UNIAMBI: UNIFIED SPATIAL AUDIO GENERATION VIA ITERATIVE EXPERT ROUTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial audio generation aims to synthesize spatial sound fields that faithfully reflect user intent under diverse conditioning signals. However, existing methods remain limited in their ability to generate accurate, high-fidelity spatial audio from heterogeneous inputs. To address this limitation, we propose UniAmbi, a unified framework for jointly synthesizing environmental audio and speech in first-order Ambisonics (FOA) format from multi-conditions. Our framework integrates three complementary designs to achieve efficient generation with semantic consistency, spatial accuracy, and perceptual fidelity. First, we introduce a Looped DiT backbone that addresses the complexity of multi-conditional FOA generation through iterative computation with shared parameters, increasing effective depth without a proportional increase in parameter count. Second, we introduce a loop aware sliding window Mixture-of-Experts (MoE) module to mitigate the representational limitations of strict parameter sharing. By incorporating expert routing on the current loop iteration and local temporal context, the module enables round-dependent transformations while encouraging consistent expert selection among neighboring tokens. Finally, we introduce per-loop on-policy self-distillation (OPSD), which converts generation-level feedback into direct supervision for predictions at every loop depth along the model's own generation trajectories. Experimental results show that our model outperforms existing methods on multiple spatial audio generation tasks. Demos can be found at `https://uniambi.github.io/`.

1 INTRODUCTION

Spatial audio characterizes both sound content and its three-dimensional distribution, providing more realistic auditory experiences for virtual reality, film, and immersive media Kim et al. (2025). To support flexible content creation, generative models must not only determine sound content and source directions but also allow users to express their creative intent through videos and descriptions, coordinating sound-event semantics, speech, audio quality, and spatial attributes within a shared sound field.

Previous works, including ViSAGe Kim et al. (2025) and OmniAudio Liu et al. (2025a), have explored generating first-order Ambisonics (FOA) directly from video. However, extending these approaches to unified multi-condition spatial audio generation requires jointly addressing constraints on both the input and output sides. On the input side, video provides information about visible sound sources and their dynamics, while text describes the overall audio content and specifies speech content; the two modalities can guide generation either jointly or independently. On the output side, sound effects and speech must maintain consistent temporal and spatial relationships. Compared with cascaded approaches that generate sounds independently before spatialization, end-to-end modeling allows these relationships to be learned jointly, reducing inter-module coordination and error propagation.

Achieving this goal presents two core challenges. First, complex conditioning interactions and coupled sound fields must be adequately modeled under a limited parameter budget. Expanding capacity by stacking independently parameterized layers also increases storage requirements and deployment costs. Second, a unified backbone must accommodate the modeling requirements of heterogeneous conditions and sound types. Speech, transient sound effects, and continuous ambient sounds exhibit distinct acoustic structures, requiring shared representations alongside specialization for local content and different stages of generation.

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

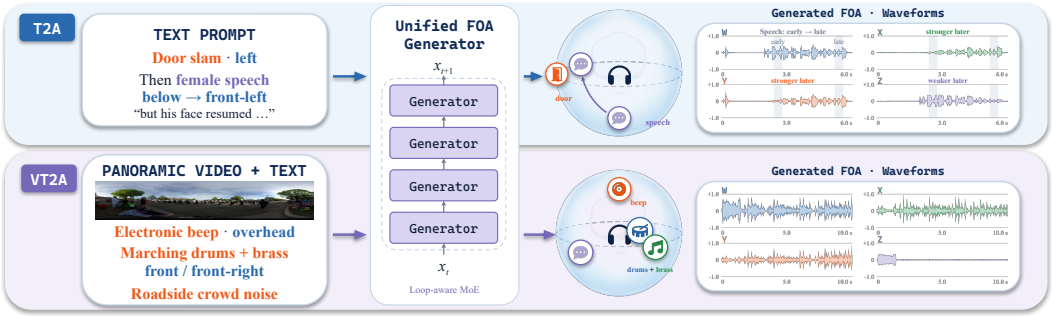

Figure 1: Text- and video–text-conditioned spatial audio generation. A shared generator produces FOA sound effects and speech from descriptions, transcripts, and optional video. Spheres illustrate the requested spatial layout; waveforms show generated W, X, Y, Z channels on an amplitude scale.

To address these challenges, we propose UNIAMBI, a unified framework for native FOA generation that supports T2FOA, V2FOA, and VT2FOA and jointly generates spatial sound effects and transcript-controlled speech. UNIAMBI uses a four-channel autoencoder to jointly represent FOA and integrates structured text with video features. For efficient modeling, we construct a Looped DiT that reuses DiT blocks for iterative refinement within each denoising step, increasing computational depth while maintaining a fixed parameter budget. To accommodate heterogeneity, we introduce a loop aware sliding window mixture-of-experts (MoE) module that dynamically selects experts based on the current loop, the center token, and its local context, providing specialized transformations for different conditioning constraints, sound events, and refinement stages. Figure 1 illustrates text-only and joint video–text generation of spatial sound effects and speech.

However, multiple loops of computation do not guarantee continued improvements in generation quality. Moreover, conventional denoising supervision is not directly aligned with final generation quality, while GRPO-style optimization based on terminal rewards lacks direct constraints on individual loop predictions. We therefore propose per-loop on-policy self-distillation (OPSD), which converts semantic, spatial, and perceptual feedback into training targets that each loop can directly fit. We further introduce a reward gate into intra-loop self-distillation (ILSD) Goyal et al. (2026), allowing cross-loop distillation only when later loops achieve a clear reward advantage, thereby encouraging progressive improvement.

Experiments demonstrate that UNIAMBI achieves competitive results across all three conditioning settings with only 385M parameters and outperforms the selected baselines on V2FOA. Compared with ReFL Xu et al. (2023), DiffusionNFT Zheng et al. (2026), and FlowGRPO Liu et al. (2025b), our per-loop OPSD with reward-gated ILSD achieves better audio quality, spatial and semantic consistency, and lower WER. Ablation studies further validate the contribution of each design.

Our main contributions are summarized as follows:

1. **A unified FOA generation framework.** UNIAMBI supports text, video, and joint conditioning within a single model to generate spatial sound effects and transcript-controlled speech.
2. **A parameter-efficient Looped DiT.** Iterative computation with shared parameters decouples effective computational depth from model size.
3. **A loop aware sliding window MoE.** Dynamic routing based on local context and loop index combines shared modeling with specialization for heterogeneous sounds.
4. **Generation-quality-driven per-loop post-training.** We combine per-loop OPSD with reward-gated ILSD to provide multidimensional generation-level feedback to internal loops.

2 RELATED WORK

Conditional spatial audio generation. Conditional spatial audio generation jointly models acoustic content and spatial attributes. SpatialSonic Sun et al. (2025) explores controllable stereo soundscapes from text and images. For native FOA, Diff-SAGe Kushwaha et al. (2025) conditions on sound

categories and source locations, ImmerseDiffusion Heydari et al. (2025) incorporates descriptions and spatial, temporal, and environmental parameters, and SonicMotion Templin et al. (2026) adds explicit control over moving sources. Visual approaches generate FOA from silent video, combine panoramic and perspective features, or incorporate reconstructed scene dynamics and acoustics Kim et al. (2025); Liu et al. (2025a); Luo et al. (2026b); Lei et al. (2026). Speech-oriented work includes image-conditioned reverberant TTS Zhang et al. (2025) and recovery of directional components from an existing omnidirectional speech track Luo et al. (2026a). These settings provide distinct forms of control through descriptions, visual cues, transcripts, and reference waveforms. UNIAMBI targets unified native FOA synthesis of sound effects and transcript-controlled speech under text, video, and joint video–text conditioning, allowing text to specify auditory content that video leaves undetermined.

Parameter sharing and mixtures of experts. Parameter sharing increases effective depth by repeatedly applying a compact set of transformations, while sparse mixtures of experts provide conditional capacity without activating all parameters for every token. In language modeling, MoEUT Csordás et al. (2024) combines depth recurrence with sparse feedforward and attention experts. LoopMoE Chen et al. (2026) couples shared iterative computation with sparse routing, using iteration- and token-conditioned modulation to differentiate recurrent updates. These studies establish parameter reuse and expert specialization as complementary choices. In visual generation, ELT Goyal et al. (2026) reuses Transformer blocks within each denoising step and supports different inference depths through intra-loop self-distillation (ILSD). Diff-MoE Cheng et al. (2025b) conditions experts on diffusion timesteps and allocates them across spatial tokens. The Thinking Pixel Sun et al. (2026) more directly combines recursive latent refinement with sparse adapter selection in diffusion attention modules. These approaches distinguish specialization across noise levels from repeated computation and expert selection within a denoising step. Building on these directions, UNIAMBI combines shared refinement with loop aware routing informed by local audio windows, accommodating different temporal events during multimodal FOA synthesis.

Generation feedback and self-distillation. Reward and preference feedback supplement denoising objectives with supervision tied to final generation quality. DDPO Black et al. (2024) optimizes downstream rewards with policy gradients, AlignProp Prabhudesai et al. (2023) directly backpropagates differentiable rewards, and Diffusion-DPO Wallace et al. (2024) learns from pairwise preferences. For audio, HarmoniDPO Peng & Zhang (2026) combines audiovisual correspondence, CLAP-based semantic consistency, and Audiobox-Aesthetics quality feedback for online preference optimization. A complementary line derives supervision from the model itself: Self Step-Distillation Ma et al. (2025) combines information across adjacent diffusion timesteps, whereas DiffusionOPSD Zhou et al. (2026b) constructs bounded, detached clean-output targets using reward gradients at queries from the model's own trajectories, then fits these targets with a finite update budget. Following this separation of target construction and fitting, we supervise every shared computation loop using FOA intensity-direction consistency, semantic alignment, and audio quality. Standard ILSD is used during pretraining, whereas reward-gated ILSD transfers knowledge across loops during per-loop OPSD post-training.

3 METHOD

3.1 FIRST-ORDER AMBISONICS AND FOA WAVVAE

First-order Ambisonics (FOA) is a multichannel audio format for representing three-dimensional sound fields. An FOA waveform of length L is represented as $\mathbf{x} \in \mathbb{R}^{4\times L}$, with four channels conventionally denoted by W, X, Y, and Z. The W channel represents the omnidirectional component, while X, Y, and Z encode first-order directional components along three orthogonal spatial axes, capturing front–back, left–right, and up–down variation, respectively. Unlike ordinary multichannel audio, the four FOA channels jointly describe a single sound field rather than independent audio signals. They largely share the semantic content of sound events, while source directions and spatial distributions are encoded through relative amplitudes, phases, and correlations across W, X, Y, Z.

Four-channel FOA WavVAE Existing FOA generation methods typically reuse VAEs trained on mono audio and encode W, X, Y, Z independently. However, the four FOA channels describe the

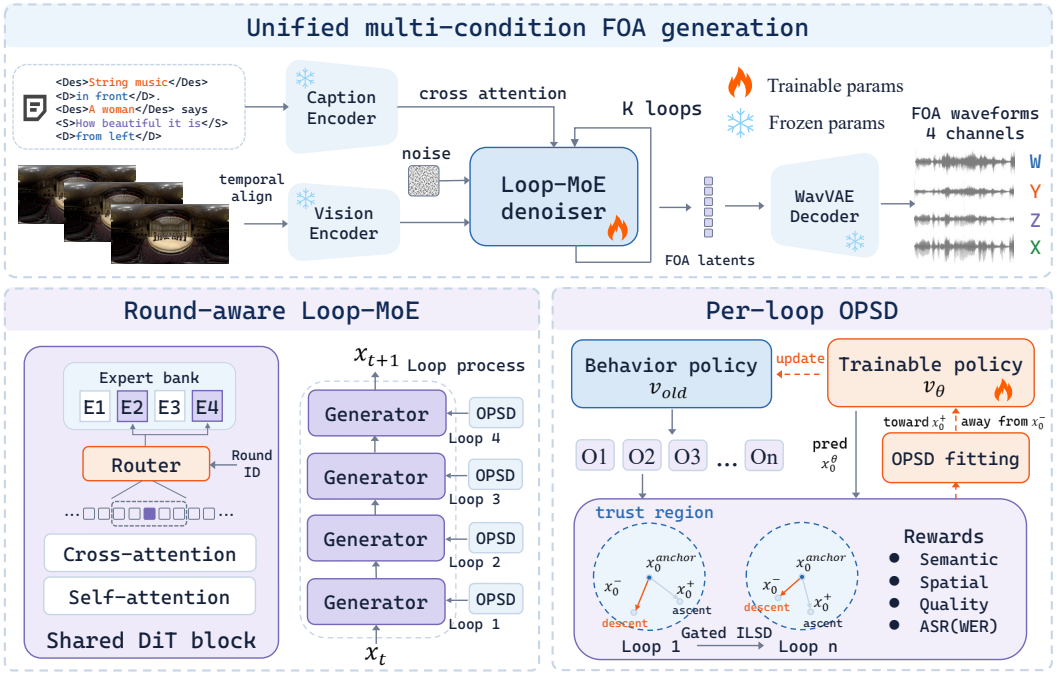

Figure 2: Unified multi-condition FOA generation. The shared backbone iterates within each denoising step, and per-loop OPSD applies generation feedback at every loop.

same acoustic events and differ primarily in their spatial encoding, so independent encoding Kumar et al. (2023) duplicates much of their semantic and acoustic information in the latent space. This introduces substantial semantic redundancy and increases the latent dimensionality at each time step. Independent encoding also disrupts the inherent coupling among FOA channels, requiring the downstream generative model to relearn these spatial relationships in the latent space and thereby increasing modeling difficulty.

We therefore train a four-channel FOA WavVAE to jointly encode and reconstruct the complete sound field. To preserve spatial information through encoding and reconstruction, we introduce two spatial objectives: an intensity loss and an axis-wise multi-resolution short-time Fourier transform (MR-STFT) loss. The intensity loss forms a three-dimensional intensity vector from the real parts of the cross-spectra between W and X, Y, Z in the time–frequency domain and minimizes the cosine distance between reconstructed and reference intensity directions. The axis-wise MR-STFT loss projects both sound fields toward the front, back, left, right, up, and down, and constrains reconstruction along each axis through spectral differences at multiple time–frequency resolutions. Initialized from the Stable Audio VAE, our VAE encodes 44.1 kHz FOA waveforms into 64-dimensional continuous latent vectors at a frame rate of 43 Hz. After training, the VAE is frozen for subsequent denoiser training and generation.

3.2 Unified spatial audio generation

We generate a four-channel FOA waveform $\mathbf{x}$ conditioned on $\mathbf{c} = (d, s, v, \mathbf{m})$, where d is an optional scene description, s an optional speech transcript, v an optional video, and $\mathbf{m}$ a condition tag identifying the modality type. UniAmbi supports text, video, and joint video–text conditioning: descriptions specify sound events and spatial attributes, while transcripts specify the words to be spoken. All data use consistent channel conventions, spherical-harmonic normalization, and spatial coordinates, with W denoting the omnidirectional component and X, Y, Z the three directional components.

As shown in Figure 2, UniAmbi comprises a four-channel FOA WavVAE, condition encoders, and a Looped DiT denoiser equipped with loop aware sliding window MoE. The model operates in the latent space of the VAE, where $\mathbf{z}_0 = \mathrm{Enc}(\mathbf{x})$ jointly represents the four FOA channels. Video is subsampled to 4 frames per second and encoded with CLIP, while captions are encoded using the

hidden states of Qwen3-0.6B Yang et al. (2025). Given these conditions, the model predicts FOA latents $\hat{\mathbf{z}}_0$, and the decoder reconstructs the sound field as $\hat{\mathbf{x}} = \mathrm{Dec}(\hat{\mathbf{z}}_0)$; we denote this decoder by D_ϕ below. To reflect the close temporal correspondence between video and audio, we proportionally repeat video tokens to match the audio-token sequence length and feed them into the backbone. Descriptions and transcripts are organized in a structured format and injected through cross-attention.

For a clean latent $\mathbf{z}_0$, we construct $\mathbf{z}_t = \alpha_t \mathbf{z}_0 + \sigma_t \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and α_t, σ_t define the noise schedule. Let $f_\theta^{(K)}$ denote the clean-latent prediction after K internal loops, converting noise or velocity predictions under the same schedule when needed. With noise-dependent weight $\lambda(t)$, pretraining minimizes

$$\mathcal{L}_{\mathrm{den}} = \mathbb{E}_{\mathbf{z}_0, \mathbf{c}, t, \boldsymbol{\epsilon}, K} \left[\lambda(t) \left\| f_\theta^{(K)}(\mathbf{z}_t, t, \mathbf{c}) - \mathbf{z}_0 \right\|_2^2 \right], \tag{1}$$

covering the intended condition combinations and loop depths.

Caption formulation We serialize scene descriptions and optional speech transcripts into a structured caption with three fields: `<Des>...</Des>` describes a sound event or speaker, `<S>...</S>` specifies the words to be spoken, and `<D>...</D>` describes the source direction. Each token receives a mark ID: 1, 2, and 3 identify the content within these respective fields, while 0 denotes padding, the tags themselves, and untagged narrative text. For example, `<Des>A man</Des> is saying <S>Hello.</S> <D>directly in front</D>.` explicitly separates the speaker description, spoken content, and direction. During caption dropout and in the unconditional branch of classifier-free guidance (CFG), an empty caption uses mark ID -1; this ID is temporarily mapped to 0 for embedding lookup, after which the corresponding mark embedding is explicitly zeroed.

We first train on large-scale text-to-FOA data to adapt the model to the FOA data distribution and enable the more challenging task of spatial speech generation. We then introduce video-to-FOA and joint video–text-to-FOA data while maintaining a proportion of speech data in the training mixture. This allows the model to progressively learn audio generation from video features while retaining its speech generation capability.

3.3 Loop aware sliding window MoE

Shared iterative refinement For fixed $(\mathbf{z}_t, t, \mathbf{c})$, the Looped DiT repeatedly applies a shared Transformer core to an initial sequence $\mathbf{h}^{(0)} = P_{\mathrm{in}}(\mathbf{z}_t)$:

$$\mathbf{h}^{(r)} = \mathcal{B}_\theta \left(\mathbf{h}^{(r-1)}, \mathbf{c}, \mathbf{e}(t), \mathbf{e}_r \right), \qquad r = 1, \ldots, K. \tag{2}$$

A shared output head yields the prediction at any loop, $f_\theta^{(r)} = P_{\mathrm{out}}(\mathbf{h}^{(r)})$. The embeddings $\mathbf{e}(t)$ and $\mathbf{e}_r$ encode the external denoising timestep and the internal loop index, respectively; internal loops do not advance sampler time. Repeating a shared B-block core K times yields an effective depth of KB while storing only one copy of the core parameters Goyal et al. (2026). The router and expert bank are also shared across loops, allowing local events to use different transformations without storing separate experts for each loop. Only k experts are activated per token per loop. Parameter sharing reduces stored parameters, but computation still grows with K: the number of block evaluations remains unchanged relative to an unshared model of the same effective depth.

Sliding window expert routing Strictly sharing feed-forward transformations can limit the updates required by different loops and local acoustic events. We therefore introduce a loop aware sliding window MoE within the shared core to increase its capacity to handle heterogeneous inputs. For an input audio token $\mathbf{u}_i^{(r)}$, a router scores the available experts, selects the top k to process the token, and combines their outputs using weights derived from the routing scores.

Prior MoE approaches to audio generation select experts independently for each audio token. However, adjacent audio frames are highly correlated, and sound events typically span multiple frames. Independent selection overlooks local temporal continuity and may produce overly fragmented expert assignments. SonicWeave Cai et al. (2026) addresses this issue with a chunk router that assigns tokens within a chunk to the same expert. Although this preserves some local continuity, it only coarsens the

granularity of expert selection and cannot accommodate sound events of variable duration. Building on this chunk-based design, we replace fixed chunks with sliding windows along the audio timeline, allowing the router to use information throughout each window when selecting experts.

3.4 PER-LOOP OPSD WITH SPATIAL, SEMANTIC, AND QUALITY FEEDBACK

Multi-dimensional feedback. A single feedback signal cannot fully characterize the spatial fidelity, semantic consistency, and perceptual quality of generated FOA audio. We therefore evaluate each generated waveform along these three complementary dimensions and combine them as

$$R_{\mathrm{fb}}(\hat{\mathbf{x}}) = \lambda_{\mathrm{sp}} R_{\mathrm{sp}}(\hat{\mathbf{x}}, \mathbf{x}^{\star}) + \lambda_{\mathrm{sem}} R_{\mathrm{sem}}(\hat{\mathbf{x}}, \mathbf{c}) + \lambda_{\mathrm{qual}} R_{\mathrm{qual}}(\hat{\mathbf{x}}), \tag{3}$$

where $\hat{\mathbf{x}}$ and $\mathbf{x}^{\star}$ denote the generated FOA waveform and its spatial reference, respectively, $\mathbf{c}$ denotes the available textual condition, and the three λ terms are non-negative weights.

The spatial feedback measures agreement in directional energy distribution while preserving the relative magnitude of the directional channels, thereby discouraging degenerate solutions that suppress spatial components. The semantic feedback measures whether the generated acoustic events and, when a reliable transcript is available, spoken content agree with the conditioning text. For video-conditioned generation tasks, we use ImageBind Girdhar et al. (2023) to provide a reward measuring audio–visual alignment. The perceptual-quality feedback evaluates complementary auditory attributes, including clarity, cleanliness, and overall production quality.

All evaluators remain frozen while retaining gradients with respect to the input waveform. The resulting feedback is used both to construct the OPSD target for each loop and to assess teacher reliability in cross-loop distillation. Further details are provided in the appendix.

Per-loop OPSD with reward-gated ILSD. Standard OPSD optimizes a single model prediction at each denoising state (Zhou et al., 2026a), whereas a looped denoiser produces a sequence of internal predictions while refining the same noisy latent. Although earlier loops influence subsequent computation, supervising only the final loop does not directly constrain the intermediate predictions. We therefore apply the same on-policy feedback to every loop and construct a separate OPSD target for each prediction.

Let $\mathcal{L}_{\mathrm{OPSD}}^{(r)}$ denote the standard OPSD loss applied to loop r. We average the per-loop losses so that the overall feedback strength remains independent of the number of supervised loops: $\mathcal{L}_{\mathrm{pOPSD}} = \frac{1}{K}\sum_{r=1}^{K} \mathcal{L}_{\mathrm{OPSD}}^{(r)}$

During flow-matching pretraining, intra-loop self-distillation (ILSD) supervises an intermediate loop using both the flow-matching target and the detached final-loop prediction. Once per-loop OPSD is introduced, each loop already receives an independent reward-derived target. We therefore remove the additional intermediate-to-target term during post-training and retain only cross-loop knowledge transfer.

Standard ILSD implicitly assumes that the deeper loop is always a reliable teacher. Additional recurrent computation, however, does not necessarily increase the multidimensional feedback for every sample. Unconditional distillation may therefore transfer a deeper prediction that is no better than the shallower one. We instead use the feedback difference between loops to determine when the deeper prediction provides a reliable teacher.

For sample i, let $\mathbf{a}_{r,i,g}$ denote the detached behavior-model anchor produced by loop r on trajectory g. We average the feedback difference between the second and first loops over G online trajectories and apply distillation only when this difference exceeds a margin δ:

$$\begin{aligned} \Delta R_i &= \frac{1}{G}\sum_{g=1}^{G} \left[R_{\mathrm{fb}}(D_\phi(\mathbf{a}_{2,i,g})) - R_{\mathrm{fb}}(D_\phi(\mathbf{a}_{1,i,g})) \right], \\ \mathcal{L}_{\mathrm{gKD}} &= \beta\, \mathbb{E}_i \left[\mathbf{1}[\Delta R_i > \delta] \left\| \mathbf{v}_\theta^{(1)} - \mathrm{sg}\left(\mathbf{v}_\theta^{(2)}\right) \right\|_2^2 \right]. \end{aligned} \tag{4}$$

The first loop is therefore distilled only when the second loop achieves a clear feedback improvement. Per-loop OPSD and reward-gated ILSD are used only during post-training and introduce no feedback evaluators at inference time. Further training details are provided in the appendix.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Evaluation metrics. We evaluate distributional fidelity with PaSST-based FD and KL Koutini et al. (2021) and VGGish-based FAD Hershey et al. (2017). For text-conditioned samples, CLAP Wu et al. (2023) measures text–audio alignment; for samples with transcripts, WER is computed from Whisper-large-v3 Radford et al. (2023) predictions. Lower FD, KL, FAD, and WER and higher CLAP indicate better performance.

For spatial consistency, ViSAGe's CC-5fps and CC-all compare generated and reference directional energy maps at five frames per second and over the first five seconds, respectively. PSELDNets Hu et al. (2025) additionally extracts a 3D event vector for each frame and class. Our weighted cosine similarity (wCS) averages the vector-pair cosine similarities using reference magnitude and valid-frame coverage as weights, normalizing within each clip before averaging across clips.

Data construction. We use pyroomacoustics to spatialize and mix single-source recordings from RWCP Nakamura et al. (2000), NIGENS Trowitzsch et al. (2020), and FSDKaggle2019 Fonseca et al. (2020), constructing single- and multi-source FOA examples. We additionally use a multimodal large language model (MLLM) Gemini-3.1-pro to produce structured content descriptions for AudioSet recordings and spatialize the audio, expanding the paired text–FOA data. For speech, we use VoxBox Wang et al. (2025) recordings spatialized through the same pyroomacoustics-based pipeline. In total, we obtain approximately 2.8M single-event audio samples and 1.7M multi-event audio mixtures.

For video-conditioned training, following Sphere360 and YT-Ambigen, we collect paired panoramic videos and spatial audio from YouTube and annotate them with an MLLM using both modalities to describe sound semantics and source directions. These annotations provide paired data for video, text, and joint conditioning. Annotation examples are provided in Appendix A. To enable text to specify off-screen sounds during joint video–text generation, we also augment the original video examples by randomly adding speech or other audio to their soundtracks and including the corresponding descriptions in the captions. The final dataset contains approximately 350K video–text–FOA triplets.

Baselines. For T2FOA, we compare with AudioLDM2 Liu et al. (2024), Tango2 Majumder et al. (2024), and MMAudio Cheng et al. (2025a). V2FOA baselines are ViSAGe and OmniAudio, which generate FOA from silent panoramic video. For VT2FOA, we use MMAudio, AudioX-MAF Tian et al. (2026), and Omni2Sound Dai et al. (2026), all of which support joint video–text conditioning. We report only metrics available under comparable task settings.

4.2 MAIN RESULTS

Table 1 examines whether UNIAMBI remains competitive across text, video, and joint video–text conditioning while retaining a compact parameter budget. Because the available T2FOA and VT2FOA baselines do not report FOA spatial metrics, comparisons in these two settings are restricted to distributional fidelity and text–audio alignment; V2FOA provides the direct comparison of both distributional and spatial fidelity.

Direct comparison on V2FOA. V2FOA provides the most complete comparison because all methods generate FOA and report both distributional and spatial metrics. Relative to OmniAudio, the stronger baseline overall, the reported UNIAMBI results reduce FD, KL, and FAD by 15.3%, 18.3%, and 31.0%, respectively, while increasing wCS, CC-5fps, and CC-all by 28.1%, 15.9%, and 4.1%. This provides the most complete evidence in Table 1 that the proposed approach is competitive under both distributional and spatial criteria.

Generality across text-bearing conditions. For T2FOA, the reported UNIAMBI results rank first on KL, FAD, and CLAP and second on FD. Relative to the strongest baseline for each metric, KL and FAD decrease by 20.2% and 1.6%, and CLAP increases by 4.0%; MMAudio retains a slightly lower FD (262.588 versus 272.823). For VT2FOA, UNIAMBI ranks first on FD, KL, and CLAP, improving them by 7.5%, 15.7%, and 3.2% over AudioX-MAF, but its FAD remains higher than all

Table 1: Main comparisons for FOA generation under text (T2FOA), video (V2FOA), and joint video–text (VT2FOA) conditioning. Param. denotes the number of model parameters. FD and KL use PaSST, and FAD uses VGGish. CLAP measures text–audio semantic alignment, while wCS, CC-5fps, and CC-all evaluate spatial consistency. Best and second-best values within each task are bolded and underlined.

Task	Model	Param.	FD ↓	KL ↓	FAD ↓	CLAP ↑	wCS ↑	CC-5fps ↑	CC-all ↑
T2FOA	AudioLDM2	712M	362.648	3.166	<u>2.768</u>	0.131	–	–	–
	Tango2	866M	716.849	2.980	9.383	0.155	–	–	–
	MMAudio	621M	**262.588**	<u>2.844</u>	4.649	<u>0.165</u>	–	–	–
	Ours	385M	<u>272.823</u>	**2.269**	**2.725**	**0.171**	0.361	0.842	0.912
V2FOA	ViSAGe	360M	731.250	2.438	11.620	–	0.080	0.291	0.380
	OmniAudio	1.2B	<u>668.860</u>	<u>2.280</u>	<u>10.514</u>	–	<u>0.143</u>	<u>0.329</u>	<u>0.423</u>
	Ours	385M	**566.415**	**1.863**	**7.253**	–	**0.184**	**0.381**	**0.440**
VT2FOA	MMAudio	621M	570.460	1.931	<u>4.230</u>	0.202	–	–	–
	AudioX-MAF	1.1B	<u>505.720</u>	<u>1.685</u>	4.338	<u>0.214</u>	–	–	–
	Omni2Sound	1.33B	531.840	1.864	**3.894**	0.203	–	–	–
	Ours	385M	**467.703**	**1.420**	6.363	**0.221**	0.193	0.385	0.475

three baselines (6.363 versus 3.894 for the best baseline). These results support broad coverage of text-only and joint conditioning in semantic alignment and PaSST-based fidelity, while also exposing a remaining mismatch in the VGGish feature space for joint conditioning. Because comparable spatial scores are unavailable for these baselines, the absolute wCS and CC values establish evaluation coverage but not comparative spatial superiority.

Parameter–quality trade-off. UNIAMBI uses 385M parameters, 38.0% fewer than the smallest T2FOA and VT2FOA baseline (MMAudio, 621M). On V2FOA, its parameter count is close to ViSAGe (360M) and 67.9% lower than OmniAudio (1.2B), while attaining stronger reported results across all comparable metrics. This comparison establishes a favorable performance–parameter trade-off rather than lower inference cost: the looped backbone reuses parameters across rounds but retains the corresponding block evaluations. The loop-configuration study below further analyzes the quality trade-offs introduced by parameter sharing under a fixed effective depth.

4.3 ABLATIONS

Ablation settings. Table 2 evaluates expert routing and post-training variants, while Table 3 compares loop configurations at a fixed effective depth. w/o SW removes sliding window routing context; w/o MoE removes sparse experts; Final-loop supervises only the last internal prediction; and Ungated ILSD removes reward-based teacher filtering. The aggregated scores summarize overall rather than task-wise behavior.

Expert capacity and temporal routing. Removing MoE degrades all eight aggregate metrics, including FD (402.825 to 499.399), FAD (4.966 to 6.113), and CLAP (0.192 to 0.145), supporting the need for conditional expert capacity in the shared backbone. Removing sliding window context has a mixed effect: six metrics worsen, whereas CC-5fps and CC-all improve slightly. The window therefore improves the overall balance, particularly wCS, rather than every spatial criterion.

Parameter sharing across loops. At a fixed 24-layer effective depth, $2 \times 12L$ is best on KL, CLAP, CC-5fps, and CC-all, while the non-looped DiT is best on FD, FAD, wCS, and WER. Three or four loops degrade most metrics relative to two loops. Thus, two loops provide the most balanced shared configuration rather than uniformly outperforming the dense model; parameter savings require separate reporting because Table 3 does not include model size or computation.

Loop-wise supervision and reward gating. The reported UNIAMBI results are numerically better than Final-loop and Ungated ILSD on all eight aggregate metrics. The largest differences are in WER (0.171 to 0.115 versus Final-loop; 0.163 to 0.115 versus Ungated ILSD) and in FD/FAD. These

Table 2: Architecture and training-method ablations across T2FOA, V2FOA, and VT2FOA. FD, KL, and FAD are sample-count-weighted means of task-level scores; CLAP, wCS, and CC are averaged over their valid samples. Best and second-best values within each group are bolded and underlined.

Method	FD ↓	KL ↓	FAD ↓	CLAP ↑	wCS ↑	CC-5fps ↑	CC-all ↑	WER ↓
Architecture components								
Ours	**402.825**	**1.900**	**4.966**	**0.192**	**0.264**	0.569	0.650	**0.115**
w/o SW	438.566	1.918	5.556	0.187	0.245	**0.570**	**0.665**	0.132
w/o MoE	499.399	2.108	6.113	0.145	0.240	0.547	0.572	0.161
Training methods								
Ours	**402.825**	**1.900**	4.966	**0.192**	**0.264**	**0.569**	0.650	**0.115**
Final-loop	426.677	1.962	5.481	0.191	0.249	0.562	0.640	0.171
Ungated ILSD	426.118	1.943	5.317	0.189	0.260	0.567	0.640	0.163
ReFL	419.453	1.925	5.546	0.189	0.256	0.562	0.652	0.194
DiffusionNFT	428.486	1.935	5.535	0.186	0.253	0.546	0.623	0.164
FlowGRPO	439.917	1.968	**4.945**	0.182	0.249	0.561	**0.656**	0.185

Table 3: Ablation study of loop configurations. All configurations have an effective depth of 24 layers. Looped configurations are reported as the number of loops $\times$ the number of layers per loop. Best and second-best values are bolded and underlined.

Configuration	FD ↓	KL ↓	FAD ↓	CLAP ↑	wCS ↑	CC-5fps ↑	CC-all ↑	WER ↓
DiT (24 layers)	**409.133**	1.940	**5.138**	0.186	**0.265**	0.582	0.656	**0.079**
$2 \times 12L$	428.280	**1.914**	5.549	**0.196**	0.260	**0.586**	**0.668**	0.156
$3 \times 8L$	457.107	2.058	6.708	0.180	0.263	0.548	0.623	0.120
$4 \times 6L$	507.717	2.271	7.234	0.157	0.245	0.555	0.637	0.112

comparisons support both intermediate supervision and reward-based teacher filtering for final-output quality, but do not establish monotonic improvement across internal loops.

Alternative post-training methods. UniAmbi is numerically best on six of eight aggregate metrics, while FlowGRPO is best on FAD (4.945 versus 4.966) and CC-all (0.656 versus 0.650). This configuration therefore offers the strongest overall balance rather than universal dominance; without uncertainty estimates, the two small gaps should not be interpreted as statistically significant.

5 Limitations and Conclusion

Conclusion We present UniAmbi, a unified framework for FOA generation that supports spatial sound effects and transcript-controlled speech under text, video, and joint video–text conditioning. Built on a four-channel FOA WavVAE, the Looped DiT increases effective depth through iterative computation with shared parameters, while the loop aware sliding window MoE uses local temporal context to specialize transformations across sound events and refinement stages. We further combine per-loop OPSD with reward-gated ILSD to refine intermediate predictions using feedback on spatial accuracy, semantic consistency, and perceptual quality. Experimental results demonstrate that UniAmbi balances parameter efficiency and generation quality across multiple conditioning settings. Ablations further support expert routing and per-loop post-training.

Limitations Shared looping reduces the number of model parameters, but computational cost still increases with loop depth. Improved parameter efficiency therefore does not necessarily translate into lower inference latency. Moreover, current large models remain limited in their ability to accurately infer source directions from spatial audio. Our feedback consequently evaluates global semantic content and spatial distributions separately, making it difficult to assess the correctness of individual sound events and their associated directions. Developing fine-grained feedback that jointly captures sound events and their spatial attributes remains a direction for future research.

AI USE STATEMENT

Generative AI tools were used to assist with audio-visual data annotation and synthetic dataset construction, as detailed in Section 4.1 and Appendix A. These tools were also used to polish the language of the manuscript. The authors conducted manual spot-checks of the annotations and constructed data, and reviewed the AI-assisted manuscript text for accuracy. The authors assume full responsibility for the manuscript, experimental results, and associated research artifacts.

REFERENCES

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, volume 2024, pp. 4965–4987, 2024.

Yunrui Cai, Xu Li, Yucheng Zhou, Jinchao Li, Dingdong Wang, Dongchao Yang, Xixin Wu, Chen Zhang, Zhiyong Wu, Pengfei Wan, et al. Sonicweave: Chunk-routed mixture-of-experts for unified audio scene generation. *arXiv preprint arXiv:2608.09571*, 2026.

Wenkai Chen, Tianshu Li, Wenyong Huang, Yichun Yin, Lifeng Shang, and Chengwei Qin. Loopmoe: Unifying iterative computation with mixture-of-experts for language modeling. *arXiv preprint arXiv:2606.04438*, 2026.

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. MMAudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *CVPR*, 2025a.

Kun Cheng, Xiao He, Lei Yu, Zhijun Tu, Mingrui Zhu, Nannan Wang, Xinbo Gao, and Jie Hu. Diff-moe: Diffusion transformer with time-aware and space-adaptive experts. In *Forty-second International Conference on Machine Learning*, 2025b.

Róbert Csordás, Kazuki Irie, Jürgen Schmidhuber, Christopher Potts, and Christopher D Manning. Moeut: Mixture-of-experts universal transformers. *Advances in Neural Information Processing Systems*, 37:28589–28614, 2024.

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

Eduardo Fonseca, Manoj Plakal, Frederic Font, Daniel P. W. Ellis, and Xavier Serra. Audio tagging with noisy labels and minimal supervision, 2020. URL https://arxiv.org/abs/1906.02975.

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 *CVPR*, 2023.

Sahil Goyal, Swayam Agrawal, Gautham Govind Anil, Prateek Jain, Sujoy Paul, and Aditya Kusupati. Elt: Elastic looped transformers for visual generation. In *European Conference on Computer Vision*, pp. 308–326. Springer, 2026.

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pp. 131–135. IEEE, 2017.

Mojtaba Heydari, Mehrez Souden, Bruno Conejo, and Joshua Atkins. Immersediffusion: A generative spatial audio latent diffusion model. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Jinbo Hu, Yin Cao, Ming Wu, Fang Kang, Feiran Yang, Wenwu Wang, Mark D Plumbley, and Jun Yang. Pseldnets: Pre-trained neural networks on a large-scale synthetic dataset for sound event localization and detection. *IEEE Transactions on Audio, Speech and Language Processing*, 33: 2845–2860, 2025.

Jaeyeon Kim, Heeseung Yun, and Gunhee Kim. ViSAGe: Video-to-spatial audio generation. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2506.12199.

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

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved RVQGAN. *arXiv preprint arXiv:2306.06546*, 2023. URL https://arxiv.org/abs/2306.06546.

Saksham Singh Kushwaha, Jianbo Ma, Mark RP Thomas, Yapeng Tian, and Avery Bruni. Diff-sage: End-to-end spatial audio generation using diffusion models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

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. *arXiv preprint arXiv:2605.30940*, 2026. URL https://arxiv.org/abs/2605.30940.

Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining, 2024. URL https://arxiv.org/abs/2308.05734.

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 *International Conference on Machine Learning*, 2025a. URL https://arxiv.org/abs/2504.14906.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl, 2025b. URL https://arxiv.org/abs/2505.05470.

Qingyu Luo, Peng Zhang, Wenwu Wang, and Philip JB Jackson. Visually-guided spatial audio generation for 360$\backslash$circ $in-the-wildspeechscenes.$ *arXiv preprint arXiv:2608.24579*, 2026a.

Ziyu Luo, Lin Chen, Qiang Qu, Xiaoming Chen, and Yiran Shen. Dynfoa: Generating first-order ambisonics with conditional diffusion for dynamic and acoustically complex 360-degree videos. *arXiv preprint arXiv:2602.06846*, 2026b.

Xinyin Ma, Runpeng Yu, Songhua Liu, Gongfan Fang, and Xinchao Wang. Diffusion model is effectively its own teacher. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12901–12911. IEEE, 2025.

Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya Poria. Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization, 2024. URL https://arxiv.org/abs/2404.09956.

Satoshi Nakamura, Kazuo Hiyane, Futoshi Asano, Takanobu Nishiura, and Takeshi Yamada. Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition. In M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis, and G. Stainhauer (eds.), *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece, May 2000. European Language Resources Association (ELRA). URL https://aclanthology.org/L00-1264/.

Wenshuo Peng and Kaipeng Zhang. Harmonidpo: Video-guided audio generation via preference-optimized diffusion: W. peng, k. zhang. *International Journal of Computer Vision*, 134(2):77, 2026.

Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. 2023.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

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 *International Conference on Learning Representations*, volume 2025, pp. 42606–42649, 2025.

Yuwei Sun, Yuxuan Yao, Hui Li, and Siyu Zhu. The thinking pixel: Recursive sparse reasoning in multimodal diffusion latents. *arXiv preprint arXiv:2604.25299*, 2026.

Christian Templin, Yanda Zhu, and Hao Wang. Generating moving 3d soundscapes with latent diffusion models. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2856–2860. IEEE, 2026.

Zeyue Tian, Zhaoyang Liu, Yizhu Jin, Ruibin Yuan, Liumeng Xue, Xu Tan, Qifeng Chen, Wei Xue, and Yike Guo. Audiox: A unified framework for anything-to-audio generation, 2026. URL https://arxiv.org/abs/2503.10522.

Ivo Trowitzsch, Jalil Taghia, Youssef Kashef, and Klaus Obermayer. The nigens general sound events database, 2020. URL https://arxiv.org/abs/1902.08314.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8228–8238. IEEE, 2024.

Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, Weizhen Bian, Zhen Ye, Sitong Cheng, Ruibin Yuan, Zhixian Zhao, Xinfa Zhu, Jiahao Pan, Liumeng Xue, Pengcheng Zhu, Yunlin Chen, Zhifei Li, Xie Chen, Lei Xie, Yike Guo, and Wei Xue. Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens, 2025. URL https://arxiv.org/abs/2503.01710.

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 *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation, 2023. URL https://arxiv.org/abs/2304.05977.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Jiawei Zhang, Tian-Hao Zhang, Jun Wang, Jiaran Gao, Ruijie Tao, Xinyuan Qian, and Xu-Cheng Yin. I 2 tts: Image-indicated immersive text-to-speech synthesis with spatial perception. In *2025 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 771–776. IEEE, 2025.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process, 2026. URL https://arxiv.org/abs/2509.16117.

Wei Zhou, Xiongwei Zhu, Lingdong Kong, Bo Chen, Lei Zhang, Yongyuan Liang, Xiaoxia Hou, Ye Tian, Xian Sun, Yingshuo Wang, Linfeng Li, Shengqiong Wu, Leigang Qu, Feng Li, Wei Liu, Julian McAuley, and Tat-Seng Chua. On-policy self-distillation in diffusion models. *arXiv preprint arXiv:2608.24646*, 2026a. URL https://arxiv.org/abs/2608.24646.

Wei Zhou, Xiongwei Zhu, Lingdong Kong, Bo Chen, Lei Zhang, Yongyuan Liang, Xiaoxia Hou, Ye Tian, Xian Sun, Yingshuo Wang, et al. On-policy self-distillation in diffusion models. *arXiv preprint arXiv:2608.24646*, 2026b.

A Dataset details

A.1 Audio data construction

We construct the audio-only corpus from isolated sound events, speech recordings, and synthesized multi-event scenes. The sound-effect sources comprise RWCP, NIGENS, FSDKaggle2019, and AudioSet. Dataset-specific labels are mapped to a shared event vocabulary, while an MLLM provides event-level content descriptions for AudioSet recordings. Speech waveforms and transcripts are drawn from VoxBox. Each source is associated with its semantic label or transcript, spatial configuration, and active time interval before being converted to four-channel FOA.

Scene composition. We construct four types of synthetic scenes. Audio-only static scenes contain two or three sound events, whereas audio-only dynamic scenes contain one to three. Speech–audio static scenes contain one speech event and one or two sound events; their dynamic counterparts contain one speech event and zero to two sound events. Every dynamic scene has exactly one moving source. When both speech and non-speech sources are present, the moving source is selected between them with equal probability. The selected moving clip must be at least 4 s long so that its trajectory remains perceptually observable.

Scene duration is determined by the constituent events and constrained to 6–20 s. For two-event scenes, we sample disjoint, partially overlapping, and containment layouts with probabilities 0.40, 0.35, and 0.25, respectively. Three-event scenes use either disjoint or containment layouts, because a single pairwise partial-overlap relation does not define an unambiguous three-event arrangement. A containing event lasts at least 4 s and extends by at least 0.4 s beyond an inner event on each side. Very short clips may be repeated up to four times, with a 0.1 s gap between repetitions, to obtain an audible event span of approximately 1.0–2.5 s. Distinct disjoint events are separated by 0.2–2.0 s of silence.

Structured captions. Captions are generated from the scene metadata with rule-based templates. We explicitly delimit event descriptions, speech transcripts, and spatial expressions with <Des>, <S>, and <D> tags. Multiple paraphrases are sampled for each direction and temporal relation so that identical scene structures do not collapse to one surface form. The basic caption templates are:

```
"<Des>{label}</Des> <D>{direction}</D>."

"<Des>A woman says</Des> <S>{transcript}</S> <D>{direction}</D>."
```

For each direction, we use multiple equivalent expressions, including:

```
front:  in front; ahead; at the front; directly in front.
back:  behind; in back; in the rear; directly behind.
left:  to the left; on the left; from the left; at the left side.
right:  to the right; on the right; at the right side.
front-left:  in front and to the left; at the front left; from the
front-left direction; toward the front left.
front-right:  in front and to the right; at the front right; from
the front-right direction; toward the front right.
back-left:  behind and to the left; at the back left; from the
back-left direction; toward the back left.
back-right:  behind and to the right; at the back right; from the
back-right direction; toward the back right.
above:  above; overhead; from above; in the upper direction.
below:  below; underneath; from below; in the lower direction.
```

Example captions include:

```
"<Des>Explosion</Des> <D>on the left</D>."

"<Des>a clip snapping shut</Des> <D>to the right</D>.  And crossing
it, <Des>cutlery clinking</Des> <D>from above</D>."

"<Des>Vehicle</Des> <D>travelling from the right across to the
back</D>."
```

A.2 Spatialization and simulation configuration

Scene geometry. For the rule-generated single- and multi-event scenes, we use a virtual room of $21 \times 10 \times 21$ m and place the listener at its center. Source coordinates are stored relative to the listener, with $+x$ pointing right, $-z$ pointing forward, and $+y$ pointing upward. Point sources are assigned to ten canonical directions: the eight horizontal directions at 45° intervals, plus above and below. The four cardinal horizontal directions and the two vertical directions lie exactly on their axes; the four horizontal diagonals receive an azimuth perturbation sampled within $\pm 12^\circ$. Static horizontal sources are placed 0.8–10.0 m from the listener, and static vertical sources are placed 1.0–4.5 m away. Direction labels are not reused within a scene when unused directions remain.

An event may instead be marked as diffuse surround sound. At each eligible draw, this option is selected with probability 0.12, with at most one surround event in a scene. Because such an event has no unique bearing or distance, its metadata stores the listener-centered origin together with an explicit surround flag rather than inventing a point-source direction.

Dynamic motion. Dynamic scenes contain either a radial distance change or an angular movement, sampled with equal probability. A radial trajectory preserves its bearing and moves between disjoint near and far ranges. The horizontal ranges are 0.5–2.0 m and 5.0–8.0 m; the vertical ranges are 0.5–1.0 m and 3.0–4.5 m. An angular trajectory preserves source distance and spans at least 90°. Horizontal turns are sampled from $\{\pm 92^\circ, \pm 135^\circ, 180^\circ\}$, while a horizontal-to-vertical or vertical-to-horizontal trajectory is also allowed. For an antipodal 180° motion, an additional waypoint specifies which half-circle is traversed. The metadata records the initial and final positions, and the waypoint when needed, so that the caption and rendered trajectory share the same geometry.

FSDK static rendering. We use a separate reproducible `pyroomacoustics` renderer for the FSDKaggle2019 augmentation. Each mono recording is rendered at up to five independently sampled positions in an $11 \times 11 \times 5$ m shoebox room. The listener is fixed at $[5.5, 5.5, 2.5]$ m; sources are sampled uniformly with a 0.001 m wall inset and remain at least 0.5 m from the listener. Combining a global seed with the stable input-row index makes position sampling independent of worker count. The renderer uses an energy-absorption coefficient of 0.80, image-source order 2, and sound speed 343 m/s. It computes the room impulse response at a colocated omnidirectional microphone and retains the response through 50 ms after the detected onset, subject to a 0.10 s upper bound.

Let $\mathbf{u} = (u_\mathrm{r}, u_\mathrm{u}, u_\mathrm{b})$ be the unit source direction in right–up–back coordinates. The scalar room response is converted to the project's W, Y, Z, X channel order with gains

$$\mathbf{g}_{\mathrm{WYZX}} = [1/\sqrt{2},\ -u_\mathrm{r},\ u_\mathrm{u},\ -u_\mathrm{b}]. \tag{5}$$

Each gain-scaled response is convolved with the resampled mono waveform. The four channels are jointly peak-normalized to -1 dBFS and saved, including the retained reverberation tail, as 44.1 kHz four-channel PCM16 FLAC. Existing outputs are accepted only after validating their sample rate, channel count, duration, and non-silence. The $11 \times 11 \times 5$ m acoustic configuration is specific to this FSDKaggle2019 augmentation; the rule-generated multi-event layouts use the $21 \times 10 \times 21$ m scene geometry described above.

A.3 Construction of video-FOA-caption triplet

For video-conditioned training, following the data construction pipelines of Sphere360 Liu et al. (2025a) and YT-Ambigen Kim et al. (2025), we collect paired panoramic videos and spatial audio from YouTube and use an MLLM to annotate these real-world scenes. Because current audio-language models do not reliably identify source directions in FOA recordings, and similar acoustic events may

Table 4: Statistics of the training data.

Dataset	Samples	Total Duration	Avg. Duration
NIGENS	5,382	28 h	18.0 s
RWCP	58,332	13 h	0.8 s
AudioSet	1,223,464	3,361 h	9.8 s
Spat-VoxBox	1,657,380	2,074 h	4.5 s
YT-Ambigen	90,547	251 h	10.0 s
Sphere360	95,507	265 h	10.0 s
FSDKaggle2019	29,820	63 h	7.65 s
multi_static	600,000	1,473 h	8.84 s
multi_dynamic	1,100,000	2,906 h	9.51 s
enhance-vt	164,321	456 h	9.99 s
Total	**5,024,753**	**10,890 h**	–

Table 5: Reconstruction performance of our FOA VAE and baseline models. PESQ is evaluated on the W channel. Best results are highlighted in bold.

Model	mrSTFT ↓	DoA ↓	PESQ ↑	ILD (dB) ↓	wCS ↑	CC-5fps ↑	CC-all ↑
Stable Audio VAE	15.9002	10.6573	1.9496	5.7275	0.3517	0.9349	0.9464
DAC	1.0588	5.9682	2.1268	1.3015	0.5518	**0.9458**	0.9527
Ours	**0.8854**	**5.4239**	**2.6433**	**1.1383**	**0.5717**	0.9431	**0.9637**

admit multiple semantic interpretations, the annotator receives both the audio and the panoramic video. It produces structured descriptions of sound semantics and source directions, yielding paired data for video-conditioned, text-conditioned, and joint video–text-conditioned training. Examples are shown below:

```
"<Des>Motorcycle engines running</Des> <D>in front</D> and
<D>below</D>.  <Des>Wind noise</Des> from forest vegetation
<D>somewhere</D>."

"<Des>River water flowing</Des> from river water <D>all around</D>.
<Des>Paddles stroking</Des> from paddles <D>all around</D>.
<Des>Items bumping canoe hull</Des> from items <D>in front</D>.
<Des>Someone speaking</Des> from speaker <D>somewhere</D>."
```

To enable text to supplement off-screen audio during joint video–text generation, we further randomly add speech or other audio to the original videos and append the corresponding descriptions to the captions as augmented data. The statistics of the final training and test sets are reported in Table 4.

B FOUR-CHANNEL FOA WAVVAE DETAILS

Given a first-order Ambisonics waveform $\mathbf{x} \in \mathbb{R}^{4\times T}$ sampled at 44.1 kHz, we use the W–Y–Z–X channel order, where W denotes the omnidirectional component and X, Y, Z denote the three first-order directional components. Our FOA WavVAE is based on the one-dimensional convolutional AutoencoderOobleck architecture. The encoder first maps the four-channel waveform to 128 hidden channels using a convolution with a kernel size of 7, followed by five residual downsampling stages with strides of $[2, 2, 4, 8, 8]$, yielding an overall temporal compression factor of 1024. The residual units employ Snake nonlinearities and weight-normalized 1-D convolutions to model periodic acoustic structures and local patterns at different temporal scales. The encoder produces 128 output channels, partitioned into a 64-dimensional mean vector $\boldsymbol{\mu}$ and 64-dimensional scale parameters $\mathbf{s}$. During training, latent variables are sampled through reparameterization as $\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The decoder mirrors the encoder to reconstruct the four-channel FOA waveform from the latent representation. The latent frame rate is approximately $44100/1024 \approx 43.1$ Hz, with each

latent frame corresponding to approximately 23.2 ms. The VAE parameters occupy approximately 155 MB.

The model is initialized from a pretrained stereo AutoencoderOobleck [Stable Audio VAE]. We adapt the encoder input layer and decoder output layer to four channels and load pretrained weights into the remaining compatible layers. The model is then jointly fine-tuned on FOA data, incorporating intensity-direction consistency and axis-wise MR-STFT losses alongside the reconstruction objective. These losses constrain the reconstructed sound-field directions and the spectral structures along the front, back, left, right, up, and down directions, respectively, preserving spatial relationships across channels. After training, the encoder and decoder are frozen and used for latent extraction and FOA waveform reconstruction, respectively, in the subsequent generative framework. Table 5 compares the reconstruction performance of our FOA VAE with Stable Audio VAE and DAC. The mrSTFT metric measures spectral reconstruction errors across multiple time–frequency resolutions, DoA measures directional deviations between reconstructed and reference sound fields, and ILD error measures discrepancies in relative level differences; lower values indicate better performance for all three metrics. PESQ evaluates perceptual speech quality on the omnidirectional W channel. wCS measures event-direction agreement weighted by reference confidence, while CC-5fps and CC-all measure spatial energy-map correlations over short temporal windows and entire clips, respectively. Higher values indicate better performance for these metrics.

Our model achieves the lowest spectral, directional, and level-difference reconstruction errors, together with the highest PESQ, wCS, and CC-all, indicating improved preservation of both acoustic content and spatial structure. DAC obtains a slightly higher CC-5fps, suggesting room for improvement in short-term spatial energy consistency. Overall, these results support the ability of the learned FOA representation to preserve acoustic and spatial information through compression and reconstruction.