per chunk	*n/a*	*n/a*	4
Optimized transitions per chunk	*n/a*	*n/a*	4
Replay passes	*n/a*	*n/a*	1
Ratio clipping	*n/a*	*n/a*	0.2
Reference weight	*n/a*	*n/a*	0.04

Stage1 uses 16 devices and Stage2 uses 8, both with a global batch of 64 clips. Supervised training uses parallel flow matching: flow time is sampled independently per chunk, with probability 0.2 of sharing one flow time across the entire clip. Stage2 jointly drops listener controls, continuous state, and all actor slots with probability 0.2 per clip; the shared mask includes the listener slot and leaves audio targets unchanged. Available controls are retained during evaluation. The actor branch is disabled throughout Stage1. Each rollout is replayed once, optimizing every transition of every valid chunk.

Distilled-policy likelihood and regularization. For a schedule $0 = \tau_0 < \cdots < \tau_N = 1$, the transition at step n is

$$\pi_\theta(z_{n+1} \mid z_n, h^c) = \mathcal{N}\big(\tau_{n+1} G_\theta(z_n, \tau_n; h^c),\, \sigma_n^2 I\big)\,. \tag{C.1}$$

For the endpoint sampler in Eq. C.1, let $\mu_\theta = \tau_{n+1} G_\theta(z_n, \tau_n; h^c)$ and let $\mathcal{I}_c$ index the valid latent coordinates of chunk c. The training noise scale is

$$\sigma_n = \begin{cases} 1 - \tau_{n+1}, & n < N-1, \\ \sigma_{\text{term}}, & n = N-1, \end{cases} \qquad \sigma_{\text{term}} = 0.02. \tag{C.2}$$

The joint transition log-density sums over valid coordinates:

$$\ell_\theta = -\frac{1}{2} \sum_{d \in \mathcal{I}_c} \left[\frac{(z_{n+1,d} - \mu_{\theta,d})^2}{\sigma_n^2} + \log(2\pi\sigma_n^2) \right], \qquad \rho_\theta = \exp(\ell_\theta - \ell_{\text{old}}). \tag{C.3}$$

All policies use the same transition variance, so the reference penalty in Eq. 6 has the closed form

$$D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) = \frac{1}{2\sigma_n^2} \sum_{d \in \mathcal{I}_c} (\mu_{\theta,d} - \mu_{\text{ref},d})^2. \tag{C.4}$$

The ratio and KL therefore use joint latent-transition distributions, with no averaging over latent coordinates. At inference, the final endpoint is deterministic, while intermediate steps retain the distilled sampler's re-noising schedule. The expectations in Eq. 6 average over all valid sampling transitions and candidates within each input, with inputs weighted equally.

Global rewards. Semantic reward is the LAION-CLAP cosine similarity between generated and reference audio after mono downmixing and matched deterministic cropping. Quality reward is the mean of the raw Audiobox-Aesthetics CE, CU, and PQ scores. The score is broadcast to all chunks as global credit, rather than interpreted as a measurement of every moment in the clip.

Event reward. PE-A-Frame queries all eight configured event classes, including classes absent from the reference annotation: gunshot, footstep, explosion, land, jump, reload, use, and attack. Predicted spans use a detection threshold of 0.3 and are merged before scoring. For class e and chunk c of valid duration d_c, let T_e^c and $P_e^{g,c}$ be the engine-derived and predicted interval sets clipped to that chunk. When T_e^c is nonempty, the local score is continuous-time IoU, $I_e^{g,c} = |P_e^{g,c} \cap T_e^c| / |P_e^{g,c} \cup T_e^c|$; otherwise, the penalty is the predicted occupancy $O_e^{g,c} = |P_e^{g,c}|/d_c$. Here $|\cdot|$ denotes duration, and empty annotation/prediction pairs receive zero rather than a positive silence reward.

Positive and negative contributions are normalized separately: first take duration-weighted averages within each class's annotated or unannotated chunks, then average equally across classes eligible for that term. This gives $r_{\text{evt}} = \overline{I} - \lambda_{\text{neg}}\overline{O}$, with $\lambda_{\text{neg}} = 0.25$. The corresponding per-chunk contributions sum exactly to r_{evt}, and their denominators depend only on annotations and valid duration. All interval operations exclude padding and use the actual duration of the final chunk.

Spatial reward. The frozen CSAVP audio direction branch projects the DSpAST azimuth pre-head representation through its learned direction-to-contrastive-space projection and normalizes the result. For each native two-second window w, its generated/reference similarity is $r_{\text{sp},w}^g = (1 + \langle a_w^g, a_w^{\text{ref}} \rangle)/2$. Reference-active duration weights the window scores, and temporal overlap maps them to generation chunks while preserving the aggregate spatial reward. Generated energy does not determine window eligibility, and this reward uses the audio direction branch rather than the full audiovisual C-sim score.

Aggregation and temporal credit. The total candidate reward is

$$R^g = r_{\text{sem}}^g + 0.6 r_{\text{qual}}^g + 4 r_{\text{evt}}^g + 6 r_{\text{sp}}^g. \tag{C.5}$$

Event and spatial contributions form the local credit $q^{g,c}$ in Section 4.4. To account for decoder lookback, each contribution is reassigned one chunk later and clamped to the last valid chunk: $\widetilde{q}^{g,c} = \sum_{j:\min(j+1,C)=c} q^{g,j}$. This preserves the sum, so the first return $Q^{g,1}$ equals the total candidate reward. Advantages subtract the candidate mean at each chunk position and divide by the root-mean-square centered residual over all valid candidate–chunk pairs within the same input. The complete return and normalization are

$$\begin{aligned} Q^{g,c} &= R_{\text{clip}}^g + \sum_{j=c}^{C} \widetilde{q}^{g,j}, \quad & \bar{Q}^c &= \frac{1}{G}\sum_{g=1}^{G} Q^{g,c}, \\ \hat{A}^{g,c} &= \frac{Q^{g,c} - \bar{Q}^c}{s_{\text{res}} + \delta}, & s_{\text{res}}^2 &= \frac{1}{GC}\sum_{g,c} (Q^{g,c} - \bar{Q}^c)^2, \end{aligned} \tag{C.6}$$

with padding excluded from all averages and $\delta > 0$ ensuring numerical stability.

C.1.2 CSAVP Training

Short-window initialization uses 2-second audio and four video frames, combining trajectory classification, cross-content contrast, and binaural perturbations. Perturbations include dual-mono conversion, channel swap, relative-channel-energy swap, and short interaural delays. The temporal model uses 8-second windows, 32 video frames at 4 FPS, and four consecutive binaural intervals. Audio and EVA parameters remain frozen during temporal training.

Temporal adaptation uses AdamW with learning rate 10^{-5}, weight decay 0.01, temperature 0.07, and a global batch of eight physical segments. Each segment supplies eight trajectories, balanced between static and dynamic views. Causal adaptation replaces temporal operations with prefix-only counterparts and uses the last frame as the clip readout. Both evaluated models are selected on validation retrieval performance.

C.1.3 Baselines and Comparison Variants

MMAudio uses the released large 44.1-kHz model with native mono output; AudioX, ThinkSound, and PrismAudio form the binaural baseline set. Our video-only configurations use empty text, with ThinkSound-Light for ThinkSound and the released PrismAudio diffusion transformer without generated CoT. MMAudio uses 25 Euler evaluations, AudioX uses 250 DPM++ 3M SDE evaluations,

and ThinkSound and PrismAudio each use 24 Euler evaluations per generated clip. NFE counts guided sampling-function evaluations, without separately counting conditional and unconditional classifier-free-guidance branches; the guidance scales are 4.5, 7, 5, and 5, respectively.

Target windows and conditioning regimes. The generation protocol uses matched 8-second target windows and one output per item, without best-of-many selection. Longer native outputs are cropped to the same interval; shorter-window adapters use fixed tiling and boundary treatment. Offline baselines receive the full target video, whereas ISAUDIO uses observations through the current chunk in Video-guided mode and only observations available at the chunk start in Interactive mode. CS2 variants use available listener and actor state; FOA variants have orientation controls but no actor inputs. Reference audio and reference-derived captions are never generation conditions, and Interactive models do not receive future states or event labels.

Model and sampling variants. Both ISAUDIO modes generate audio autoregressively. Within each mode, SFT models are evaluated at NFE$=25$ and NFE$=4$, and distilled students at NFE$=4$. GRPO further refines the distilled Interactive student at the same NFE$=4$. The Video-guided and Interactive variants use 8-frame chunks with clean generated history ($t = 1$). Sampling and post-training comparisons within the Interactive mode keep the observation cutoff, chunk size, and history setting fixed. Supervised models use Euler integration; distilled and GRPO-refined models use the consistency endpoint sampler. Each sampling step makes one generator call without an additional classifier-free-guidance evaluation, so NFE equals the number of sampling steps per chunk. NFE excludes conditioning encoders and audio decoding; total generator calls per clip equal NFE multiplied by the number of chunks.

Controlled ablations. Visual ablations remove DINO, CLIP, or causal CSAVP in both supervised stages while retaining chunk-start conditioning and listener/actor branches. These models are evaluated before distillation or GRPO using Euler sampling at NFE$=25$, with Full SFT as the reference. IoU and FPO use InteractSpatial only. Single-reward GRPO activates only Semantic, Aesthetic, Spatial, or Event, using a common distilled initialization, 4,000 training examples, and 1,000 updates. All reward variants and the unmodified no-RL initialization are evaluated on the same 5,144 InteractSpatial samples with a matched sampling seed and four-step consistency sampler. Event ablations compare chunk-local versus clip-level credit and remove the absent-event occupancy penalty while retaining the overlap reward.

C.1.4 STREAMING EVALUATION

Complete-segment RTF is $\mathrm{RTF} = t_{\mathrm{elapsed}}/T_{\mathrm{audio}}$, with values below one indicating faster-than-real-time throughput. Service RTF excludes observation waiting, whereas TTFA measures time to the first stable audio output, including observation waiting and decoder buffering. All models use one NVIDIA A100 80GB, batch size one, and two CPU threads on the same 32 eight-second videos. Timing covers video decoding and resizing, online feature extraction, condition preparation and transfer, sampling, waveform decoding, and CPU delivery and cropping. Model loading, warm-up, input staging, loading existing control/state streams, and output writing are excluded. Native samplers and precision are retained, and visual features are recomputed for each input.

C.2 EVALUATION METRICS

Event IoU and false-positive occupancy (FPO) apply to annotated InteractSpatial windows; the remaining evaluation axes are shared by ViewSpatial and InteractSpatial. C-sim and S-CLAP assess binaural outputs and are not reported for the mono MMAudio baseline.

Semantic and temporal evaluation. CLAP is cosine similarity between normalized embeddings of generated and reference audio from a frozen MS-CLAP encoder. The semantic reward instead uses LAION-CLAP. Event evaluation compares PE-A-Frame predictions with engine-derived CS2 intervals for eight event classes, using a fixed detection threshold of 0.3. Engine-derived onsets are expanded using fixed class-specific durations; predicted and target intervals are merged, clipped to each eight-second window, and evaluated on a 20 ms occupancy grid. For window i and class e, let $P_{i,e}$ and $T_{i,e}$ be the predicted and target interval sets, d_i the window duration, and $\mathcal{N}_e = \{i :$

$T_{i,e} = \varnothing\}$ the windows without that event class. We accumulate intersections and unions across all windows before averaging classwise IoU, and report false-positive occupancy separately:

$$\begin{aligned} \text{IoU} &= \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \frac{\sum_i |P_{i,e} \cap T_{i,e}|}{\sum_i |P_{i,e} \cup T_{i,e}|}, \\ \text{FPO} &= \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \frac{\sum_{i \in \mathcal{N}_e} |P_{i,e}|}{\sum_{i \in \mathcal{N}_e} d_i}. \end{aligned} \tag{C.7}$$

Here $|\cdot|$ denotes duration and $\mathcal{E}$ contains the eight evaluated classes, each with nonzero denominators on InteractSpatial. Empty prediction–target pairs contribute neither intersection nor union; any class with an empty aggregate union is omitted from the IoU average. Predictions in GT-empty windows still enter the IoU union, while FPO isolates their occupancy rather than being subtracted from IoU. Both metrics are reported as percentages: higher IoU and lower FPO are preferred, and neither is the composite Event reward used for post-training.

DeSync uses a fixed Synchformer evaluator with the same audiovisual crops for all methods, averaging the first and last context windows when a clip exceeds its context. It measures misalignment in seconds, with lower values preferred.

Aesthetic quality and distribution. Audiobox-Aesthetics predicts four dimensions. We report production quality (PQ), content enjoyment (CE), and content usefulness (CU) on their native higher-is-better scales. Production complexity (PC) describes the number and layering of sound components and is omitted from quality ranking. FD is the Fréchet distance between split-level VGGish feature distributions. KL (Copet et al., 2023) averages $D_{\mathrm{KL}}(p_{\mathrm{ref}} \| p_{\mathrm{gen}})$ over matched clips, where p is the softmax distribution over 527 PaSST (Koutini et al., 2021) event classes and natural logarithms are used. All methods share preprocessing and feature implementations, with lower FD and KL preferred.

Spatial evaluation. C-sim measures temporally aligned audiovisual correspondence using the frozen bidirectional CSAVP evaluator. Each eight-second evaluation window is divided into $K = 4$ consecutive two-second intervals, which need not coincide with the generator's chunk boundaries. For window n and interval k, let $\mathbf{a}_{n,k}$ and $\mathbf{v}_{n,k}$ denote the generated-audio and temporally corresponding video embeddings, respectively. We first compute interval-wise cosine similarities and then average them over intervals and evaluation windows:

$$\text{C-sim} = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{K} \sum_{k=1}^{K} \frac{\mathbf{a}_{n,k}^{\top} \mathbf{v}_{n,k}}{\|\mathbf{a}_{n,k}\|_2 \|\mathbf{v}_{n,k}\|_2}. \tag{C.8}$$

Video interval embeddings are obtained from a single bidirectional forward pass over the full eight-second video and can incorporate cross-interval context. This full-context evaluator is used for offline scoring; generation retains the chunk-start conditioning constraint. Scores are raw cosine similarities, with higher values indicating closer audiovisual correspondence.

Reward and evaluation models. The reward configuration in Section 4.4 combines LAION-CLAP semantic similarity, PE-A-Frame event IoU and absent-event occupancy, the CSAVP audio direction branch, and Audiobox-Aesthetics. Event IoU/FPO and PQ/CE/CU therefore assess axes whose predictors also contribute to reward optimization. C-sim compares generated audio with video through the full CSAVP representation, whereas spatial reward compares generated and reference audio through its direction branch. These comparisons differ but share learned components, so C-sim is a related representation-level assessment rather than an independent spatial predictor. External Spatial-CLAP, distributional metrics, and MOS-Q/C provide complementary assessments without contributing to the rewards. MS-CLAP supplies a semantic encoder distinct from the LAION-CLAP reward model.

Subjective listening protocol. We conduct subjective evaluations using the Mean Opinion Score (MOS), with the listening interface shown in Figure 3. For each dataset, all models are evaluated on the same 100 test clips, with each generated clip rated by five expert listeners on a five-point Likert scale from 1 (Bad) to 5 (Excellent). MOS-Q assesses overall audio quality, while MOS-C measures overall audiovisual consistency across semantic, temporal, and spatial dimensions. For each clip, we average the five listener ratings to obtain its MOS.

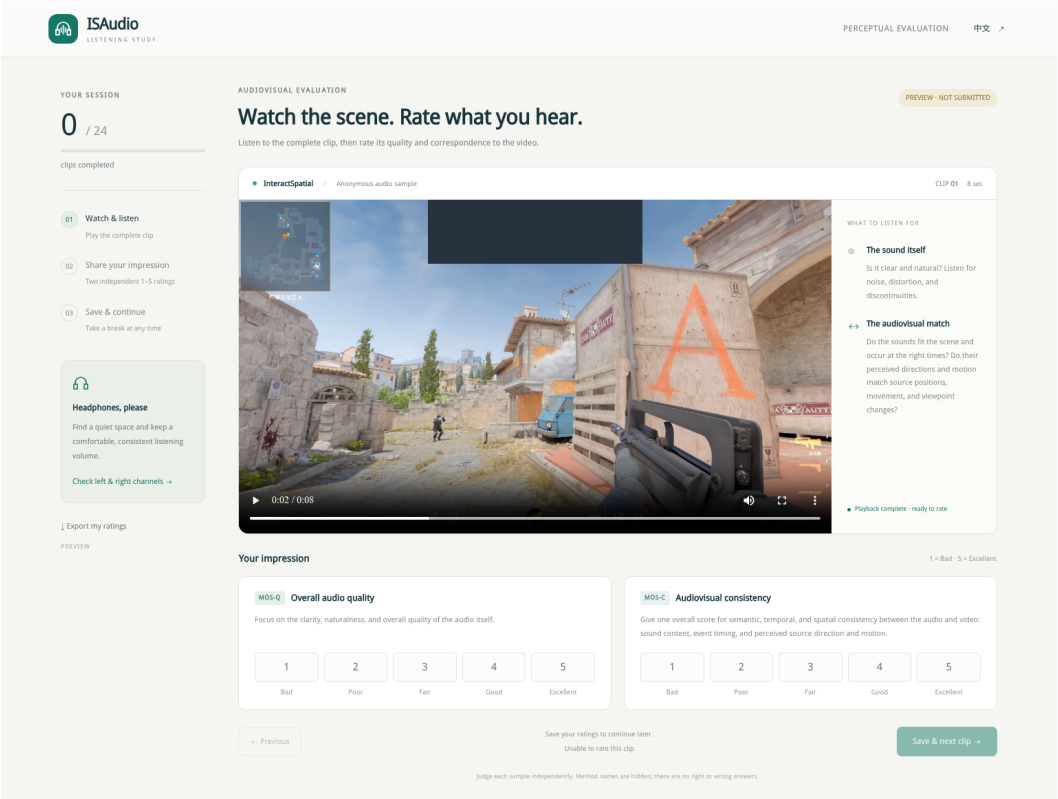

Figure 3: Preview of the listening interface for MOS-Q and MOS-C evaluation.

C.3 Additional Results

C.3.1 Spatial Retrieval on FOA-T

Table C.2: **Spatial retrieval on FOA-T.** Top-1 accuracy (%). Full denotes bidirectional video encoding. Mean averages both retrieval directions.

	Chunk			Clip		
Encoder	V→A	A→V	Mean	V→A	A→V	Mean
Full	**33.87**	**31.45**	**32.66**	**35.87**	**33.96**	**34.91**
Causal	32.22	30.62	31.42	35.49	33.51	34.50

Evaluation uses the same decoded inputs and audio-tower weights for both visual encoders. Of 372 test segments, 360 have all eight candidates valid and cover all 36 unseen sources. Validity excludes near-black video and near-silent audio, identically for both models. Table C.2 reports bidirectional retrieval; chance is 12.50%. Both models distinguish same-scene trajectories above chance.

C.3.2 Case Studies

Figures 4 and 5 compare generated audio with the reference and baseline outputs on ViewSpatial and InteractSpatial. In the ViewSpatial example, ISAudio captures the broad transition from left-dominant to right-dominant energy as the viewpoint changes. In the InteractSpatial example, it preserves the early right-dominant activity and the subsequent shift toward the left. Together, these examples illustrate the time-varying spatial structure of our generated audio under both video-guided synthesis and action-conditioned interactive prediction.

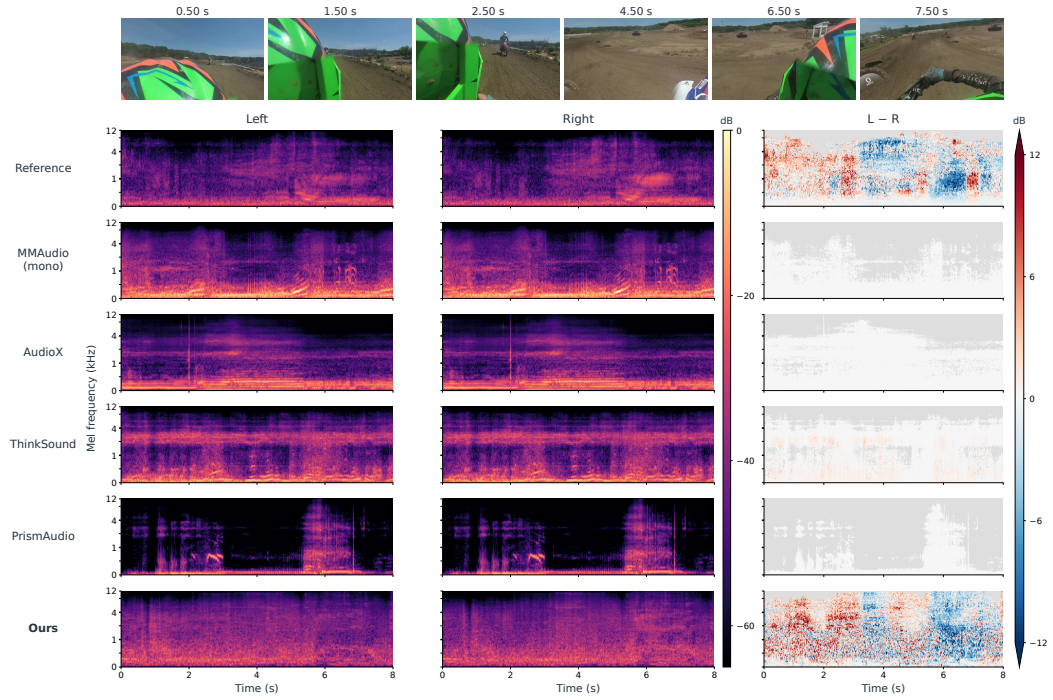

Figure 4: Qualitative comparison on ViewSpatial. Columns show the left-channel and right-channel mel spectrograms and their level difference, with a shared scale across the spectrograms. Red and blue indicate stronger left- and right-channel energy, respectively, while gray masks low-energy bins.

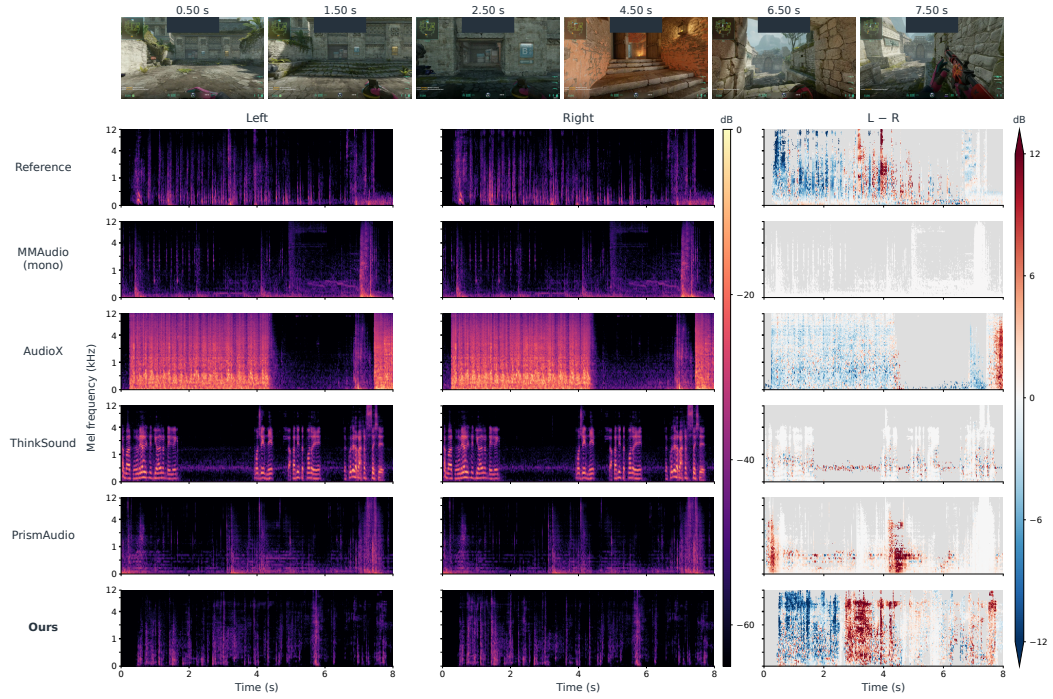

Figure 5: Qualitative comparison on InteractSpatial. Rows compare the reference audio, four baselines, and our model, using the visualization conventions in Figure 4.

C.3.3 FEW-STEP GENERATION

Table C.3: **Four-step generation on ViewSpatial.** SFT uses direct Euler sampling; distillation uses a learned consistency sampler at NFE$=4$.

		Semantic	Temporal	Aesthetic			Spatial Accuracy		Distribution	
Model	NFE	CLAP↑	DeSync↓	PQ↑	CE↑	CU↑	C-sim↑	S-CLAP↑	FD↓	KL↓
ISAUDIO – Video-guided										
SFT	25	**0.487**	1.143	5.18	**3.42**	4.47	**0.2488**	**0.621**	**8.59**	**1.436**
SFT	4	0.415	1.213	4.94	3.06	4.43	0.1727	0.556	14.40	1.825
+ Distillation	4	0.450	1.152	**5.21**	3.38	4.50	0.1918	0.576	10.25	1.506
ISAUDIO – Interactive										
SFT	25	0.464	**1.131**	5.11	3.37	4.45	0.1697	0.582	8.94	1.520
SFT	4	0.394	1.248	4.95	3.14	**4.59**	0.0934	0.520	15.02	1.986
+ Distillation	4	0.435	1.146	5.17	3.33	4.39	0.1133	0.539	9.83	1.620

Table C.3 compares direct four-step sampling with consistency distillation for both observation regimes on ViewSpatial, using the objective metrics in Table 3. The SFT rows at NFE$=25$ provide references for Video-guided and Interactive generation; none of these variants uses GRPO. For Interactive prediction, distillation raises CLAP from 0.394 to 0.435, C-sim from 0.0934 to 0.1133, and S-CLAP from 0.520 to 0.539, while reducing FD from 15.02 to 9.83 and KL from 1.986 to 1.620. Video-guided generation shows the same direction of change across these metrics (e.g., C-sim increases from 0.1727 to 0.1918 and S-CLAP from 0.556 to 0.576), demonstrating the benefit of learned few-step sampling in both settings.

Table C.4: **Video-guided generation on InteractSpatial.** SFT at NFE$=25$ is the reference for direct and distilled generation at NFE$=4$. IoU and FPO are percentages; no variant uses GRPO.

		Semantic	Temporal			Aesthetic			Spatial Accuracy	
Model	NFE	CLAP↑	DeSync↓	IoU↑	FPO↓	PQ↑	CE↑	CU↑	C-sim↑	S-CLAP↑
ISAUDIO – Video-guided										
SFT	25	**0.591**	0.736	10.36	9.68	**6.91**	**3.39**	**6.56**	0.1890	**0.830**
SFT	4	0.530	**0.734**	9.16	**8.42**	6.73	3.34	6.47	**0.2101**	0.820
+ Distillation	4	0.569	0.742	**11.00**	8.84	6.74	**3.39**	6.23	0.1947	0.807

Table C.4 completes the Video-guided comparisons on InteractSpatial. At NFE$=4$, distillation improves CLAP from 0.530 to 0.569, IoU from 9.16% to 11.00%, and PQ from 6.73 to 6.74 over direct SFT sampling; the corresponding Interactive results appear in Table 4.

C.3.4 GENERATION EFFICIENCY

Table C.5 compares end-to-end generation time under the protocol in Appendix C.1.4. Relative to 25-step Interactive SFT, the distilled and GRPO-refined models achieve $2.47\times$ and $2.49\times$ speedups, respectively. At four steps, direct SFT sampling and the distilled model have similar generation time; distillation improves the quality obtained at this sampling budget, as shown in Table 4.

Table C.5: **End-to-end generation efficiency.** All models use one A100 80GB and batch size one; time is the mean per eight-second output. NFE is per clip for baselines and per chunk for ISAUDIO.

Model	NFE	RTF↓	Time (s)↓
MMAudio	25	0.45	3.56
AudioX	250	1.53	12.21
ThinkSound	24	0.30	2.37
PrismAudio	24	**0.28**	**2.28**
ISAUDIO – Video-guided			
SFT	25	1.31	10.50
ISAUDIO – Interactive			
SFT	25	1.31	10.48
SFT	4	0.53	4.22
+ Distillation	4	0.53	4.24
+ GRPO	4	0.53	4.21

C.3.5 DISTRIBUTIONAL EVALUATION

Table C.6: **Distributional quality on InteractSpatial.** Outputs match Tables 4 and C.4. GRPO follows distillation.

Model	NFE	FD↓	KL↓
MMAudio	25	**9.28**	1.992
AudioX	250	15.24	2.303
ThinkSound	24	36.60	3.200
PrismAudio	24	30.87	2.434
ISAUDIO – Video-guided			
SFT	25	15.43	**1.108**
SFT	4	13.67	1.175
+ Distillation	4	23.81	1.210
ISAUDIO – Interactive			
SFT	25	17.91	1.418
SFT	4	17.36	1.441
+ Distillation	4	24.90	1.652
+ GRPO	4	23.73	1.504

Table C.7: **Distributional metrics for ablations.** Outputs match Table 5.

Variant	FD↓	KL↓
Visual (NFE$=25$)		
ViewSpatial + InteractSpatial		
Full SFT	**14.18**	**1.460**
w/o DINO	15.19	1.486
w/o CLIP	14.98	1.495
w/o causal CSAVP	15.47	1.493
Reward comparison (NFE$=4$) – InteractSpatial		
No RL	24.90	1.652
Semantic only	**23.66**	1.529
Aesthetic only	24.54	1.691
Spatial only	25.50	1.532
Event only	24.44	1.537
Clip-level	24.32	1.566
No penalty	23.78	**1.477**
Full GRPO	23.73	1.504

Tables C.6 and C.7 report FD and KL for the generation and ablation comparisons, using the same generated outputs. Each dataset-level FD is computed from the complete split-level feature distributions. For the cross-dataset visual ablations, FD is the sample-count-weighted mean of the two dataset-level FD scores, and KL uses the same sample-count weighting.