# WORLDSONUS: BRINGING SOUND TO WORLDS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Recent advances in world models have enabled increasingly realistic visual synthesis. However, these generated environments remain largely silent. Bringing sound to world models poses three core challenges: **real-time generation** to keep pace with interactive video streams, **interactive control** to respond to mid-stream sound instructions, and **spatially aligned stereo** to reflect scene geometry and camera motion. To address these demands, we introduce WorldSonus, an interactive video-to-audio framework designed for real-time spatial sound synthesis in world models. For real-time generation, WorldSonus employs a streaming causal autoregressive diffusion architecture that synthesizes audio chunks at a low real-time factor (RTF) of $0.41$. For interactive control, we incorporate an audio-centric captioning pipeline with chunk-indexed prompt scheduling, enabling dynamic manipulation of sound events during generation. For spatial alignment, we leverage high-quality stereo supervision curated from diverse stereo and ambisonic data. Extensive experiments demonstrate that while tailored for world models, WorldSonus generalizes effectively to open-domain video-to-audio benchmarks, matching or outperforming state-of-the-art bidirectional models in both acoustic quality and spatial alignment.

## 1 INTRODUCTION

Recent advances in video generation (Wang et al., 2025; Cai et al., 2025; HaCohen et al., 2025; Kong et al., 2024; Yang et al., 2025; Team, 2025b) have accelerated the development of generative world models (Valevski et al., 2025; He et al., 2025; Sun et al., 2025b; Gao et al., 2026). While these models dynamically synthesize visual environments in response to user actions, the resulting virtual worlds remain predominantly silent. Bringing realistic sound to interactive world models requires audio that seamlessly tracks an evolving visual stream, maintains camera-aligned spatial balance under dynamic viewpoints, and immediately responds to updated text instructions during an ongoing session.

Existing approaches to world-model sound generation fall into two primary categories. The first direction builds upon joint audiovisual foundation models (HaCohen et al., 2026) extended to long-horizon streaming (Duan et al., 2026; Zhang et al., 2026a; Su et al., 2026). However, coupling audio synthesis directly to the visual backbone prevents these systems from serving as standalone audio modules for visual-only world models, while causal distillation or long autoregressive (AR) rollouts often accumulate error over time. The second direction employs dedicated video-to-audio (V2A) generators, allowing modular integration with external visual engines. Yet current streaming V2A models remain constrained: V-AURA (Viertola et al., 2025) uses bidirectional visual windows and outputs mono audio. SoundReactor (Saito et al., 2025) achieves real-time stereo generation but lacks mid-stream text controllability. SwanSphere (Lei et al., 2026) focuses on panoramic video and first-order ambisonics (FOA) soundfields rather than standard perspective streams. Consequently, there remains a need for a modular V2A framework that simultaneously delivers causal real-time generation, mid-stream interactive text control, and spatially aligned perspective stereo audio.

We present WorldSonus, a modular V2A framework designed to endow generative world models with real-time, controllable stereo audio. For **real-time generation**, WorldSonus employs a streaming AR-diffusion architecture emitting $100$ ms audio chunks conditioned only on past and current visual context. A bounded persistent state ensures constant memory usage and compute cost as the stream expands. For **interactive control**, training on time-varying prompt schedules combined with chunk-level cross-attention enables users to dynamically revise text instructions for on-screen or off-screen sounds without session resets or past-frame recomputation. For **spatially aligned stereo**, we use clear-stereo clips and convert panoramic FOA data into camera-aligned stereo supervision.



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.

Generating precise audiovisual alignment within causal contexts introduces unique challenges, as future visual frames are unavailable. To enforce temporal synchronization under limited lookahead, we introduce two-timescale visual conditioning and ShiftNCE. Chunk-level visual summaries guide the AR backbone to maintain semantic continuity, while frame-aligned local tokens condition the flow head with fine-grained spatial and temporal cues. To refine event timing without inference cost, ShiftNCE trains the AR state using a contrastive objective supervised by a frozen synchronization expert, eliminating ambiguous temporal negatives during training.

Experiments demonstrate that despite causal operation, WorldSonus matches or exceeds state-of-the-art offline bidirectional V2A baselines across both open-domain VGGSound (Chen et al., 2020) videos and our Interactive benchmark, which comprises gameplay and real-world footage. WorldSonus achieves VGGish FAD (Hershey et al., 2017; Kilgour et al., 2019) scores of $1.73$ on $5$ s clear-stereo VGGSound (Chen et al., 2020) clips and $2.03$ on $30$ s interactive videos, with an execution latency of $41.2$ ms per $100$ ms audio chunk (real-time factor, $\mathrm{RTF} = 0.41$) on a single NVIDIA H100 GPU.

Our main contributions are threefold:

- **Real-time, interactive stereo generation.** We introduce WorldSonus, a modular V2A framework that unifies real-time causal generation, mid-stream prompt responsiveness, and camera-aligned stereo synthesis for interactive world models.
- **Causal audiovisual synchronization.** We propose two-timescale visual conditioning and a training-only ShiftNCE distillation objective to maintain temporal synchronization under causal streaming without adding runtime latency.
- **State-of-the-art performance.** Extensive evaluations show that WorldSonus achieves competitive or superior acoustic quality and stereo balance compared with offline bidirectional baselines, with counterfactual tests confirming robust interactive control.

## 2 Related work

**Bidirectional video-to-audio generation.** Bidirectional video-to-audio (V2A) synthesis has evolved from text-to-audio latent diffusion models (Liu et al., 2023a) and early V2A diffusion networks (Luo et al., 2023; Zhang et al., 2026b) to time-aligned models and rectified flow architectures (Wang et al., 2024; Cheng et al., 2025), improving semantic and temporal fidelity through audio-visual pretraining and explicit synchronization features. Recent models such as AudioX (Tian et al., 2025), ThinkSound (Liu et al., 2025a), and PrismAudio (Liu et al., 2025b) render stereo via flow matching and reward alignment. AC-Foley (Fang et al., 2026) adds reference-audio conditioning, while Omni2Sound (Dai et al., 2026) investigates video-text-to-audio diffusion across modality combinations. Meanwhile, spatial audio generation has advanced from diffusion-based binaural synthesis (Leng et al., 2022) toward object-aware stereo and ambisonics (Karchkhadze et al., 2025; Kim et al., 2025; Liu et al., 2025c). However, these approaches require full-clip visual context and do not support streaming video input or dynamic control within a persistent causal session.

**Streaming video-to-audio generation.** V-AURA (Viertola et al., 2025) uses AR audio-token prediction with chunked waveform decoding, while SwanSphere (Lei et al., 2026) extends causal generation to spatial audio conditioned on panoramic video and text. Neither establishes fully end-to-end causal streaming: V-AURA conditions each audio step on a bidirectional visual window and non-causal DAC (Kumar et al., 2023) decoding, and SwanSphere's variational autoencoder (VAE) (Kingma & Welling, 2014), based on Stable Audio Open (Evans et al., 2025), does not support stateful stream decoding. SoundReactor (Saito et al., 2025) combines a causal AR backbone with a diffusion head for streaming audio from gameplay video. Its V2A generator weights are not public, so we adopt only the released causal VAE. Building on this line of work, we target open-domain, interactive video with bounded persistent state and sound instructions that can change mid-stream.

**Interactive world models and joint audio–video generation.** Interactive video world models such as GameGen-X (Che et al., 2025) and LingBot-World 2.0 (Gao et al., 2026) support controllable visual rollouts and increasingly persistent interaction. Complementary approaches, including Omni-Forcing (Su et al., 2026) and Ripple (Ding et al., 2026), jointly generate streaming audio and video. We instead treat audio as a separate causal module conditioned on an externally produced visual stream: the upstream model exposes only realized frames, without requiring access to its action space or internal architecture, while users or agents independently update audio prompts.

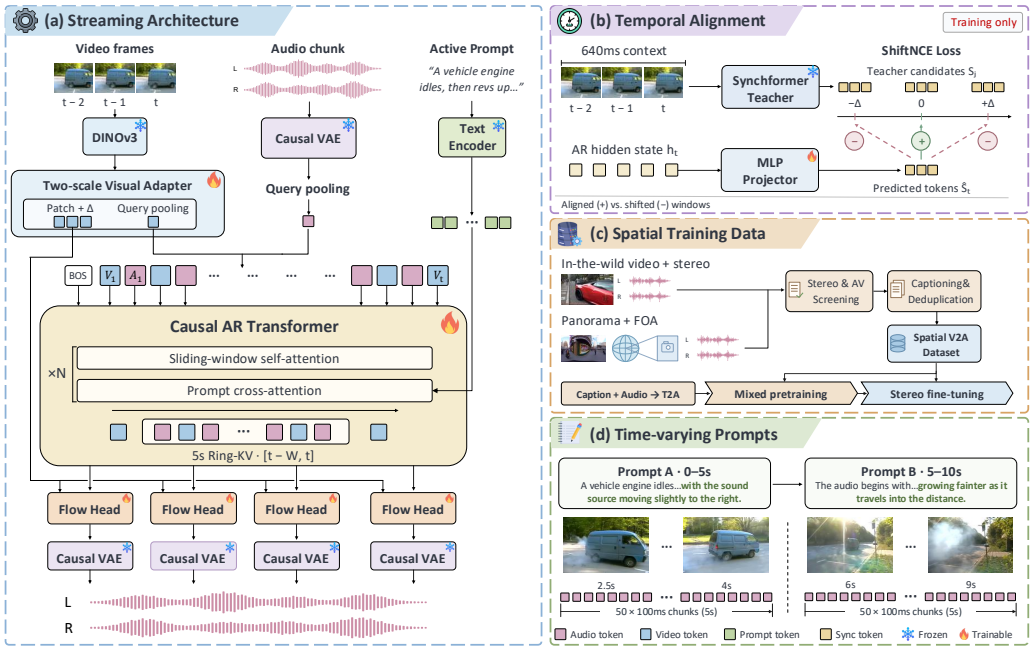



Figure 1: Overview of WorldSonus. (a) **Causal streaming pipeline**: Video frames map to two-timescale visual tokens conditioning an AR transformer and a rectified-flow head. (b) **Training-only ShiftNCE**: A training-only frozen Synchformer teacher guides temporal alignment via contrastive window matching. (c) **Spatial stereo supervision**: In-the-wild stereo filtering and panoramic FOA view-decoding provide directional audio grounding. (d) **Interactive prompt control**: Text prompts update dynamically at chunk boundaries, preserving session state and acoustic continuity.

## 3 WorldSonus

WorldSonus synthesizes real-time stereo audio from streaming video and dynamic prompts via chunk-wise causal AR generation (Figure 1). It couples a bounded recurrent memory with multi-scale visual conditioning and contrastive temporal alignment.

### 3.1 Streaming Formulation

At each streaming chunk $t \in \{1, \ldots, T\}$, the model receives visual frames $\mathbf{v}_t$ and active prompt $\mathbf{p}_t$ to predict the corresponding audio latent chunk $\mathbf{a}_t$. Conditioned on a bounded history state $\mathbf{M}_{t-1}$, the joint distribution factorizes causally over time:

$$p_\theta(\mathbf{a}_1, \ldots, \mathbf{a}_T) = \prod_{t=1}^{T} p_\theta(\mathbf{a}_t \mid \mathbf{M}_{t-1}, \mathbf{v}_t, \mathbf{p}_t), \tag{1}$$

where $\mathbf{M}_{t-1}$ is maintained by a sliding-window Key-Value cache bounded by maximum length $W$. This factorization uses only frames of past and current chunks, without access to frames of future blocks. Audio is synthesized in the continuous latent space of a frozen causal stereo VAE (Saito et al., 2025), which encodes 48 kHz stereo audio into latents advancing at 30 Hz. Its stateful decoder converts predicted chunks to waveform during streaming, with no re-decoding of past latents. Each 100 ms generation chunk corresponds to three latent frames $\mathbf{a}_t \in \mathbb{R}^{3 \times d_a}$, time-aligned with the three visual frames in $\mathbf{v}_t$ (30 FPS).

### 3.2 Autoregressive Diffusion

To balance long-range semantic continuity with high-fidelity continuous generation, we adopt a chunked AR-diffusion framework (Li et al., 2024; Saito et al., 2025; Lei et al., 2026). At each step, a decoder-only Transformer processes two chunk-aggregated tokens (one from the previous

audio chunk and one from the current video chunk) to maintain session state. Trained with a causal sliding-window attention mask, the model deploys a fixed-capacity Ring-KV cache over context window $W$ (a circular buffer overwriting the oldest slots in place). This ensures bounded memory while matching the context-window policy between training and inference.

The resulting AR hidden state $\mathbf{h}_t$ conditions a compact flow head, which denoises the three latent frames of chunk $t$ using a rectified-flow objective (Lipman et al., 2023; Liu et al., 2023b). The flow head attends bidirectionally inside the current chunk, without access to frames of future blocks, and it reads the frame-aligned visual tokens of Section 3.3 as finer-grained spatial and temporal evidence. Appendix A gives the remaining network details.

### 3.3 Two-Timescale Visual Conditioning

To reconcile long-horizon semantic continuity in the bounded AR backbone with precise acoustic rendering in the flow head, we decouple visual conditioning into two complementary timescales: **chunk-level semantics** and **frame-aligned local features**. For each chunk (spanning three frames), a frozen DINOv3 (Siméoni et al., 2026) encoder maps video frames into spatial patch grids $\mathbf{G}_n$. Mixed-aspect inputs are assigned to resolution buckets and resized without stretching at a fixed pixel budget. To capture motion cues, each grid is concatenated with its temporal difference $\Delta\mathbf{G}_n = \mathbf{G}_n - \mathbf{G}_{n-1}$ (Chen et al., 2026). A learned query then aggregates each frame's spatial grid into a single feature token, yielding three frame tokens per chunk.

These frame tokens are then processed into the two target timescales: For **chunk-level semantics**, a learned summary query aggregates the three frame tokens into a single chunk-level visual token, which is routed to the AR backbone to maintain a compact footprint. For **frame-aligned local features**, the three refined frame tokens bypass the AR backbone to condition the flow head directly. Our dual-path architecture provides the flow head with frame-aligned visual tokens without further chunk-level aggregation, which is essential for preserving the temporal granularity of individual video frames.

### 3.4 Interactive Prompt Control

To support interactive control during streaming, we introduce an in-place prompt control mechanism that allows text instructions to be updated mid-stream.

Given a sound instruction, a T5Gemma 2 (Zhang et al., 2025) text encoder and compressor project it into a compact token set that conditions the AR backbone via cross-attention, which is cached across streaming steps. When an instruction updates during streaming, the system performs an in-place replacement of the cross-attention cache at the nearest chunk boundary, preserving the Ring-KV cache and existing visual and audio states. To train the model for these dynamic transitions, we incorporate time-varying prompt schedules including mid-sequence instruction switches and drops. This enables seamless steering in a persistent causal session.

### 3.5 Training

**Teacher-forced pretraining.** The generator is pretrained under a rectified flow objective on a mixture of stereo video-audio and audio-only data, using teacher forcing over the bounded context window $W$. For a ground-truth latent chunk $\mathbf{a}_t$, noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and flow timestep $s \sim \mathcal{U}[0, 1]$, the linear interpolation state is $\mathbf{x}_{t,s} = (1 - s)\boldsymbol{\epsilon} + s\mathbf{a}_t$. The velocity matching loss is formulated as:

$$\mathcal{L}_{\text{flow}}(\boldsymbol{\epsilon}) = \| f_\theta(\mathbf{x}_{t,s}, s \mid \mathbf{C}_t) - (\mathbf{a}_t - \boldsymbol{\epsilon}) \|_2^2 , \tag{2}$$

where $\mathbf{C}_t$ aggregates the AR history state and local frame-aligned visual slots. We apply Explorative Modeling (Gladstone et al., 2026) to flow-matching training by evaluating three candidate source noises per step and optimizing the best-matching draw.

**Teacher-guided temporal alignment.** While existing synchronization models (Iashin et al., 2024) provide strong temporal cues, deploying them at inference introduces extra latency and windowing constraints. We instead employ a frozen synchronization encoder as a training-only teacher to distill temporal alignment directly into the generator.

Let $\mathbf{S}_t$ denote the teacher's target embedding for a video window ending at chunk $t$, and let a projector $P(\cdot)$ map the generator's internal AR state $\mathbf{h}_t$ to predict $\mathbf{S}_t$. To enforce temporal synchronization, our ShiftNCE objective contrasts the aligned target $\mathbf{S}_t$ against temporally shifted embeddings $\mathbf{S}_{t+\delta}$ from the same video across a set of non-zero chunk offsets $\delta \in \mathcal{D}_t$:

$$\mathcal{L}_{\text{sync}} = -\log \frac{\exp\left(\text{sim}(P(\mathbf{h}_t), \mathbf{S}_t)/\tau\right)}{\exp\left(\text{sim}(P(\mathbf{h}_t), \mathbf{S}_t)/\tau\right) + \sum_{\delta \in \mathcal{D}_t} \exp\left(\text{sim}(P(\mathbf{h}_t), \mathbf{S}_{t+\delta})/\tau\right)}, \tag{3}$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity, and $\tau$ is the temperature hyperparameter. To prevent false negatives during static visual segments, temporally shifted candidates with high visual similarity under the teacher model are filtered out. The final pretraining objective is the joint loss $\mathcal{L} = \mathcal{L}_{\text{flow}} + \lambda_{\text{sync}}\mathcal{L}_{\text{sync}}$, where $\lambda_{\text{sync}}$ balances the two objectives. Following pretraining, we fine-tune the network on a high-quality stereo video-audio subset. Complete training hyperparameter settings are detailed in Appendix B.

# 4 Data Curation

**Data sources.** Our dataset contains $1{,}465$ hours of audio, including $999$ hours of paired stereo video-audio data and $466$ hours of audio-only data. The *video-audio subset* uses open-source datasets including VGGSound (Chen et al., 2020), AudioSet (Gemmeke et al., 2017), Kinetics-700 (Carreira et al., 2019), and HD-EPIC (Perrett et al., 2025) to cover a wide range of acoustic events. To support audio generation for interactive world models, we collect additional standard stereo video with clear camera motion and dynamic objects. We also incorporate panoramic recordings from the existing Sphere360 (Liu et al., 2025c) and YT-AmbiGen (Kim et al., 2025) datasets. For panoramic clips, we estimate horizontal sound directions from the acoustic energy field to sample perspective visual crops and render view-aligned stereo audio. The *audio-only subset* includes AudioCaps (Kim et al., 2019) and isolated audio tracks from weakly-aligned video clips, where the visual content is not aligned with the audio, but the soundtrack retains high-quality acoustic events.

**Data filtering.** To ensure spatial quality and video-audio alignment, we use a two-stage filtering pipeline. First, automated signal processing removes silent, phase-corrupted, narrow-stereo, and dual-mono samples. Second, Qwen3-Omni (Team, 2025a) acts as a verifier to filter out severe cross-modal mismatches and non-diegetic audio like voiceovers and background music. Complete details appear in Appendix C.

**Captioning.** We generate dense audio-centric captions using Qwen3-Omni (Team, 2025a). Prompts focus on sound categories, acoustic texture, temporal timing, and spatial ambiance, while excluding visual details unrelated to sound to avoid introducing irrelevant conditioning cues. Captions are generated from both video and audio for video-audio data, and from audio alone for audio-only data. Annotation details are in Appendix C.3.

# 5 Experiments

## 5.1 Experimental Setup

**Evaluation sets.** We evaluate mainly on five test splits: two clear-stereo VGGSound (Chen et al., 2020) splits ($5$ s and $10$ s; $4{,}096$ clips each) and three splits comprising interactive gameplay and real-world footage with dynamic camera motion ($5$ s and $10$ s, with $4{,}096$ clips each; $30$ s, with $1{,}024$ clips). We additionally evaluate on physical collision events from Greatest Hits (Owens et al., 2016).

**Baselines.** WorldSonus is a causal model, using only frames from past and current chunks without access to future frames. In the absence of causal streaming stereo baselines, we evaluate against bidirectional stereo models including AudioX (Tian et al., 2025), ThinkSound (Liu et al., 2025a), and PrismAudio (Liu et al., 2025b), alongside the streaming monophonic baseline V-AURA (Viertola et al., 2025).

Table 1: Quantitative comparison across five held-out splits (4,096 clips for 5 s/10 s; 1,024 for 30 s). Best results are in bold, and second-best are underlined. Acc.: Bi (Bidirectional), Str (Stream), Caus (Causal). Cond.: VT (video+text), V (video-only). Chunk/$t$: chunk length and compute time in ms. IB and CLAP are multiplied by 100. S-FD$_O$ is omitted for monophonic V-AURA.

| | | | | | Distribution (mid) | | | | Spatial | Semantic | | Temporal |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Set | Model | Acc. | Cond. | Chunk/$t$ | FAD↓ | FD$_P$↓ | FD$_O$↓ | KL$_P$↓ | S-FD$_O$↓ | IB↑ | CLAP↑ | DeSync↓ |
| VGG 5 s | AudioX | Bi | VT | – | 3.00 | 153.29 | **38.19** | 1.66 | 87.60 | 26.96 | 37.87 | 0.986 |
| | ThinkSound | Bi | VT | – | 2.94 | 141.40 | 47.32 | 1.91 | 62.95 | 24.12 | 33.62 | **0.481** |
| | PrismAudio | Bi | VT | – | 2.09 | **132.58** | 50.72 | 1.67 | 94.77 | 26.28 | **39.20** | 0.539 |
| | V-AURA | Str | V | 640/636 | 4.06 | 304.06 | 51.01 | 2.03 | – | 26.48 | 24.91 | 1.287 |
| | WorldSonus (Ours) | Caus | VT | 100/41.2 | **1.73** | 159.74 | 39.03 | **1.48** | **43.47** | **27.06** | 38.35 | 0.686 |
| VGG 10 s | AudioX | Bi | VT | – | 3.86 | 194.72 | **41.50** | 1.54 | 83.53 | 27.84 | 34.56 | 0.944 |
| | ThinkSound | Bi | VT | – | 2.18 | **144.17** | 46.49 | 1.73 | 55.02 | 25.76 | 30.46 | 0.514 |
| | PrismAudio | Bi | VT | – | 2.64 | 154.69 | 52.38 | 1.47 | 118.08 | 28.16 | 33.26 | **0.427** |
| | V-AURA | Str | V | 640/636 | 4.35 | 318.13 | 57.71 | 1.95 | – | 26.73 | 23.35 | 1.208 |
| | WorldSonus (Ours) | Caus | VT | 100/41.2 | **1.79** | 167.52 | 41.90 | **1.32** | **45.40** | **28.82** | **36.77** | 0.687 |
| Inter. 5 s | AudioX | Bi | VT | – | 4.44 | 186.49 | 66.77 | 1.56 | 109.21 | 24.88 | 32.53 | 1.059 |
| | ThinkSound | Bi | VT | – | 8.49 | 256.38 | 61.84 | 1.98 | 58.68 | 18.39 | 22.15 | 0.733 |
| | PrismAudio | Bi | VT | – | 6.22 | 222.83 | 62.74 | 1.60 | 83.85 | 23.08 | 32.42 | **0.651** |
| | V-AURA | Str | V | 640/636 | 8.18 | 436.10 | 83.78 | 2.01 | – | **26.84** | 16.07 | 1.344 |
| | WorldSonus (Ours) | Caus | VT | 100/41.2 | **2.68** | **163.74** | **40.99** | **1.29** | **34.26** | 26.48 | **35.90** | 0.831 |
| Inter. 10 s | AudioX | Bi | VT | – | 5.05 | 352.02 | 69.19 | 1.57 | 119.16 | 25.39 | 27.89 | 1.077 |
| | ThinkSound | Bi | VT | – | 6.27 | 332.49 | 60.93 | 1.85 | 56.47 | 21.43 | 24.10 | 0.738 |
| | PrismAudio | Bi | VT | – | 5.16 | 338.17 | 61.23 | 1.42 | 96.12 | 25.20 | 31.08 | **0.599** |
| | V-AURA | Str | V | 640/636 | 9.75 | 532.74 | 86.46 | 1.96 | – | 27.60 | 15.47 | 1.192 |
| | WorldSonus (Ours) | Caus | VT | 100/41.2 | **2.62** | **262.25** | **42.91** | **1.30** | **37.42** | **28.63** | **34.22** | 0.820 |
| Inter. 30 s | AudioX | Bi | VT | – | 5.52 | 427.00 | 81.14 | 1.88 | 121.62 | **24.78** | 30.34 | 1.125 |
| | ThinkSound | Bi | VT | – | 7.00 | 473.40 | 73.46 | 2.31 | 81.88 | 11.71 | 19.50 | 0.901 |
| | PrismAudio | Bi | VT | – | 6.08 | 375.22 | 62.99 | 1.66 | 95.26 | 22.32 | 33.70 | **0.765** |
| | V-AURA | Str | V | 640/636 | 13.33 | 666.31 | 108.01 | 2.74 | – | 19.93 | 8.54 | 1.266 |
| | WorldSonus (Ours) | Caus | VT | 100/41.2 | **2.03** | **253.21** | **26.67** | **1.22** | **25.44** | 22.93 | **34.21** | 0.867 |

### 5.2 METRICS

**Audio quality and audiovisual alignment.** For *acoustic fidelity*, we report Fréchet distances in VGGish (Hershey et al., 2017) (FAD (Kilgour et al., 2019)), PaSST (Koutini et al., 2022) (FD$_P$), and OpenL3 (Cramer et al., 2019) (FD$_O$) feature spaces, together with PaSST-based KL divergence (KL$_P$). FAD and FD$_O$ use the mid-channel signal. S-FD$_O$ applies the OpenL3 distance to the side-channel signal to evaluate stereo distribution quality. For *audiovisual alignment*, we use ImageBind (Girdhar et al., 2023) (IB) similarity for semantic consistency and Synchformer (Iashin et al., 2024) offset (DeSync) for synchronization, supplemented by onset accuracy, F1, and AP on Greatest Hits following MMAudio (Cheng et al., 2025).

**Stereo balance agreement.** BiasSkill measures whether generated and reference audio favor the same left/right channel within 1 s windows. It reports the normalized gain over a permutation baseline that accounts for each model's fixed channel bias. Details are provided in Appendix E.

### 5.3 MAIN RESULTS

**Audio quality.** We conduct quantitative comparisons across all five held-out test splits. As shown in Table 1, WorldSonus achieves competitive or superior acoustic quality (FAD, FD$_P$, and KL$_P$) compared to bidirectional models on both VGGSound and interactive video datasets, despite using only visual frames from past and current chunks. Compared with the streaming baseline V-AURA, WorldSonus achieves lower distributional divergence and stronger audiovisual alignment while running at a much finer temporal resolution (100 ms vs. 640 ms), synthesized in just 41.2 ms per chunk on a single NVIDIA H100 GPU (RTF $= 0.41$) to comfortably satisfy real-time streaming requirements. Together, these results demonstrate that bounded-memory causal generation can combine competitive audio quality with real-time synthesis across the evaluated 5–30 s durations.

**Spatial and temporal alignment.** For spatial alignment, we complement distributional stereo metrics (S-FD$_O$ in Table 1) with stereo balance agreement measured by BiasSkill over 1 s windows (Table 2). WorldSonus achieves the highest BiasSkill across all evaluated splits. Figure 2 provides complementary qualitative evidence from two interactive clips with sound sources located predominantly on one side: beach surf on the left and a passing truck on the right. The left/right spectrograms and signed energy-balance curves $(E_R - E_L)/(E_R + E_L)$ show that WorldSonus reproduces the corresponding channel dominance, consistent with the ground-truth audio, whereas the baselines exhibit weaker or less consistent spatial alignment.

Table 2: Stereo balance agreement (BiasSkill, %). Full table in App. E.

| Model | VGG 10 s | Inter. 10 s | Inter. 30 s |
|---|---|---|---|
| AudioX | <u>3.79</u> | <u>3.08</u> | −0.01 |
| ThinkSound | 3.22 | −1.78 | 0.63 |
| PrismAudio | 1.10 | −0.06 | <u>4.19</u> |
| WorldSonus (Ours) | **4.42** | **6.03** | **15.90** |

Table 3: Onset detection on Greatest Hits.

| Model | Acc.↑ | F1↑ | AP↑ |
|---|---|---|---|
| AudioX | **0.858** | 0.770 | 0.790 |
| ThinkSound | 0.706 | 0.758 | <u>0.904</u> |
| PrismAudio | 0.783 | **0.808** | **0.906** |
| w/o finetune | <u>0.805</u> | <u>0.796</u> | 0.884 |
| w/o ShiftNCE | 0.678 | 0.560 | 0.681 |
| WorldSonus (Ours) | 0.803 | 0.785 | 0.871 |

For temporal alignment, bidirectional baselines that optimize directly on Synchformer features achieve lower DeSync (Table 1). Nevertheless, WorldSonus maintains competitive event timing without test-time guidance, as reflected in Table 3 on the model-free Greatest Hits onset benchmark.

Together, these results show that WorldSonus combines competitive temporal alignment with stronger spatial alignment under causal visual conditioning.

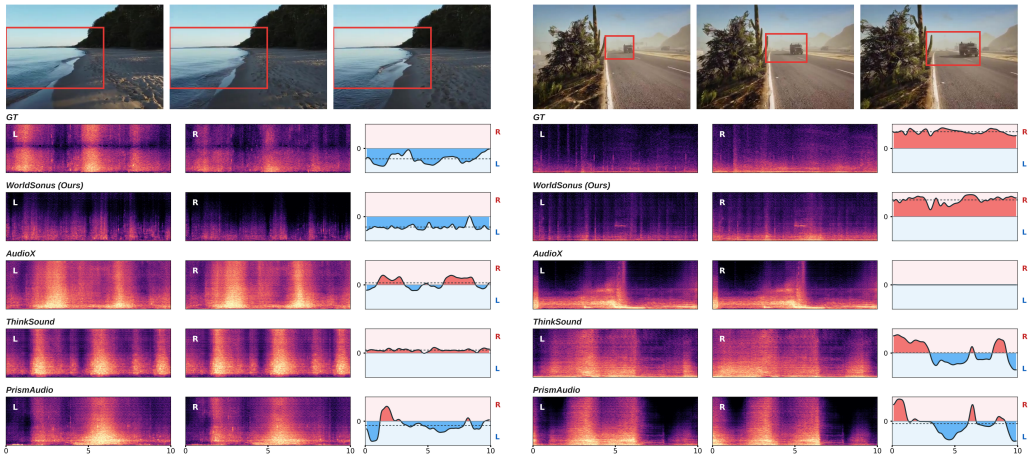



Figure 2: Qualitative stereo layout versus bidirectional baselines on two interactive clips. Red boxes mark sources. Energy-balance curves show left (negative) versus right (positive) channel dominance.

**Long-horizon stability.** To analyze potential drift during continuous streaming, we evaluate 1,024 Interactive 30 s clips under two regimes: *Direct* (generating the final 25–30 s window from a cold start) and *Rollout tail* (extracting the final window from an uninterrupted 30 s stream). As shown in Table 4, the rollout tail closely tracks the direct baseline in acoustic fidelity and synchronization (FAD 2.51 vs. 2.63, DeSync 0.827 vs. 0.880), with minimal shifts in OpenL3 and ImageBind. It indicates that under our bounded 5 s Ring-KV cache, WorldSonus maintains stable, collapse-free streaming across extended horizons without periodic resets.

Table 4: Long-horizon stability on Interactive 30 s (1,024 clips). Final 5 s window (25–30 s) generated directly versus via continuous rollout.

| Protocol | FAD↓ | FD$_P$↓ | FD$_O$↓ | KL$_P$↓ | S-FD$_O$↓ | IB↑ | DeSync↓ |
|---|---|---|---|---|---|---|---|
| Direct last 5 s | 2.63 | 271.68 | **35.16** | **1.22** | **39.24** | **21.38** | 0.880 |
| Rollout tail (25–30 s) | **2.51** | **246.69** | 40.46 | 1.31 | 41.77 | 19.99 | **0.827** |

**Text control.** To evaluate interactive prompt control, we measure text to audio agreement using CLAP (Wu et al., 2023) between each 5 s audio segment and its caption. For 10 s clips, let $P_1$ and $P_2$ denote the sequential instructions for the first and second 5 s halves ($P_1 \rightarrow P_2$). We report the dual relative match rate, defined as the percentage of clips where both 5 s segments score higher with their own caption than with the other. As shown in Table 5, WorldSonus achieves dual match rates of 25.68% on VGGSound and 23.44% on Interactive 10 s, outperforming all baseline architectures.

Table 5: Dual relative match rate (%). Both halves prefer their own reference prompt.

| Model | VGG 10 s | Inter. 10 s |
|---|---|---|
| AudioX | 15.41 | <u>14.04</u> |
| ThinkSound | 13.92 | 13.43 |
| PrismAudio | <u>16.16</u> | 11.99 |
| V-AURA (video-only) | 12.16 | 12.11 |
| WorldSonus (Ours) | **25.68** | **23.44** |

Table 6: Paired prompt-switch control. Gain $G$ is switch-minus-hold preference. $\pm$ is half the 95% percentile bootstrap interval (App. F).

| Set | Switch pref. | Hold pref. | Gain $G$ |
|---|---|---|---|
| VGG 10 s | +0.024 | −0.027 | 0.051 $\pm$0.004 |
| Inter. 10 s | +0.022 | −0.030 | 0.053 $\pm$0.003 |

However, strong text-audio alignment alone does not fully isolate text responsiveness from natural visual transitions. To rule out visual conflation, we perform paired counterfactual interventions by comparing a Switch setting ($P_1 \rightarrow P_2$ at 5 s) against a Hold baseline ($P_1 \rightarrow P_1$) under identical visual features and seeds. The intervention gain $G$ measures how much switching increases the second half's CLAP preference for $P_2$ over $P_1$. As shown in Table 6, WorldSonus yields positive net intervention gains $G$ (+0.0509 and +0.0525) with 95% bootstrap confidence intervals excluding zero. These combined results confirm that WorldSonus achieves active, causal text steering over generated audio with visual context held fixed. Details are provided in Appendix F.

**Subjective comparison.** To complement objective evaluations, we perform a blind, randomized human A/B study on 40 representative Interactive 10 s clips. Twenty assessors provided 400 pairwise evaluations across four criteria: spatial alignment, temporal alignment, semantic match, and overall preference, with ties counted as half a vote (Appendix G). As shown in Figure 3, WorldSonus is preferred over all three generated baselines across all criteria, while ground-truth audio remains preferred. Overall preference reaches 58.8% against AudioX, 80.0% against ThinkSound, and 65.0% against PrismAudio, indicating stronger overall listener preference.

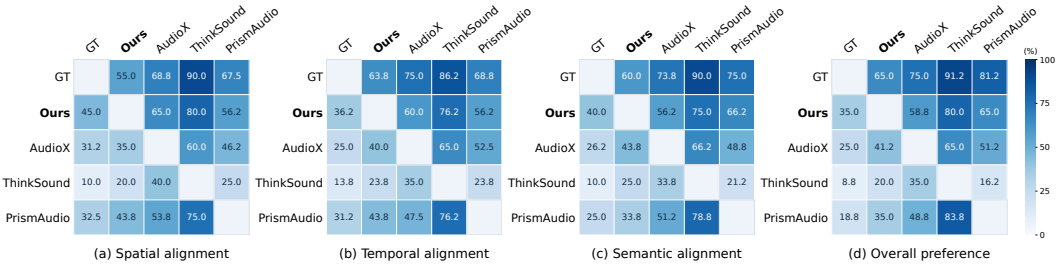



Figure 3: User-study preference matrices across (a) spatial alignment, (b) temporal alignment, (c) semantic alignment, and (d) overall preference. Cell $(i, j)$ is the percentage preferring the row method over the column method, counting each tie as half a vote (40 ratings per pair).

### 5.4 Ablation Studies

All ablation checkpoints are evaluated directly after pretraining without fine-tuning on the Interactive 10 s set (4,096 clips) under fixed classifier-free guidance (Ho & Salimans, 2022). Table 7 isolates architectural and alignment components, while Table 8 compares generation-chunk granularities.

**Temporal alignment objective.** To investigate whether the proposed ShiftNCE objective explicitly drives fine-grained synchronization, and whether external temporal representations can replace it, we compare our default model against a variant trained without ShiftNCE and another incorporating direct features from a 640 ms causal Synchformer (Iashin et al., 2024) window. As shown in Table 7, removing ShiftNCE degrades DeSync from 0.831 to 1.067, while direct feature injection yields worse timing and acoustic quality, possibly reflecting a mismatch between its 640 ms feature window and our 100 ms streaming chunks. These results confirm that ShiftNCE provides the necessary temporal supervision to sharpen alignment without requiring additional sync encoders or inference overhead.

**Visual feature representation.** To examine whether our DINOv3 patch-based representation outperforms the pooled CLIP-family features commonly used in prior V2A models (Zhang et al., 2026b; Cheng et al., 2025; Liu et al., 2025a; Shan et al., 2025), we compare it with pooled SigLIP 2 (Tschannen et al., 2025) features. We also evaluate a variant without temporal deltas ($\Delta \mathbf{G}_n$). Table 7 shows that our DINOv3 features substantially improve ImageBind alignment from 24.56 to 28.05 over the Pooled SigLIP 2 baseline. Furthermore, removing temporal deltas worsens both DeSync (0.958) and FAD (3.08). These results demonstrate the effectiveness of our DINOv3-based visual conditioning and explicit temporal cues for video-to-audio synthesis.

Table 7: Architecture ablations on Interactive 10 s.

| Variant | FAD↓ | $FD_P$↓ | $FD_O$↓ | $KL_P$↓ | S-$FD_O$↓ | IB↑ | DeSync↓ |
|---|---|---|---|---|---|---|---|
| w/o ShiftNCE | 2.49 | 276.68 | 48.50 | 1.44 | 54.08 | 27.20 | 1.067 |
| Direct sync features | 2.81 | 278.28 | 46.33 | 1.40 | 48.68 | 26.22 | 1.043 |
| w/o temporal delta | 3.08 | 267.51 | 49.37 | 1.40 | 68.76 | **28.77** | 0.958 |
| Pooled SigLIP 2 | 3.29 | 291.60 | **31.96** | 1.33 | **29.81** | 24.56 | 0.944 |
| w/o flow-head DINO | 15.82 | 906.49 | 117.12 | 2.25 | 127.59 | 12.65 | 1.020 |
| WorldSonus (Ours) | **2.42** | **266.41** | 40.97 | **1.31** | 40.17 | 28.05 | **0.831** |

Table 8: Chunk-length comparison on Interactive 10 s.

| Chunk | FAD↓ | $FD_P$↓ | $FD_O$↓ | $KL_P$↓ | S-$FD_O$↓ | IB↑ | DeSync↓ |
|---|---|---|---|---|---|---|---|
| 33 ms | 3.79 | 333.05 | 62.17 | 1.68 | 62.16 | 27.18 | 0.915 |
| 200 ms | 2.51 | 274.86 | **39.03** | **1.26** | **35.12** | 26.35 | **0.819** |
| 100 ms | **2.42** | **266.41** | 40.97 | 1.31 | 40.17 | **28.05** | 0.831 |

**Two-timescale visual conditioning.** To evaluate the role of feeding direct frame-level visual features to the flow head, we compare our model with a variant that removes this shortcut path and passes visual information only through the autoregressive state. As shown in Table 7, removing the direct visual connection leads to a severe degradation in audio quality, with FAD increasing from 2.42 to 15.82. This confirms that in our streaming design, combining direct frame-level visual inputs with higher-level autoregressive context is essential for generating clear and synchronized audio.

**Generation-chunk granularity.** To determine the optimal balance between streaming latency and generation quality, we evaluate chunk sizes across 33 ms, 100 ms, and 200 ms. Table 8 shows that shrinking the chunk size to 33 ms worsens acoustic quality (FAD 3.79), whereas increasing it to 200 ms increases latency without consistent improvements across metrics. We therefore choose 100 ms as a practical balance between streaming granularity and generation quality.

## 6 Conclusion

We present WorldSonus, a causal spatial audio synthesis framework for real-time sound generation in interactive world models. By factorizing generation into chunk-wise autoregressive flow matching, WorldSonus synthesizes 48 kHz stereo audio in 100 ms chunks with an execution latency of 41.2 ms per chunk on a single NVIDIA H100 GPU (RTF $= 0.41$). Our decoupled visual conditioning couples long-term planning via a bounded 5 s Ring-KV cache with frame-aligned local guidance in the flow head, while training-only ShiftNCE enforces precise temporal synchronization with zero inference overhead. Evaluated on open-domain videos and interactive environments, WorldSonus achieves acoustic fidelity and spatial localization competitive with or exceeding offline bidirectional models, while seamlessly supporting real-time prompt updates.

## 7 Limitations

**Causal audio representations.** WorldSonus generates audio in the continuous latent space of a frozen causal stereo VAE adapted from SoundReactor (Saito et al., 2025). SoundReactor (Saito et al., 2025) reports lower reconstruction quality for its causal VAE decoder than for its non-causal counterpart. Developing more expressive, large-scale causal audio codecs will be crucial to close this performance gap in future work.

**Stereo data scale.** High-quality spatial audio data remains scarce compared to massive monophonic corpora. Many public stereo videos contain artificial or noisy channel separation that does not accurately reflect visual object motion. While our two-stage filtering pipeline and panoramic soundfield decoding extract reliable spatial supervision, expanding the scale of authentic stereo data remains a key goal for future work.

### ETHICS STATEMENT

Generated audio may be mistaken for recorded sound or used to create misleading audiovisual content. Deployments should disclose synthetic audio and avoid recreating identifiable voices without consent. Use and redistribution of source data should respect applicable licenses and privacy.

### REPRODUCIBILITY STATEMENT

We specify the causal state, chunk duration, guidance decomposition, training objectives, evaluation populations, and streaming latency protocol in the main paper and appendix. The appendix also provides model configurations, training settings, and data composition.

### AI USE STATEMENT

We used Qwen3-Omni to filter audiovisual data and generate audio captions. Generative AI tools also assisted with writing, coding and manuscript review. The authors are responsible for all claims, experiments, citations, and artifacts.

## REFERENCES

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 813–824. PMLR, 2021. URL http://proceedings.mlr.press/v139/bertasius21a.html.

Xunliang Cai, Qilong Huang, Zhuoliang Kang, Hongyu Li, Shijun Liang, Liya Ma, Siyu Ren, Xiaoming Wei, Rixu Xie, and Tong Zhang. Longcat-video technical report. *CoRR*, abs/2510.22200, 2025. doi: 10.48550/ARXIV.2510.22200. URL https://doi.org/10.48550/arXiv.2510.22200.

João Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *CoRR*, abs/1907.06987, 2019. URL http://arxiv.org/abs/1907.06987.

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.

Zehua Chen, Junyou Wang, Yuxuan Jiang, Zhenying Fang, Yusheng Dai, Jianfei Chen, Ziwei Liu, and Jun Zhu. Visual representation matters: Exploiting temporal differences in video-to-audio generation. *CoRR*, abs/2608.04902, 2026. doi: 10.48550/ARXIV.2608.04902. URL https://doi.org/10.48550/arXiv.2608.04902.

Ziyang Chen, David F. Fouhey, and Andrew Owens. Sound localization by self-supervised time delay estimation. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXVI*, volume 13686 of *Lecture Notes in Computer Science*, pp. 489–508. Springer, 2022. doi: 10.1007/978-3-031-19809-0_28. URL https://doi.org/10.1007/978-3-031-19809-0_28.

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.

Jason Cramer, Ho-Hsiang Wu, Justin Salamon, and Juan Pablo Bello. Look, listen, and learn more: Design choices for deep audio embeddings. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pp. 3852–3856. IEEE, 2019. doi: 10.1109/ICASSP.2019.8682475. URL https://doi.org/10.1109/ICASSP.2019.8682475.

Yusheng Dai, Zehua Chen, Yuxuan Jiang, Baolong Gao, Qiuhong Ke, Jun Zhu, and Jianfei Cai. Omni2sound: Towards unified video-text-to-audio generation. *CoRR*, abs/2601.02731, 2026. doi: 10.48550/ARXIV.2601.02731. URL https://doi.org/10.48550/arXiv.2601.02731.

Yanbo Ding, Zhizhi Guo, Quanyue Song, Yishan He, Zhixiang He, Yongxiang Li, and Yali Wang. Ripple: Real-time streaming audio-video generation with cross-modal recurrent memory. *CoRR*, abs/2607.26818, 2026. doi: 10.48550/ARXIV.2607.26818. URL https://doi.org/10.48550/arXiv.2607.26818.

Nan Duan, Haoyang Huang, Weiyang Jin, Haoran Li, Yaowei Li, Yuming Li, Yijun Liu, Xin Lu, Xiaoxiao Ma, Yanwen Ma, Yaofeng Su, Yilang Sun, Haoyu Wang, Zeyue Xue, Songchun Zhang, and Junhao Zhuang. Long-horizon audio-visual generation for persistent stories and interactive worlds. *CoRR*, abs/2608.23383, 2026. doi: 10.48550/ARXIV.2608.23383. URL https://doi.org/10.48550/arXiv.2608.23383.

Zach Evans, Julian D. Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2025, Hyderabad, India, April 6-11, 2025*, pp. 1–5. IEEE, 2025. doi: 10.1109/ICASSP49660.2025.10888461. URL https://doi.org/10.1109/ICASSP49660.2025.10888461.

Pengjun Fang, Yingqing He, Yazhou Xing, Qifeng Chen, Ser-Nam Lim, and Harry Yang. Ac-foley: Reference-audio-guided video-to-audio synthesis with acoustic transfer. *CoRR*, abs/2603.15597, 2026. doi: 10.48550/ARXIV.2603.15597. URL https://doi.org/10.48550/arXiv.2603.15597.

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.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pp. 776–780. IEEE, 2017. doi: 10.1109/ICASSP.2017.7952261. URL https://doi.org/10.1109/ICASSP.2017.7952261.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 15180–15190. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01457. URL https://doi.org/10.1109/CVPR52729.2023.01457.

Alexi Gladstone, Heng Ji, and Yilun Du. Explorative modeling: Unlocking a third pretraining axis and end-to-end generation. *CoRR*, abs/2607.27372, 2026. doi: 10.48550/ARXIV.2607.27372. URL https://doi.org/10.48550/arXiv.2607.27372.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. Ltx-video: Realtime video latent diffusion. *CoRR*,

abs/2501.00103, 2025. doi: 10.48550/ARXIV.2501.00103. URL https://doi.org/10.48550/arXiv.2501.00103.

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.

Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, Baixin Xu, Hao-Xiang Guo, Kaixiong Gong, Cyrus Wu, Wei Li, Xuchen Song, Yang Liu, Eric Li, and Yahui Zhou. Matrix-game 2.0: An open-source, real-time, and streaming interactive world model. *CoRR*, abs/2508.13009, 2025. doi: 10.48550/ARXIV.2508.13009. URL https://doi.org/10.48550/arXiv.2508.13009.

Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. 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 X*, volume 15068 of *Lecture Notes in Computer Science*, pp. 289–305. Springer, 2024. doi: 10.1007/978-3-031-72684-2_17. URL https://doi.org/10.1007/978-3-031-72684-2_17.

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.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *CoRR*, abs/2207.12598, 2022. doi: 10.48550/ARXIV.2207.12598. URL https://doi.org/10.48550/arXiv.2207.12598.

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.

Tornike Karchkhadze, Kuan-Lin Chen, Mojtaba Heydari, Robert Henzel, Alessandro Toso, Mehrez Souden, and Joshua Atkins. Stereofoley: Object-aware stereo audio generation from video. *CoRR*, abs/2509.18272, 2025. doi: 10.48550/ARXIV.2509.18272. URL https://doi.org/10.48550/arXiv.2509.18272.

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.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 119–132. Association for Computational Linguistics, 2019. doi: 10.18653/V1/N19-1011. URL https://doi.org/10.18653/v1/n19-1011.

Jaeyeon Kim, Heeseung Yun, and Gunhee Kim. Visage: Video-to-spatial audio 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=8bF1Vaj9tm.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6114.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Daquan Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models. *CoRR*, abs/2412.03603, 2024. doi: 10.48550/ARXIV.2412.03603. URL https://doi.org/10.48550/arXiv.2412.03603.

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.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved RVQGAN. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/58d0e78cf042af5876e12661087bea12-Abstract-Conference.html.

Ke Lei, Yu Zhang, Changhao Pan, Xueyi Pu, Wenxiang Guo, Ruiqi Li, and Zhou Zhao. Towards streaming synchronized spatial audio generation via autoregressive diffusion transformer. *CoRR*, abs/2605.30940, 2026. doi: 10.48550/ARXIV.2605.30940. URL https://doi.org/10.48550/arXiv.2605.30940.

Yichong Leng, Zehua Chen, Junliang Guo, Haohe Liu, Jiawei Chen, Xu Tan, Danilo P. Mandic, Lei He, Xiangyang Li, Tao Qin, Sheng Zhao, and Tie-Yan Liu. Binauralgrad: A two-stage conditional diffusion probabilistic model for binaural audio synthesis. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/95f03faf3763e1b1ce2c3de62da8f090-Abstract-Conference.html.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. 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/66e226469f20625aaebddbe47f0ca997-Abstract-Conference.html.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo P. Mandic, Wenwu Wang, and Mark D. Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett

(eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 21450–21474. PMLR, 2023a. URL https://proceedings.mlr.press/v202/liu23f.html.

Huadai Liu, Kaicheng Luo, Jialei Wang, Wen Wang, Qian Chen, Zhou Zhao, and Wei Xue. Thinksound: Chain-of-thought reasoning in multimodal llms for audio generation and editing. 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*, 2025a. URL http://papers.nips.cc/paper_files/paper/2025/hash/710f3f8473b93394505a082f9a8f3ba2-Abstract-Conference.html.

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, 2025b. doi: 10.48550/ARXIV.2511.18833. URL https://doi.org/10.48550/arXiv.2511.18833.

Huadai Liu, Tianyi Luo, Kaicheng Luo, Qikai Jiang, Peiwen Sun, Jialei Wang, Rongjie Huang, Qian Chen, Wen Wang, Xiangtai Li, Shiliang Zhang, Zhijie Yan, Zhou Zhao, and Wei Xue. Omniaudio: Generating spatial audio from 360-degree video. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025c. URL https://proceedings.mlr.press/v267/liu25as.html.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023b. URL https://openreview.net/forum?id=XVjTT1nw5z.

Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/98c50f47a37f63477c01558600dd225a-Abstract-Conference.html.

Andrew Owens, Phillip Isola, Josh H. McDermott, Antonio Torralba, Edward H. Adelson, and William T. Freeman. Visually indicated sounds. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 2405–2413. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.264. URL https://doi.org/10.1109/CVPR.2016.264.

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.

Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Kumar Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, Jacob Chalk, Zhifan Zhu, Rhodri Guerrier, Fahd Abdelazim, Bin Zhu, Davide Moltisanti, Michael Wray, Hazel Doughty, and Dima Damen. HD-EPIC: A highly-detailed egocentric video dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 23901–23913. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02226. URL https://openaccess.thecvf.com/content/CVPR2025/html/Perrett_HD-EPIC_A_Highly-Detailed_Egocentric_Video_Dataset_CVPR_2025_paper.html.

Seyedmorteza Sadat, Otmar Hilliges, and Romann M. Weber. Eliminating oversaturation and artifacts of high guidance scales in 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=e2ONKX6qzJ.

Koichi Saito, Julian Tanke, Christian Simon, Masato Ishii, Kazuki Shimada, Zachary Novack, Zhi Zhong, Akio Hayakawa, Takashi Shibuya, and Yuki Mitsufuji. Soundreactor: Frame-level online video-to-audio generation. *CoRR*, abs/2510.02110, 2025. doi: 10.48550/ARXIV.2510.02110. URL https://doi.org/10.48550/arXiv.2510.02110.

Sizhe Shan, Qiulin Li, Yutao Cui, Miles Yang, Yuehai Wang, Qun Yang, Jin Zhou, and Zhao Zhong. Hunyuanvideo-foley: Multimodal diffusion with representation alignment for high-fidelity foley audio generation. *CoRR*, abs/2508.16930, 2025. doi: 10.48550/ARXIV.2508.16930. URL https://doi.org/10.48550/arXiv.2508.16930.

Noam Shazeer. GLU variants improve transformer. *CoRR*, abs/2002.05202, 2020. URL https://arxiv.org/abs/2002.05202.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seung Eun 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. *Trans. Mach. Learn. Res.*, 2026, 2026. URL https://openreview.net/forum?id=2NlGyqNjns.

Jianlin Su, Murtadha H. M. Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/J.NEUCOM.2023.127063. URL https://doi.org/10.1016/j.neucom.2023.127063.

Yaofeng Su, Yuming Li, Zeyue Xue, Jie Huang, Siming Fu, Haoran Li, Haoyang Huang, and Nan Duan. Omniforcing: Unleashing real-time joint audio-visual generation. In Paolo Favaro, Zuzana Kukelova, Atsuto Maki, Anna Rohrbach, Konrad Schindler, and Federico Tombari (eds.), *Computer Vision - ECCV 2026 - 19th European Conference, Malmö, Sweden, September 8-12, 2026, Proceedings, Part XIII*, volume 17013 of *Lecture Notes in Computer Science*, pp. 566–584. Springer, 2026. doi: 10.1007/978-3-032-37271-0_31. URL https://doi.org/10.1007/978-3-032-37271-0_31.

Peiwen Sun, Sitong Cheng, Xiangtai Li, Zhen Ye, Huadai Liu, Honggang Zhang, Wei Xue, and Yike Guo. Both ears wide open: Towards language-driven spatial audio generation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025a. URL https://openreview.net/forum?id=qPx3i9sMxv.

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, 2025b. doi: 10.48550/ARXIV.2512.14614. URL https://doi.org/10.48550/arXiv.2512.14614.

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

Step-Video Team. Step-video-t2v technical report: The practice, challenges, and future of video foundation model. *CoRR*, abs/2502.10248, 2025b. doi: 10.48550/ARXIV.2502.10248. URL https://doi.org/10.48550/arXiv.2502.10248.

Zeyue Tian, Yizhu Jin, Zhaoyang Liu, Ruibin Yuan, Xu Tan, Qifeng Chen, Wei Xue, and Yike Guo. Audiox: Diffusion transformer for anything-to-audio generation. *CoRR*, abs/2503.10522, 2025. doi: 10.48550/ARXIV.2503.10522. URL https://doi.org/10.48550/arXiv.2503.10522.

Michael Tschannen, Alexey A. Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa,

Olivier J. Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *CoRR*, abs/2502.14786, 2025. doi: 10.48550/ARXIV.2502.14786. URL https://doi.org/10.48550/arXiv.2502.14786.

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.

Ilpo Viertola, Vladimir Iashin, and Esa Rahtu. Temporally aligned audio for video with autoregression. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2025, Hyderabad, India, April 6-11, 2025*, pp. 1–5. IEEE, 2025. doi: 10.1109/ICASSP49660.2025.10890587. URL https://doi.org/10.1109/ICASSP49660.2025.10890587.

Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Xiaofeng Meng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *CoRR*, abs/2503.20314, 2025. doi: 10.48550/ARXIV.2503.20314. URL https://doi.org/10.48550/arXiv.2503.20314.

Xihua Wang, Yuyue Wang, Yihan Wu, Ruihua Song, Xu Tan, Zehua Chen, Hongteng Xu, and Guodong Sui. Tiva: Time-aligned video-to-audio generation. In Jianfei Cai, Mohan S. Kankanhalli, Balakrishnan Prabhakaran, Susanne Boll, Ramanathan Subramanian, Liang Zheng, Vivek K. Singh, Pablo César, Lexing Xie, and Dong Xu (eds.), *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, pp. 573–582. ACM, 2024. doi: 10.1145/3664647.3681027. URL https://doi.org/10.1145/3664647.3681027.

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.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=LQzN6TRFg9.

Biao Zhang, Paul Suganthan, Gaël Liu, Ilya Philippov, Sahil Dua, Ben Hora, Kat Black, Gus Martins, Omar Sanseviero, Shreya Pathak, Cassidy Hardin, Francesco Visin, Jiageng Zhang, Kathleen Kenealy, Qin Yin, Xiaodan Song, Olivier Lacombe, Armand Joulin, Tris Warkentin, and Adam Roberts. T5gemma 2: Seeing, reading, and understanding longer. *CoRR*, abs/2512.14856, 2025. doi: 10.48550/ARXIV.2512.14856. URL https://doi.org/10.48550/arXiv.2512.14856.

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.

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. *Int. J. Comput. Vis.*, 134(1):46, 2026b. doi: 10.1007/S11263-025-02649-3. URL https://doi.org/10.1007/s11263-025-02649-3.

# A Implementation Configuration and Architectural Details

Table 9 summarizes the concrete configuration and architectural choices of WorldSonus, supplementing the high-level system description in Section 3.

Table 9: System implementation and architectural configuration of WorldSonus.

| **Component** | Configuration and Architectural Specification |
|---|---|
| **Streaming Unit** | 100 ms chunk: 3 latent frames and 3 video frames at 30 FPS |
| **Causal Stereo VAE** | Frozen SoundReactor (Saito et al., 2025) VAE; maps 48 kHz stereo audio to 30 Hz latents |
| **AR Context Window** | Ring-KV cache over $W = 50$ chunks (5 s) |
| **AR Backbone** | 36 layers, width 1,024, 16 attention heads, SwiGLU (Shazeer, 2020) hidden size 4,096 |
| **Visual Encoder** | Frozen DINOv3 S+ (Siméoni et al., 2026); aspect-preserving resolution buckets; patch grid pooled $2\times$ |
| **Visual Adapter** | Linear bottleneck to 128 channels, concatenated with $\Delta\mathbf{G}_n$; 2D-RoPE (Su et al., 2024; Heo et al., 2024); 1 learned query per frame plus 1 chunk-summary query |
| **Text Prompt Encoder** | Frozen T5Gemma 2 (Zhang et al., 2025), truncated at 192 tokens |
| **Prompt Compressor** | 1 global token +31 learned queries, cached until the prompt updates |
| **Flow Head** | Width 1,024, SwiGLU hidden size 5,120; bidirectional attention within chunk; cross-attention to 3 frame tokens; AdaLN-Zero (Peebles & Xie, 2023) conditioned on $\mathbf{h}_t$ |
| **Sampler** | 15 Euler steps per 100 ms chunk |
| **Guidance** | AR stage: $s_v = s_p = 3.0$. Flow head: AR-summary scale 1.0, local-visual scale 8.0 with APG (Sadat et al., 2025) ($\eta = 0$, momentum $+0.25$, radius 0.124) |
| **ShiftNCE** | Training-only objective, using a frozen Synchformer (Iashin et al., 2024) teacher |

**AR token pairing.** At each step, the model receives the audio summary from the previous chunk and the visual summary from the current chunk. The audio summary is causally masked from attending to current or future visual tokens. Positional encoding uses chunk time indices via 2D-RoPE, maintaining continuous temporal representations when the Ring-KV buffer wraps.

**Guidance formulations.** Visual and textual controls are guided independently using classifier-free guidance (Ho & Salimans, 2022). Sampling evaluates a conditional branch $f$, a null-video branch $f_{\neg v}$, and a null-prompt branch $f_{\neg p}$:

$$\hat{f} = f + (s_v - 1)(f - f_{\neg v}) + (s_p - 1)(f - f_{\neg p}). \tag{4}$$

Guidance scales are listed in Table 9. On the flow head, APG (Sadat et al., 2025) is applied to the clean latent estimate $\hat{\mathbf{a}}_t = \mathbf{x}_{t,s} + (1 - s)\mathbf{u}$, where $\mathbf{u}$ is the predicted velocity.

**Baseline evaluations and clip windows.** For fair comparison, AudioX (Tian et al., 2025), ThinkSound (Liu et al., 2025a), and PrismAudio (Liu et al., 2025b) receive identical audio-centric captions as WorldSonus and are executed via their official inference pipelines with chain-of-thought reasoning disabled. V-AURA (Viertola et al., 2025) operates in video-only mode. On 5 s and 10 s benchmarks, all baseline models generate full-length clips in a single inference call. On the 30 s benchmark, AudioX and PrismAudio generate three contiguous 10 s clips matching their maximum context limits, whereas ThinkSound and V-AURA process the entire 30 s sequence in a single call. In contrast, WorldSonus continuously streams 100 ms chunks under bounded Ring-KV state with six sequential 5 s audio prompts. For 10 s caption-conditioned baseline runs, prompt pairs are concatenated in chronological order (e.g., "First five seconds: … Next five seconds: …").

# B Training Recipe and Optimization Details

Training consists of two distinct stages (Table 10). Full optimization hyperparameters are listed in Table 11. The audio VAE (Saito et al., 2025), DINOv3 S+ (Siméoni et al., 2026), T5Gemma 2 (Zhang et al., 2025), and the Synchformer (Iashin et al., 2024) teacher remain frozen throughout training.

Table 10: Training curriculum: corpora, task mixtures, and view lengths.

| Stage | Corpus Composition | Task Ratio (video-audio:audio-only) | View Representation |
|---|---|---|---|
| Pretraining (0–140k) | Full video-audio/audio-only mixture: base stereo datasets plus mono-dominant expansion | 2:1 | 5 s blocks, with 10 s paired views on 4 of 5 epochs |
| Fine-tuning (140–150k) | High-synchrony filtered stereo video-audio only (expansion excluded) | 1:0 | 5 s blocks, with 10 s paired views on 4 of 5 epochs |

Table 11: Training configuration and hyperparameter settings.

| **Hyperparameter** | Setting / Value |
|---|---|
| **Training Steps** | 140,000 pretraining steps +10,000 fine-tuning steps (150,000 total) |
| **Optimizer** | AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.99$, weight decay 0.05) |
| **Learning Rate** | Linear warmup from 0 to $5 \times 10^{-5}$ over 4,000 steps; constant $5 \times 10^{-5}$ thereafter |
| **Gradient Clipping** | Maximum gradient norm 1.0 |
| **EMA Rate** | Decay rate 0.9999 |
| **Distributed Setup** | 16 NVIDIA H100 GPUs (2 nodes); batch size 16 per GPU (256 global batch) |
| **Flow Batch Multiplier** | 4 independent flow timesteps sampled per audio instance |
| **Explorative Modeling** | $K = 3$ candidate source noise vectors per example; backpropagation through best candidate |
| **AR Causal Context** | 5.0 s sliding window (50 chunks) under teacher forcing |
| **Guidance Dropout** | Video dropout 0.1, prompt dropout 0.1; flow-head categorical dropout 0.2 (0.1 AR summary, 0.1 local visual); most recent teacher-forced audio token nulled with 0.2 probability |
| **Loss Objectives** | Joint Flow Matching ($\mathcal{L}_{\text{flow}}$) and Causal ShiftNCE ($\mathcal{L}_{\text{sync}}$) |

### B.1 Explorative Modeling for Flow Matching (XM3)

During pretraining, we apply Explorative Modeling (Gladstone et al., 2026) to the flow-matching objective. For each training instance, we sample $K = 3$ noises with the same timestep $s \sim \mathcal{U}[0, 1]$ and dropout mask, then compute their vector field errors without gradients:

$$\ell^{(k)} = \left\| \hat{\mathbf{u}}^{(k)} - (\mathbf{a}_t - \boldsymbol{\epsilon}^{(k)}) \right\|_2^2, \tag{5}$$

and execute backpropagation solely on $k^* = \arg\min_k \ell^{(k)}$. Standard Gaussian noise sampling is used during inference.

### B.2 Ambiguity-Aware Causal ShiftNCE

ShiftNCE distills synchronization signals from a frozen Synchformer (Iashin et al., 2024) teacher into the AR backbone during training, and is disabled during inference.

**Teacher representations.** The teacher processes a causal 640 ms visual window ending on a chunk boundary, outputting eight ordered tokens $\mathbf{S}_t$. A 2-layer MLP maps the AR representation $\mathbf{h}_t$ to this alignment space. Pairwise similarity averages cosine similarity across the eight token slots:

$$\text{sim}(P(\mathbf{h}_i), \mathbf{S}_j) = \frac{1}{8} \sum_{k=1}^{8} \cos\big(P(\mathbf{h}_i)_k, (\mathbf{S}_j)_k\big). \tag{6}$$

**Negative mining and anchor selection.** For chunk $t$, candidate negatives comprise same-clip temporal offsets $\delta \in \{\pm 1, \pm 2, \pm 4\}$. Candidates exhibiting high teacher similarity ($\text{sim}(\mathbf{S}_t, \mathbf{S}_{t+\delta}) \geq 0.97$) are filtered out to avoid false-negative penalization. Among valid sequence anchors, we select the top 40% most teacher-separable anchors ranked by contrastive margin.

**Objective formulation.** With a fixed temperature $\tau = 0.07$, the alignment loss is averaged uniformly over selected valid anchors $\mathcal{A}_t$:

$$\mathcal{L}_{\text{sync}} = -\frac{1}{|\mathcal{A}_t|} \sum_{i \in \mathcal{A}_t} \log \frac{\exp\left(\text{sim}(P(\mathbf{h}_i), \mathbf{S}_i)/\tau\right)}{\exp\left(\text{sim}(P(\mathbf{h}_i), \mathbf{S}_i)/\tau\right) + \sum_{\delta \in \mathcal{D}_i} \exp\left(\text{sim}(P(\mathbf{h}_i), \mathbf{S}_{i+\delta})/\tau\right)}. \quad (7)$$

**Gradient balancing.** The loss weight $\lambda_{\text{sync}}$ dynamically maintains the synchronization gradient norm at approximately $5\%$ of the flow matching gradient norm, bounded within $[0.02, 0.08]$ after a $5{,}000$-step linear warmup. For audio-only clips, $\mathcal{L}_{\text{sync}}$ is set to zero.

### B.3 Dynamic Prompt Schedules

To facilitate mid-stream steering during continuous generation sessions, each 10-s paired training sequence is assigned one of four prompt schedules:

- **Mid-stream switch** ($60\%$): Prompt $P_1$ conditions the first 5-s segment, switching to prompt $P_2$ at the 5-s boundary.
- **Instruction withdrawal** ($10\%$): Prompt $P_1$ conditions the first 5-s segment, followed by generation without text conditioning ($P = \emptyset$) for the second 5 seconds.
- **Instruction arrival** ($10\%$): Generation begins without text conditioning ($P = \emptyset$) for the first 5 seconds, followed by prompt $P_1$ for the remaining 5 seconds.
- **Sustained hold** ($20\%$): Prompt $P_1$ is maintained continuously across the full 10-s sequence.

## C Training Corpus and Data Curation

### C.1 Corpus Accounting and Composition

The curated training dataset comprises $993{,}920$ total entries amounting to $1{,}464.93$ hours of audio. Table 12 presents the complete breakdown across all four data pools. An independent held-out validation set of $6{,}106$ entries is isolated from training.

Table 12: Training corpus breakdown by pool and audio duration.

| Data Pool | 5-s Clips | 10-s Paired Views | Total Entries | Audio Duration (h) |
|---|---|---|---|---|
| Base stereo video-audio | 590,510 | 139,079 | 590,510 | 820.15 |
| Mono-dominant video-audio | 128,885 | 48,105 | 128,885 | 179.01 |
| Base audio-only | 66,859 | 60,832 | 127,691 | 261.84 |
| Recovered audio-only | 146,834 | – | 146,834 | 203.94 |
| Total Training Inventory | 933,088 | 60,832 | 993,920 | 1,464.93 |

Video-audio $10$ s paired views reuse existing $5$ s clips and are not counted again in the entry or duration totals. The total row's $10$ s column counts only the $60{,}832$ independent audio-only entries, which are included in both totals.

**Source-family breakdown.** The base stereo video-audio pool ($820.15$ h) combines four sources:

- *Open-stereo in-the-wild* ($183{,}271$ clips / $254.54$ h): VGGSound (Chen et al., 2020) and AudioSet (Gemmeke et al., 2017);
- *Interactive world model* ($135{,}478$ clips / $188.16$ h): gameplay and simulated scenes;
- *Panoramic ambisonics* ($151{,}298$ clips / $210.13$ h): Sphere360 (Liu et al., 2025c) ($112{,}502$ clips / $156.25$ h) and YT-AmbiGen (Kim et al., 2025) ($38{,}796$ clips / $53.88$ h); and
- *Action-centric and egocentric video* ($120{,}463$ clips / $167.31$ h): Kinetics-700 (Carreira et al., 2019) ($92{,}474$ clips / $128.44$ h) and HD-EPIC (Perrett et al., 2025) ($27{,}989$ clips / $38.87$ h).

The mono-dominant video-audio pool (179.01 h) supplements events from VGGSound (98,357 clips / 136.61 h) and AudioSet (30,528 clips / 42.40 h). These clips retain two-channel audio, with "mono-dominant" referring to duplicated-mono or narrow-stereo signals with limited channel separation. This expansion is screened with Qwen3-Omni and used only during pretraining. It is excluded from spatial fine-tuning. Base audio-only data combines AudioCaps (Kim et al., 2019) and IF-Caps (Tian et al., 2025) (127,691 clips / 261.84 h). The recovered audio-only pool (146,834 clips / 203.94 h) retains weakly-aligned clips lacking direct visual spatial cues but containing clean acoustic events.

### C.2 Signal-Level Stereo Screening and Ambisonic Decoding

The following signal-level screening applies to the base stereo video-audio pool, rather than the separately retained, Qwen3-Omni-screened mono-dominant expansion:

1. **Channel energy and DC centering**: Audio is resampled to 48 kHz and DC-centered per channel. Any clip exhibiting root-mean-square energy $\mathrm{RMS} < 10^{-4}$ on either channel is discarded as silent.
2. **Lateral energy contrast and correlation**: We compute the signed energy contrast $d = (E_R - E_L)/(E_R + E_L)$ and normalized inter-channel cross-correlation $\rho \in [-1, 1]$. Clips with $|\rho| > 0.98$ or stereo width $< 0.15$ are excluded from the base stereo video-audio pool.

**Ambisonic-to-stereo rendering.** For Sphere360 and YT-AmbiGen, the soundfield is rotated to align with the primary horizontal source and decoded at $\pm 45°$, matching a $90°$ perspective crop:

$$L = W + \frac{\sqrt{2}}{2}(X + Y), \qquad R = W + \frac{\sqrt{2}}{2}(X - Y). \tag{8}$$

### C.3 Multimodal Verification and Audio-Centric Captioning

Qwen3-Omni (Team, 2025a) scores audio-visual consistency from 1 to 5 and flags non-diegetic music. Video-audio clips below 4, or with such music, are removed. Captions name category, material, direction, reverberation, and envelope, and they omit visual appearance. Video-audio captions see the video and the audio, whereas audio-only captions see the audio only.

### C.4 Evaluation Set Isolation and Leakage Audit

Evaluation covers five main held-out test splits: VGGSound (5 s, 10 s) and Interactive Gameplay and Real-World (5 s, 10 s with 4,096 clips each; 30 s with 1,024 clips). We additionally evaluate on Greatest Hits (Owens et al., 2016) physical collisions (244 clips).

We audited the evaluation populations against our training inventory to verify strict data isolation:

- On VGGSound, our evaluation adheres to the official test split, ensuring clip-level isolation against the training corpus.
- On Interactive benchmarks, source-group checks confirmed zero overlap with the training and rollout inventories. The 4,096 clips in the 5-s and 10-s sets derive from 3,111 unique source groups, with at most two non-overlapping windows per source. The 30 s benchmark contains 1,024 distinct held-out sources.

## D Evaluation Protocols and Metric Specifications

### D.1 Objective Metrics

Metric $\mathrm{FD}_O$ computes the Fréchet Audio Distance directly on the mid-channel signal $(L + R)/2$, rather than averaging per-channel distances. Similarly, $\text{S-FD}_O$ evaluates the side-channel signal $(L - R)/2$. DeSync is the mean absolute audiovisual synchronization offset, measured in seconds. Lower values indicate better synchronization. Segment-aligned CLAP score details and BiasSkill specifications are presented in Appendix F and Appendix E, respectively. All other objective metrics utilize standard feature extractors as described in Section 5.

### D.2 Physical Onset Evaluation Protocol

Greatest Hits (Owens et al., 2016) onset settings follow MMAudio (Cheng et al., 2025): 244 clips, 8 s, 22.05 kHz; threshold $\delta = 0.3$; 50 ms pre/post-peak context; 300 ms matching tolerance.

### D.3 Hardware Benchmark and Streaming Latency Protocol

Inference latency is measured on a single NVIDIA H100 GPU using batch size 1 and 15 Euler sampling steps per chunk. Steady-state synthesis per 100 ms chunk requires 41.21 ms at p50 and 41.38 ms at p95, achieving a Real-Time Factor (RTF) of 0.41. Over a continuous 10 s streaming session including cold-start initialization, the overall RTF is 0.466. During interactive prompt updates, chunk latency increases slightly to 45.28 ms at p50 and 47.40 ms at p95 due to state cache refresh.

### D.4 Baseline Reproduction and Access Mode Analysis

We compare WorldSonus against state-of-the-art bidirectional stereo models (AudioX (Tian et al., 2025), ThinkSound (Liu et al., 2025a), PrismAudio (Liu et al., 2025b)) and the streaming monophonic AR baseline V-AURA (Viertola et al., 2025). All baselines are evaluated using official publicly released checkpoints, default samplers, and recommended classifier-free guidance parameters.

Baselines fall into three temporal access modes:

- **Bidirectional access (AudioX, ThinkSound, PrismAudio)**: These models require full-clip visual context prior to inference. They process video frames via full-sequence bidirectional attention, rendering them unsuitable for streaming or live-deployment scenarios.
- **Stream mode (V-AURA)**: For long-form generation, V-AURA operates with 640 ms strides over a 2.56 s visual window. Within this localized window, frame encoding via TimeSformer (Bertasius et al., 2021) remains non-causal, and DAC (Kumar et al., 2023) neural decoding incorporates future context.
- **Causal mode (WorldSonus)**: Visual conditioning is limited to past and current chunk frames, maintaining strict causal alignment as detailed in Section 3.

### D.5 ITD-Based Metrics and Limitations

In the production stereo used in our data, channel differences do not reliably reflect physical arrival-time delays. In our diagnostic experiment, duplicating a mono signal across both channels ($L = R$) achieved the lowest ITD-MSE (Chen et al., 2022) among the tested outputs despite having no stereo separation. We therefore omit ITD-MSE from our spatial evaluation. We also exclude FSAD (Sun et al., 2025a) because it uses ITD-related StereoCRW features. Instead, we report S-FD$_O$ and BiasSkill to evaluate stereo energy distribution and left/right channel agreement.

## E Stereo Balance Agreement (BiasSkill)

Stereo balance agreement (BiasSkill) measures whether generated and reference audio favor the same left/right channel within 1 s windows. It normalizes the gain over a permutation baseline to account for a model's fixed channel bias.

**Windowing and energy contrast.** Audio signals are resampled to 24 kHz and segmented into non-overlapping 1 s windows. Each window is DC-centered per channel. For channel mean-square energies $E_L = \frac{1}{|W|}\sum_{n\in W} x_L[n]^2$ and $E_R = \frac{1}{|W|}\sum_{n\in W} x_R[n]^2$, the left–right energy contrast is:

$$d = \frac{E_R - E_L}{E_R + E_L + 10^{-12}}. \tag{9}$$

A window is designated right-dominant if $d > 0$ and left-dominant if $d < 0$, with $|d|$ reflecting the magnitude of directional separation.

**Gating and eligibility criteria.** To prevent scoring ambiguous or silent intervals, a ground-truth window is considered eligible only if it satisfies:

$$|d_{\mathrm{GT}}| \geq 0.3, \qquad \mathrm{RMS}_{\mathrm{GT}} = \sqrt{\frac{E_L + E_R}{2}} \geq 10^{-4}. \tag{10}$$

For generated audio, we enforce an absolute activity threshold of $\mathrm{RMS}_{\mathrm{gen}} \geq 0.001$ ($-60$ dBFS). A generated window is marked as correct if and only if:

$$|d_{\mathrm{gen}}| \geq 0.1 \quad \text{and} \quad d_{\mathrm{gen}} \cdot d_{\mathrm{GT}} > 0. \tag{11}$$

Critically, near-mono generations ($|d_{\mathrm{gen}}| < 0.1$) that fail to establish stereo separation are counted as incorrect (retained in the evaluation denominator), preventing models from achieving artificially inflated accuracy by producing centered outputs.

**Balanced accuracy.** Accuracy is computed over eligible windows with ground-truth side $s \in \{L, R\}$ in clip $i$, macro-averaged across clips $C_s$, and balanced between channels:

$$A_s = \frac{1}{|C_s|} \sum_{i \in C_s} \frac{1}{|W_{i,s}|} \sum_{w \in W_{i,s}} \mathbb{1}[\hat{y}_{i,w} = s], \qquad A_{\mathrm{bal}} = \frac{1}{2}(A_L + A_R). \tag{12}$$

Table 13 summarizes the eligible window counts and retention rates across test sets.

Table 13: Eligible 1 s window counts and retention rates for BiasSkill evaluation.

| Set | Clips | Cand. Win. | Elig. Clips | Elig. Win. (L / R) | Keep (%) |
|---|---|---|---|---|---|
| VGGSound 10 s | 4,096 | 40,960 | 1,589 | 3,314 / 3,356 | 16.3 |
| Interactive 10 s | 4,096 | 40,960 | 2,275 | 6,759 / 3,316 | 24.6 |
| Interactive 30 s | 1,024 | 30,720 | 679 | 2,870 / 3,200 | 19.8 |
| Total | 9,216 | 112,640 | – | 12,943 / 9,872 | 20.3 |

**Permutation null and normalized BiasSkill.** Generative models often exhibit fixed channel biases stemming from training data or architecture (e.g., ThinkSound consistently generates right-heavy energy, whereas PrismAudio is left-heavy). To eliminate these baseline preferences, we draw 10,000 whole-clip derangements shared across all models, pairing generated audio clips with mismatched ground-truth videos while maintaining internal window alignment. The permutation null score $\mu_{\mathrm{null}}$ represents the expected balanced accuracy under these shuffled pairings. BiasSkill is defined as the normalized gain over this model-specific null:

$$\mathrm{BiasSkill} = \frac{A_{\mathrm{bal}} - \mu_{\mathrm{null}}}{1 - \mu_{\mathrm{null}}} \times 100\%. \tag{13}$$

A BiasSkill score of $0\%$ indicates performance no better than random chance under the model's fixed channel bias, while $100\%$ represents perfect left/right channel agreement on eligible windows. The one-sided Monte Carlo $p$-value is given by:

$$p = \frac{1 + \sum_{b=1}^{B} \mathbb{1}[A_{\mathrm{bal}}^{(b)} \geq A_{\mathrm{bal}}]}{B + 1}, \tag{14}$$

with $B = 10{,}000$. Table 14 reports the detailed stereo balance agreement results.

# F Interactive Text Control and Prompt-Switch Analysis

## F.1 CLAP Implementation and Segment Alignment

Semantic audio-text correspondence is evaluated using LAION-CLAP (Wu et al., 2023). Audio tracks are converted to mono at a 48 kHz sampling rate. Each waveform is divided into consecutive, non-overlapping 5 s segments and evaluated against its corresponding audio caption. Embeddings are extracted in FP16 precision and normalized in FP32. We report raw cosine similarities, macro-averaged within individual clips and across the test evaluation set.

Table 14: Detailed stereo balance agreement results under the $-60$ dBFS generation-RMS gate. L/R Rec. denotes per-side recall; Bal. Acc. is $A_{\text{bal}}$; RMS Cov. is the coverage of $\text{RMS}_{\text{gen}} \geq 0.001$; Act. Cov. requires both RMS and $|d_{\text{gen}}| \geq 0.1$; $\Delta = A_{\text{bal}} - \mu_{\text{null}}$; BiasSkill represents the normalized effect size (%); and $p$ indicates the one-sided permutation test value.

| Set | Model | L Rec. | R Rec. | Bal. Acc. | RMS Cov. | Act. Cov. | Null $\mu_{\text{null}}$ | $\Delta$ | BiasSkill (%) | $p$ |
|---|---|---|---|---|---|---|---|---|---|---|
| VGG 10 s | AudioX | 0.273 | 0.186 | 0.229 | 0.990 | 0.430 | 0.199 | $+0.030$ | 3.79 | 0.0002 |
| | ThinkSound | 0.181 | 0.386 | 0.283 | 0.956 | 0.554 | 0.260 | $+0.024$ | 3.22 | 0.0035 |
| | PrismAudio | 0.277 | 0.093 | 0.185 | 0.946 | 0.364 | 0.176 | $+0.009$ | 1.10 | 0.1052 |
| | WorldSonus (Ours) | 0.256 | 0.211 | 0.233 | 0.976 | 0.439 | 0.198 | $+0.035$ | **4.42** | 0.0001 |
| Int. 10 s | AudioX | 0.256 | 0.237 | 0.246 | 0.967 | 0.460 | 0.222 | $+0.024$ | 3.08 | 0.0001 |
| | ThinkSound | 0.159 | 0.395 | 0.277 | 0.983 | 0.594 | 0.290 | $-0.013$ | $-1.78$ | 0.9568 |
| | PrismAudio | 0.387 | 0.036 | 0.211 | 0.995 | 0.447 | 0.212 | $-0.001$ | $-0.06$ | 0.5390 |
| | WorldSonus (Ours) | 0.317 | 0.273 | 0.295 | 0.995 | 0.559 | 0.250 | $+0.045$ | **6.03** | 0.0001 |
| Int. 30 s | AudioX | 0.247 | 0.225 | 0.236 | 0.999 | 0.468 | 0.236 | 0.000 | $-0.01$ | 0.4974 |
| | ThinkSound | 0.101 | 0.455 | 0.278 | 0.997 | 0.559 | 0.273 | $+0.005$ | 0.63 | 0.3088 |
| | PrismAudio | 0.493 | 0.046 | 0.270 | 0.997 | 0.519 | 0.238 | $+0.032$ | 4.19 | 0.0004 |
| | WorldSonus (Ours) | 0.447 | 0.397 | 0.422 | 0.995 | 0.716 | 0.313 | $+0.109$ | **15.90** | 0.0001 |

### F.2 Paired Counterfactual Intervention Protocol

To rigorously verify that dynamic acoustic changes during streaming stem from active prompt steering rather than underlying visual transitions in the video, we construct a paired counterfactual intervention across all 4,096 clips in VGGSound 10 s and Interactive 10 s.

For each video clip with consecutive reference prompts $(P_1, P_2)$, we synthesize two paired generations sharing identical model parameters (150 k EMA checkpoint), visual feature inputs, random seeds, Euler solver settings (15 steps), and classifier-free guidance scales:

- **Switch Arm**: Conditioned on prompt $P_1$ for 0–5 s, dynamically switching to prompt $P_2$ at the 5 s boundary without clearing model state cache.
- **Hold Arm**: Conditioned continuously on prompt $P_1$ throughout the full 0–10 s duration.

Let $A_1$ and $A_2$ denote the generated 5 s audio segments, and let $s_{ij} = \cos(e_A(A_i), e_T(P_j))$ represent the pairwise cosine similarity between audio segment $A_i$ and text prompt $P_j$. Segment alignment $S$, matching advantage $M$, and dual relative match indicator $R$ are formulated as:

$$S = \frac{1}{2}(s_{11} + s_{22}), \tag{15}$$

$$M = \frac{1}{2}\big[(s_{11} - s_{12}) + (s_{22} - s_{21})\big], \tag{16}$$

$$R = \mathbb{1}[s_{11} > s_{12} \wedge s_{22} > s_{21}]. \tag{17}$$

The second segment's preference for the updated prompt is given by $\Delta = s_{22} - s_{21}$. The net paired intervention gain $G_i$ for clip $i$ is defined as:

$$G_i = \Delta_i^{\text{switch}} - \Delta_i^{\text{hold}}. \tag{18}$$

Table 15 summarizes the full paired intervention experimental results.

### F.3 Confidence Intervals and Prompt Stratification

Confidence intervals for the mean intervention gain $G$ are derived via 20,000 paired bootstrap resamples of clips within each benchmark dataset (random seed 5031). We report the 2.5th and 97.5th percentiles as $[l, u]$. The compact $\pm$ margin corresponds to half of this interval width.

To evaluate how semantic contrast influences prompt control responsiveness, clips are ranked by text-to-text cosine similarity $\cos(e_T(P_1), e_T(P_2))$ and partitioned into four equal quartiles of 1,024 clips each (Table 16). Quartile Q1 contains the most semantically distinct prompt pairs, while Q4 contains the most similar. As shown, $G$ remains positive across all strata and scales directly with the degree of prompt divergence.

Table 15: Full paired-control intervention results (4,096 clips per set). The $s_{ij}$ entries are raw CLAP cosine similarities. $S$ is the mean of the aligned similarities $s_{11}$ and $s_{22}$, $M$ is the mean matching advantage, and $R$ is reported as a percentage (%). Hold is evaluated against the same reference prompt pair $(P_1, P_2)$ as Switch.

| Set | Arm | $s_{11}$ | $s_{12}$ | $s_{21}$ | $s_{22}$ | $S$ | $M$ | $R$ (%) |
|---|---|---|---|---|---|---|---|---|
| VGG 10 s | Switch | 0.3823 | 0.3312 | 0.3293 | 0.3530 | 0.3677 | 0.0374 | 25.68 |
| | Hold | 0.3823 | 0.3312 | 0.3543 | 0.3270 | 0.3546 | 0.0119 | 14.43 |
| Int. 10 s | Switch | 0.3560 | 0.3062 | 0.3062 | 0.3284 | 0.3422 | 0.0360 | 23.44 |
| | Hold | 0.3560 | 0.3062 | 0.3316 | 0.3013 | 0.3286 | 0.0097 | 12.82 |

Table 16: Stratification of paired prompt gain $G$ across text-similarity quartiles (1,024 clips per quartile). Brackets denote 95% clip-level bootstrap percentile intervals.

| Evaluation Set | Quartile | Text Cosine Range | Intervention Gain $G$ |
|---|---|---|---|
| VGGSound 10 s | Q1 (Most Distinct) | $[-0.244, 0.525]$ | $0.1393\ [0.1287, 0.1499]$ |
| | Q2 | $[0.525, 0.727]$ | $0.0441\ [0.0388, 0.0494]$ |
| | Q3 | $[0.727, 0.842]$ | $0.0157\ [0.0127, 0.0187]$ |
| | Q4 (Most Similar) | $[0.842, 0.990]$ | $0.0047\ [0.0030, 0.0064]$ |
| Interactive 10 s | Q1 (Most Distinct) | $[-0.287, 0.490]$ | $0.1277\ [0.1187, 0.1367]$ |
| | Q2 | $[0.491, 0.688]$ | $0.0514\ [0.0459, 0.0569]$ |
| | Q3 | $[0.688, 0.820]$ | $0.0247\ [0.0210, 0.0284]$ |
| | Q4 (Most Similar) | $[0.820, 0.990]$ | $0.0063\ [0.0046, 0.0080]$ |

## G Subjective User Study Protocol

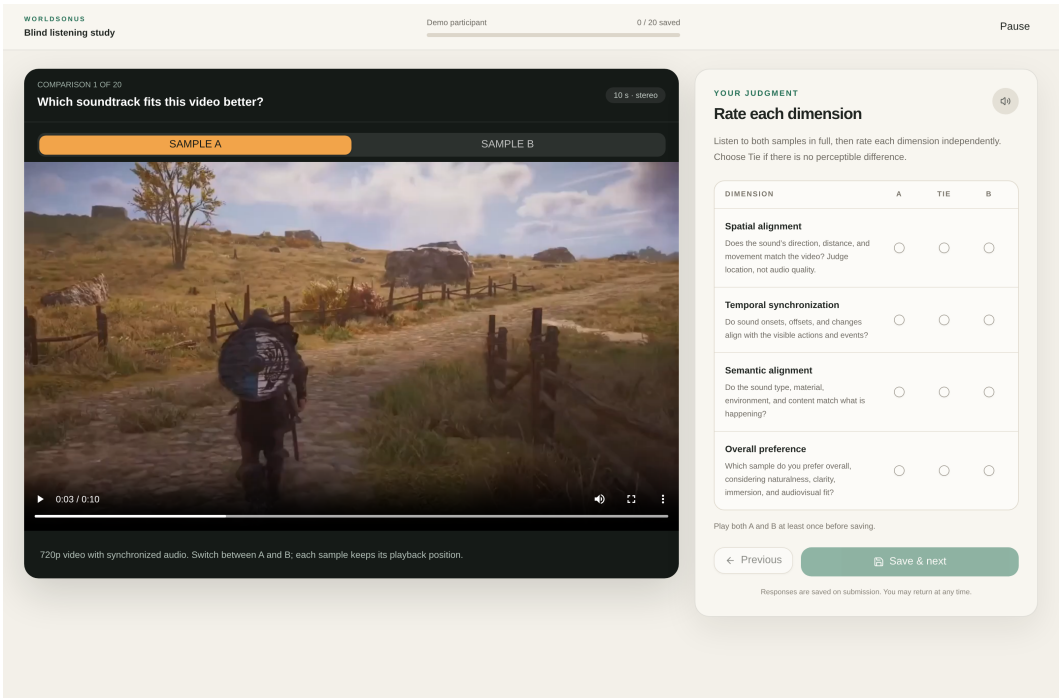
The user study evaluates 40 clips from Interactive 10 s, comprising 20 gameplay and 20 real-world scenes. Participants evaluate randomized, anonymized A/B pairs (with a tie option) across four dimensions: spatial alignment, temporal synchronization, semantic consistency, and overall preference. The five evaluated audio sources comprise Ground Truth, WorldSonus, AudioX, ThinkSound, and PrismAudio (V-AURA is excluded due to its monophonic output). Twenty independent evaluators produced a total of 400 pairwise evaluations, yielding 40 ratings per paired comparison.

Figure 4 shows the evaluation interface for gameplay and real-world clips, with paired samples labeled A and B and a separate A/Tie/B choice for each criterion.

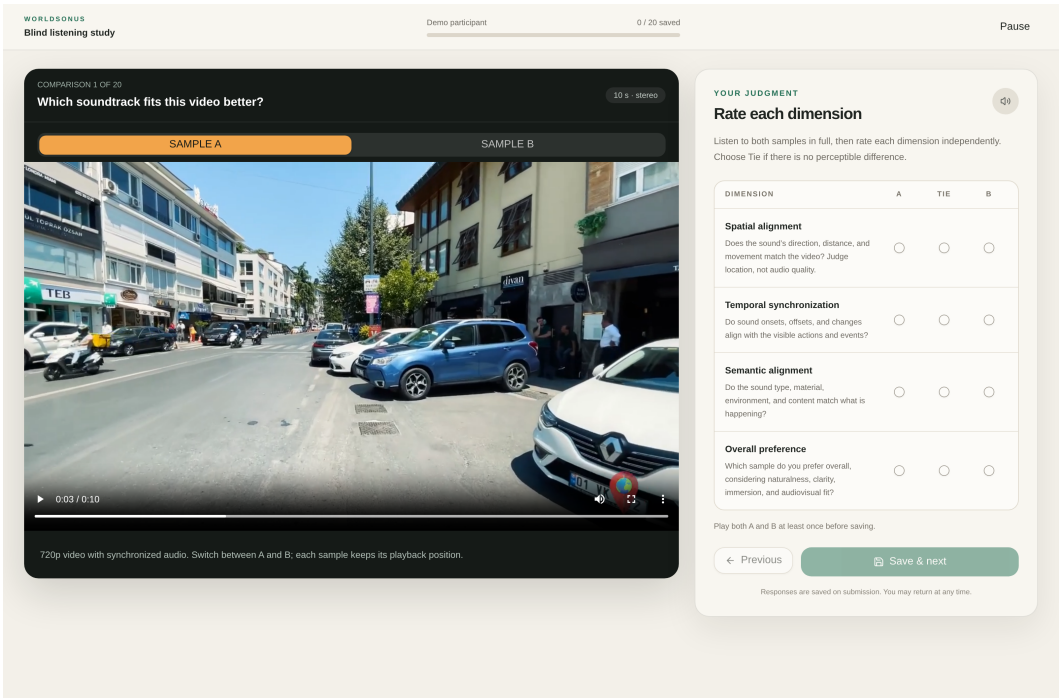
The preference score $P_{ij}^{(d)}$ of method $i$ over method $j$ on criterion $d$ is calculated as:

$$P_{ij}^{(d)} = 100 \times \frac{W_{ij}^{(d)} + \frac{1}{2} T_{ij}^{(d)}}{W_{ij}^{(d)} + T_{ij}^{(d)} + L_{ij}^{(d)}}, \tag{19}$$

where $W_{ij}^{(d)}$, $T_{ij}^{(d)}$, and $L_{ij}^{(d)}$ denote total wins, ties, and losses, respectively. Ties contribute half a vote (0.5), ensuring symmetric scoring $P_{ij}^{(d)} + P_{ji}^{(d)} = 100\%$. All evaluation matrices are reported on a normalized 0–100% scale.




(a) Gameplay example




(b) Real-world example

Figure 4: User-study interface for gameplay and real-world clips. Participants compare samples A and B and select A, Tie, or B for spatial alignment, temporal synchronization, semantic alignment, and overall preference.