SPATIAL AUDIO GENERATION AND EDITING WITH EXECUTABLE PLANS AND SELF-DISTILLING TRANSFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generating and editing multi-source spatial audio requires fine-grained control over source semantics, temporal activity, and 3D trajectories, while preserving unmodified acoustics during edits. Unifying these tasks requires explicit source-level instruction grounding and preservation of multichannel spatial relationships. We present AMBIT, an end-to-end framework for native First-Order Ambisonics (FOA) generation and editing using an executable SCENEPLAN that binds source descriptions to activity intervals and 3D trajectories. Built on a unified Transfusion backbone, our framework introduces three contributions: (1) autoregressive planning coupled with rectified-flow rendering, using contrastive reference features for instruction grounding and full reference latents for acoustic preservation; (2) an FOA-aware continuous VAE with shared-decoder omnidirectional recovery, asymmetric grouped KL, and spatial covariance matching; and (3) on-policy self-distillation (OPSD) tailored to Transfusion, jointly refining planning and rendering with target-conditioned teacher predictions on student-generated prefixes and sampling states. Target audio is used only for teacher conditioning during refinement, not inference. Evaluations across speech, music, and sound scenes show improved generation content and spatial quality over the compared baselines, alongside better acoustic fidelity and unedited-source preservation during targeted edits.

1 INTRODUCTION

Spatial audio authoring supports immersive media, virtual acoustics, and interactive environments through text-driven generation and iterative instruction-guided refinement. Generation requires controlling each source's content, activity, and 3D trajectory, whereas editing interprets instructions against an existing soundfield to identify target sources and modify their attributes while preserving unmodified content and spatial structure. These complementary tasks motivate a unified framework linking instruction grounding, discrete planning, and continuous multichannel synthesis.

Spatial audio research has advanced substantially. ImmerseDiffusion synthesizes soundscapes in First-Order Ambisonics (FOA) from text and spatial, temporal, and environmental conditions using latent diffusion (Heydari et al., 2025). SonicMotion extends FOA generation to moving sound sources with explicit trajectory control (Templin et al., 2026). For editing, SwanWeave follows natural language instructions to modify existing FOA scenes across content, timing, and spatial attributes (Lei et al., 2026a). However, unified scene authoring requires maintaining agreement between language intent, source attributes, and rendered acoustics without disturbing unmodified content.

This task mainly faces three challenges. **(i) Grounding language in source-level 3D controls.** AudioChat explores stereo planning (Chen et al., 2026); native FOA authoring additionally requires binding source content, timing, and 3D motion, including identifying the edited target source. **(ii) Preserving acoustics and spatial structure in compact latents.** Continuous FOA autoencoders support generation, while discrete codecs provide spatial-consistency supervision (Heydari et al., 2025; Sudarsanam et al., 2025). Yet, channel-wise reconstruction fails to explicitly supervise inter-channel energy, coherence, and phase. **(iii) Refining discrete planning and continuous rendering.** Paired training uses reference prefixes and target plans, whereas inference relies on predicted plans

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.

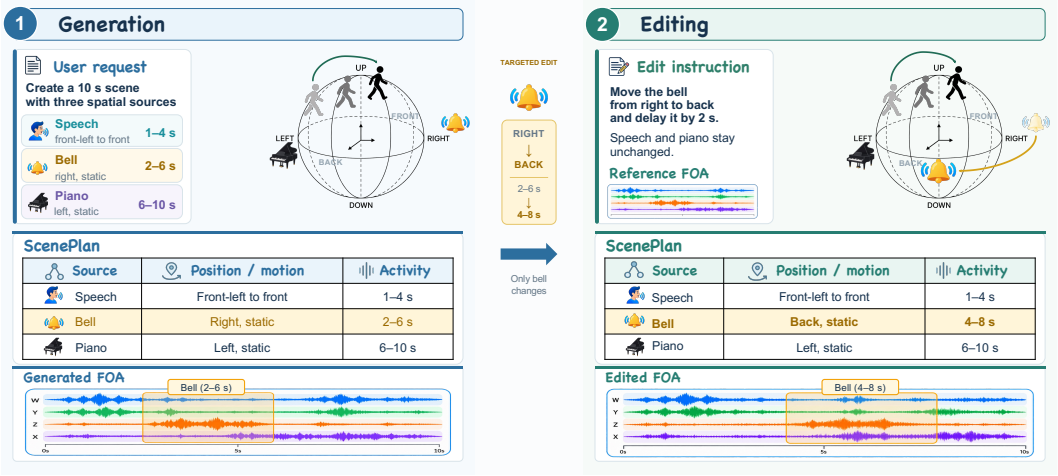


Figure 1: Overview of AMBIT capabilities. **(1) Spatial audio generation** creates multi-source 3D scenes from text, controlling content, timing, and trajectories. **(2) Instruction-guided spatial and temporal editing** modifies selected sources, preserving unmodified content and spatial structure.

and visited states. DiffusionOPSD (Zhou et al., 2026) refines continuous generators without discrete token supervision; joint self-distillation requires complementary targets for both modes.

To address these challenges, we introduce AMBIT, a framework for native FOA generation and editing through Transfusion (Zhou et al., 2025) (shown in Figure 1), with three contributions: **(i) Source-grounded Transfusion planning and rendering.** We use an executable SCENEPLAN as a common interface, specifying source-level content, timing, and motion before synthesis. A shared Transfusion backbone couples autoregressive planning with rectified-flow rendering (Ma et al., 2025), using source-indexed controls deterministically compiled from plans. For editing, we combine semantic and spatial contrastive features for instruction grounding with full reference latents to preserve acoustic details absent from the plan. **(ii) FOA-aware continuous Variational Autoencoder.** Accurate omnidirectional W recovery alone does not constrain inter-channel geometry. We pair shared-decoder auxiliary W reconstruction with asymmetric grouped KL, assigning one latent group a recovery role under weaker regularization. Spatial covariance matching complements this by supervising relative channel energy, coherence, and phase. **(iii) Joint Transfusion On-Policy Self-Distillation.** To address the coupled distribution shift in planning and rendering, we adapt on-policy self-distillation (Zhao et al., 2026b; Jiang et al., 2026a) to the student's generation process. Student-generated plans condition DiT trajectories, while frozen teachers use target plan memory and target audio reference tokens to supervise planning and rendering, respectively. Token distribution and clean latent distillation jointly update the shared Transformer and task-specific modules, coupling the branches through shared parameters and fresh student rollouts. The student uses only source audio and edit instructions during rollouts and inference.

For music and sound generation, AMBIT achieves higher paired CLAP and lower FAD than AudioX-Turbo (Tian et al., 2026a), while achieving the lowest speech WER and CER. Spatially, it reduces directional and binaural errors over TangoFlux (Hung et al., 2026) with spatialization. For editing, it outperforms SwanWeave (Lei et al., 2026a) in acoustic fidelity and spatial accuracy. Fixed-renderer comparisons support our planner, VAE ablations show better reconstruction with retained spatial accuracy, and Transfusion OPSD outperforms SFT, ReFL, and GRPO across all nine metrics.

2 RELATED WORK

Spatial audio generation and representation. General-purpose and multimodal generators broaden audio content and conditioning (Liu et al., 2024; Evans et al., 2025; Tian et al., 2026b;a; 2025; Hung et al., 2026), while timing-aware methods introduce event-level temporal control (Xie et al., 2024; Jiang et al., 2025; 2026b). Diff-SAGe, DualSpec, and PhysWave (Kushwaha et al., 2025; Zhao et al., 2026a; Yao et al., 2026) explore category and location conditioning, joint spectrogram modeling, and direction- and distance-consistency priors, respectively. Video-conditioned and

trajectory-aware models further extend FOA synthesis (Kim et al., 2025; 2026; Lei et al., 2026b). Neural audio codecs (Kumar et al., 2023; Défossez et al., 2022; Ji et al., 2025) and continuous FOA autoencoders (Heydari et al., 2025) provide compact representations, while FOA Tokenizer incorporates spatial consistency supervision into discrete compression (Sudarsanam et al., 2025). AMBIT combines source-level 3D control with acoustic fidelity across native FOA generation and editing. Our continuous VAE supervises acoustic reconstruction and inter-channel spatial consistency.

Audio editing and executable scene planning. AudioSpa and TAS offer text-guided mono-to-binaural spatialization (Feng et al., 2025; Pan et al., 2025). Diffusion editing uses noise perturbation, inversion, or instruction conditioning (Meng et al., 2021; Manor & Michaeli, 2024; Jia et al., 2025; Wang et al., 2023). MMEdit, Audio-Omni, and SmartDJ (Tao et al., 2025; Tian et al., 2026c; Lan et al., 2025) extend language-guided editing, while SwanWeave targets native FOA editing (Lei et al., 2026a). Scripting and structured planning support compositional synthesis (Liu et al., 2025c; Wang et al., 2026); AudioChat unifies stereo generation and editing (Chen et al., 2026). Following Transfusion (Zhou et al., 2025), AMBIT unifies autoregressive planning and native FOA flow rendering. Executable, grammar-constrained SCENEPLANs bind content, activity, and 3D paths, compiling deterministically into source-indexed semantic and frame-aligned spatial controls. Audio and text contrastive learning aids spatial grounding (Elizalde et al., 2023; Wu et al., 2023; Seki et al., 2026); our FOA features ground edits; full reference latents preserve unedited-source acoustics.

On-policy refinement and self-distillation. Preference optimization and reinforcement learning improve audio generation (Majumder et al., 2024; Li et al., 2026); ReFL adds differentiable reward feedback (Xu et al., 2023). GRPO's group-relative advantages (Shao et al., 2024) extend to flow and diffusion generators (Liu et al., 2025b; Xue et al., 2025). On-policy distillation supervises student-generated sequences (Agarwal et al., 2024), with self-distillation for LLM reasoning (Zhao et al., 2026b) and continuous generators (Zhang et al., 2026; Zhou et al., 2026; Jiang et al., 2026a; Liu et al., 2026). Transfusion OPSD instead jointly refines a shared FOA editor's discrete planner and continuous renderer. Frozen planning and rendering teachers use target plan memory and target audio to supply token distributions on student prefixes and clean-latent targets at student-visited flow states. Target conditioning is teacher-only; student inference inputs remain unchanged.

3 METHODOLOGY

Overview. Let u, e denote generation and editing instructions, x_s a reference First-Order Ambisonics (FOA) waveform, and $z_s = E_V(x_s)$ its latent, with VAE encoder/decoder E_V, D_V. The autoregressive (AR) planner π_θ predicts an executable SCENEPLAN $\mathcal{S}$ token by token, specifying source-indexed content, activity, and 3D motion. Compilation and condition encoding produce semantic tokens and frame-aligned spatial controls $\mathcal{C}(\mathcal{S})$. We adapt the Diffusion Transformer (DiT) architecture (Peebles & Xie, 2023) for audio rendering. Our renderer v_θ predicts audio latent velocities, which a rectified flow sampler $\Phi[v_\theta]$ integrates before VAE decoding. AR and DiT are two modes of a Transfusion-style Transformer, with shared blocks and task-specific interfaces collectively parameterized by θ. For editing, frozen AMBIT CLAP features $F(z_s)$ ground instructions, while full reference latents preserve acoustic detail. With $\xi = u$ for generation or $\xi = (z_s, F(z_s), e)$ for editing, inference is

$$\widehat{\mathcal{S}} = \pi_\theta(\xi), \qquad \widehat{x} = D_V\Big(\Phi[v_\theta]\big(\epsilon; \mathcal{C}(\widehat{\mathcal{S}}), z_s\big)\Big). \tag{1}$$

Here $\epsilon \sim \mathcal{N}(0, I)$ is noise and $\widehat{x}$ is FOA output; generation omits $z_s, F(z_s)$ (Figure 2).

3.1 EXECUTABLE SCENE PLANS

A SCENEPLAN $\mathcal{S}$ binds source content, activity, and spatial attributes before synthesis. It specifies duration, room type, and source types (speech, music, or sound), description, optional speech transcript, active interval, and a static or linear 3D trajectory in azimuth, elevation, and distance. The planner predicts the full edited scene with modified and retained sources.

SCENEPLAN compilation. A deterministic compiler converts source descriptions into text clauses and aligns activity and pose fields to latent frames. A frozen text backbone, learned projection, and source role embeddings produce cross-attention tokens c_{attn}. Event embeddings mark latent frames

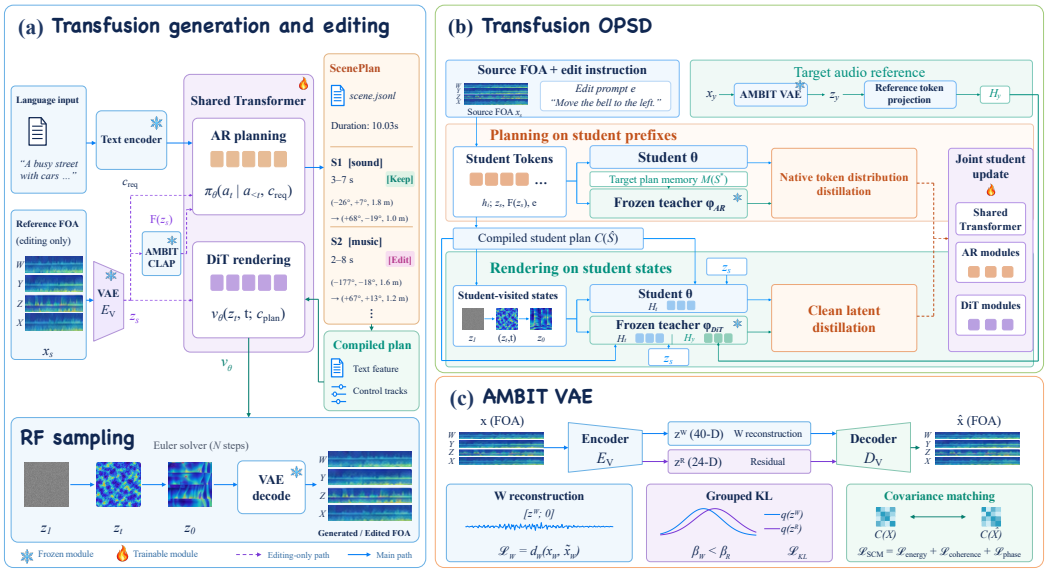


Figure 2: Overview of AMBIT. **(a)** A shared Transformer predicts executable SCENEPLANs and renders native FOA through latent flow conditioned on compiled plans. Editing uses source latents in both modes and AMBIT CLAP features in planning. **(b)** OPSD distills token and clean latent predictions on student prefixes and visited states, jointly updating the shared Transformer and AR/DiT modules. Target plans and audio condition only the frozen teachers. **(c)** The VAE combines shared-decoder auxiliary W reconstruction, grouped KL, and spatial covariance supervision.

whose sample intervals overlap source activity; a trajectory MLP encodes poses at frame midpoints. The spatial fields form $c \in \mathbb{R}^{256\times T}$ over T latent frames. Thus $\mathcal{C}(\mathcal{S}) = (c_{\text{attn}}(\mathcal{S}), c(\mathcal{S}))$ combines semantic conditioning with temporally aligned source control tracks (Appendix B.1).

3.2 AMBIT VARIATIONAL AUTOENCODER

We adapt the continuous audio VAE of Stable Audio Open (Evans et al., 2025) to encode four-channel waveforms x as latents $z = E_V(x)$ and reconstruct FOA through $D_V(z)$. Extending the input and output to four channels does not itself prioritize omnidirectional recovery or add supervision of FOA inter-channel relationships. We therefore combine an auxiliary W channel reconstruction task, asymmetric grouped KL regularization, and spatial covariance matching (SCM).

Omnidirectional recovery and capacity allocation. The omnidirectional W channel provides a scene-wide acoustic reconstruction target. We partition $z = [z^W; z^R]$, where z^R contains the remaining latent channels. The full latent reconstructs all four channels. An auxiliary pass masks z^R and supervises only W through the same decoder; we combine this auxiliary reconstruction objective with grouped KL regularization:

$$\widetilde{x}_W = \left[D_V([z^W; \mathbf{0}])\right]_W, \qquad \mathcal{L}_{\text{KL}} = \beta_W K_W + \beta_R K_R, \quad \beta_W < \beta_R. \tag{2}$$

The auxiliary loss $\mathcal{L}_W$ combines multiresolution short-time Fourier transform (STFT) and SI-SDR losses (Le Roux et al., 2019); K_W, K_R are groupwise KL penalties. The W task assigns z^W a reconstruction role under weaker KL regularization. Decoder reuse adds no parameters; the partition encourages specialization without enforcing exact content and spatial disentanglement.

Spatial covariance matching. To constrain FOA inter-channel structure, spatial covariance matching (SCM) compares local STFT covariance through energy fractions p_i, coherence magnitudes γ_{ij}, and stabilized inter-channel phase factors U_{ij} between reference and reconstructed audio:

$$\mathcal{L}_{\text{SCM}} = \left\langle \sum_i |\widehat{p}_i - p_i| \right\rangle_w + \left\langle \sum_{i<j} |\widehat{\gamma}_{ij} - \gamma_{ij}| \right\rangle_w + \left\langle 1 - \text{Re}(\widehat{U}_{ij}\overline{U_{ij}}) \right\rangle_v. \tag{3}$$

Hats denote reconstructed statistics; brackets denote weighted averages. The phase component retains magnitude dependence from stabilization. The full VAE training objective is

$$\mathcal{L}_{\mathrm{VAE}} = \mathcal{L}_{\mathrm{base}} + \mathcal{L}_{W} + \mathcal{L}_{\mathrm{KL}} + \lambda_{\mathrm{SCM}}\mathcal{L}_{\mathrm{SCM}}. \tag{4}$$

Here $\mathcal{L}_{\mathrm{base}}$ contains spectral, waveform, and adversarial terms, and λ_{SCM} weights spatial supervision. Appendix B.2 details the covariance statistics, stabilization, and weighted reductions.

3.3 TRANSFUSION-STYLE PLANNING AND RENDERING

AMBIT unifies FOA generation and editing in a shared Transfusion Transformer (Zhou et al., 2025). AR planning predicts executable SCENEPLANs; DiT renders FOA from compiled plans. Mode-specific projections and masks enable causal attention over plan tokens and bidirectional attention over rendering latents. Generation uses text instructions; editing adds source audio.

AMBIT CLAP for source understanding. Following CLAP (Elizalde et al., 2023; Wu et al., 2023), we learn separate semantic and scene alignment heads from frozen AMBIT VAE latents:

$$\mathcal{L}_{\mathrm{align}} = \mathcal{L}_{\mathrm{sem}} + \lambda_{\mathrm{scene}}\mathcal{L}_{\mathrm{scene}}, \tag{5}$$

where both terms use symmetric audio and text contrastive losses with multiple positives. Semantic alignment matches source content and counts; scene alignment additionally matches source timing, spatial attributes, and room properties, weighted by λ_{scene}. The frozen encoder supplies global and temporal features $F(z_s)$ to the editing planner. Appendix B.3 details encoder training.

Autoregressive planning. A grammar constrained codec serializes SCENEPLANs into tokens. We train the planner with next token cross entropy $\mathcal{L}_{\mathrm{AR}}$ on reference prefixes and task inputs. Planning heads predict source inventory, copy literal spans, and select qualitative controls for deterministic completion (Appendix B.4). Editing uses a bidirectional source latent prefix with AMBIT CLAP features and instruction cross attention. A source slot decoder predicts edit targets, operations, and updated attributes; its memory guides AR decoding via gated residual attention.

Flow rendering. A compiled SCENEPLAN specifies scene structure; full reference latents z_s provide acoustic details of retained sources for editing. For target waveform x_y, let $z_0 = z_y = E_V(x_y)$ be its clean latent. We sample flow time $t \in [0, 1]$ from $p(t)$, construct $z_t = (1 - t)z_0 + t\epsilon$, and minimize the base rectified flow objective (Liu et al., 2022):

$$\mathcal{L}_{\mathrm{RF}} = \mathbb{E}\left[\frac{1}{\|m\|_1}\left\|m \odot \left(v_\theta([z_t; c(\mathcal{S}); z_s], t \mid c_{\mathrm{attn}}(\mathcal{S})) - (\epsilon - z_0)\right)\right\|_2^2\right]. \tag{6}$$

Brackets concatenate channels; $\odot$ multiplies elementwise. The binary mask m selects valid latent coordinates across all channels, excluding padded frames; $\|m\|_1$ counts all valid latent coordinates. Generation omits z_s; sampling integrates from noise at $t = 1$ to the output latent at $t = 0$. Generation training uses normalized speech activity weighting and an auxiliary latent temporal difference loss for sound only scenes. Appendices B.4 and C detail objectives and time distributions.

Joint planning and rendering. Paired editing jointly supervises plan tokens and audio velocities alongside recognition, edit binding, and attribute preservation under a common source assignment:

$$\mathcal{L}_{\mathrm{joint}} = \mathcal{L}_{\mathrm{AR}} + \mathcal{L}_{\mathrm{RF}} + \lambda\mathcal{L}_{\mathrm{struct}}, \tag{7}$$

where λ weights the structured source supervision $\mathcal{L}_{\mathrm{struct}}$ (Appendix B.5). AR and DiT losses jointly update the shared Transformer and their respective interfaces. Paired rendering uses ground truth target plans; inference uses student predictions. The pretrained text backbones, AMBIT CLAP, and AMBIT VAE remain frozen; student conditioning projections and embeddings are trainable.

3.4 ON-POLICY SELF-DISTILLATION FOR PLANNING AND RENDERING

Paired training uses reference prefixes and ground-truth plans, whereas inference uses student-generated counterparts. This discrepancy propagates to rendering, because student-generated plans determine rendering conditions and influence sampling states. We adapt on-policy self-distillation (Zhao et al., 2026b; Jiang et al., 2026a) to supervise both modes on student-generated contexts, with both losses updating the shared Transformer. From a common supervised model, we prepare and freeze planning and rendering teachers with target plan memory and audio reference tokens, respectively. The student uses source audio and instructions throughout refinement and inference for editing.

Planning on student prefixes. The student and planning teacher ϕ_{AR} share editing conditions $c = (z_s, F(z_s), e)$ and a student generated prefix h_i. The teacher additionally reads target plan memory $M(\mathcal{S}^\star)$. For non-forced decisions $\mathcal{I}$ and student distribution $p_{\theta,i} = p_\theta(\cdot \mid h_i, c)$, we use

$$q_i = \mathrm{sg}[p_{\phi_{\mathrm{AR}}}(\cdot \mid h_i, c, M(\mathcal{S}^\star))], \qquad \ell_{\mathrm{AR}} = \sum_{i \in \mathcal{I}} \eta_i D_{\mathrm{KL}}(q_i \,\|\, p_{\theta,i}). \tag{8}$$

Both distributions are normalized over the same tokens allowed by the grammar at h_i. The stop-gradient operator sg treats q_i as a fixed teacher target, so the KL loss updates only the student. Detached weights η_i balance decision families and exclude positions where aligned target facts are illegal or conflict with teacher predictions. A reference loss ℓ_{ref} anchors selected plan fields to predictions of the frozen initial supervised model on h_i under ordinary conditions.

Rendering on student states. Compiling the AR student's plan $\widehat{\mathcal{S}}$ provides conditions for DiT trajectories alongside source latent z_s. We collect visited states (z_j, t_j), with queries $\mathcal{Q}$ stratified across sampling times. The teacher ϕ_{DiT} receives the same states, source latent, and plan, plus temporally aligned reference tokens $H_y = P_{\phi_{\mathrm{DiT}}}(z_y)$ projected from clean target audio latents as additional attention context. Matching the models' clean latent predictions gives

$$\ell_{\mathrm{DiT}} = \sum_{j \in \mathcal{Q}} \frac{\omega_j t_j^2}{\|m\|_1} \Big\| m \odot \Big(v_\theta(z_j, t_j; \widehat{\mathcal{S}}, z_s) - \mathrm{sg}\Big[v_{\phi_{\mathrm{DiT}}}(z_j, t_j; \widehat{\mathcal{S}}, z_s, H_y) \Big] \Big) \Big\|_2^2. \tag{9}$$

Masking and normalization follow Equation 6. The factor t_j^2 follows from $\widehat{z}_0 = z_j - t_j v$; ω_j averages noise trajectories and time strata. Sampled states, plans, and teacher outputs are detached.

Joint refinement through shared parameters. For training records $(x_s, e, \mathcal{S}^\star, x_y) \sim \mathcal{D}$,

$$\mathcal{L}_{\mathrm{OPSD}} = \mathbb{E}_{\mathcal{D}}[\lambda_{\mathrm{AR}} \ell_{\mathrm{AR}} + \lambda_{\mathrm{DiT}} \ell_{\mathrm{DiT}} + \lambda_{\mathrm{ref}} \ell_{\mathrm{ref}}]. \tag{10}$$

Gradients from these losses are accumulated for the shared Transformer and the task-specific modules before each optimizer step. Updating this shared representation couples the refinement of both modes in subsequent rollouts. Only student predictions and trainable conditions receive gradients, while rollouts, teacher outputs, reference models, and pretrained backbones remain detached or frozen. Appendix E details teacher configurations, loss reductions, and optimization hyperparameters.

4 Experiments

4.1 Experimental Setup

Datasets. We train on 1.6M FOA examples from synthetic scenes and real recordings, and 1M editing pairs covering speech, music, and environmental sounds. The AMBIT VAE is pretrained on 1M FOA clips. Evaluation uses 8,000 generation requests, 5,000 editing pairs, and 2,000 VAE test clips. Appendices A and D detail data construction, partitions, and evaluation protocols.

Model and training. AMBIT uses 15 shared Transfusion blocks (width 1,024, 16 attention heads) for native FOA generation and editing. The shared backbone, native AR/DiT interfaces, and structured heads total 332.6M trainable parameters. Joint supervised training uses token and rectified flow objectives with AdamW at 10^{-4} for adapters and heads and 2×10^{-6} for shared blocks. Paired rendering uses ground-truth plans; inference uses AR predictions. The Qwen3.5-0.8B text backbone (Qwen Team, 2026), AMBIT CLAP, and AMBIT VAE remain frozen. OPSD uses AdamW with a global batch of 64 requests and loss weights $\lambda_{\mathrm{AR}} = 0.1$, $\lambda_{\mathrm{DiT}} = 1$, and $\lambda_{\mathrm{ref}} = 0.1$, with prepared teachers frozen. Architecture and training settings appear in Appendices B, C, and E.

Evaluation metrics. We evaluate generation and editing with the same AMBIT checkpoint. For spatial audio, we follow SwanSphere (Lei et al., 2026b) in reporting FD-PaSST and KL-PaSST, supplemented by paired audio CLAP, FAD, KL-PANNs, and FD-OpenL3. Speech evaluation uses Distil-Whisper WER/CER (Gandhi et al., 2023) and UTMOS (Saeki et al., 2022). For editing, we report FD, FAD, KL, and LSD following SwanWeave (Lei et al., 2026a), with CLAP and PANNs features for FD, and the paired CLAP score (Wu et al., 2023). Both tasks report direction errors on eligible isolated intervals: spherical and azimuth MAE, and angular trajectory RMSE. Complementary GCC, CRW (Chen et al., 2022), FSAD (Sun et al., 2025), and wCS from PSELDNets (Lei et al.,

2026b; Hu et al., 2025), together with scene direction and unedited source preservation, are detailed in Appendix D. Codec evaluation reports reconstruction quality, intensity-vector direction error, and framewise intensity-coherence error in Eq. 25. Appendix D.4 reports our subjective listening study.

4.2 AMBIT VAE EVALUATION

FOA reconstruction must preserve acoustic content and spatial structure. We compare codecs and train cumulative objective variants from scratch for matched updates (Table 1). AMBIT outperforms DAC (Kumar et al., 2023), EnCodec (Défossez et al., 2022), and WavTokenizer (Ji et al., 2025) across all four metrics. SCM yields the largest cumulative W SI-SDR and direction-error gains; grouped KL with auxiliary W recovery further improves all four metrics. These gains support combined supervision of omnidirectional content and interchannel geometry (Appendix D.3).

Table 1: FOA reconstruction and cumulative ablation. The upper block compares codecs on native scenes; each "+" below adds to the previous row, with grouped KL replacing uniform KL. Ablations use matched updates from scratch and 24 kHz evaluation. Bold marks the best per block.

System/configuration	W SI-SDR (dB) ↑	Direction (°) ↓	Intensity coherence error ↓	W paired CLAP ↑
DAC	5.57	5.13	0.08	0.88
EnCodec	0.41	12.25	0.15	0.85
WavTokenizer	-19.41	56.66	0.35	0.63
AMBIT VAE	**6.39**	**4.96**	**0.04**	**0.90**
Phase (base)	5.63	6.51	0.06	0.86
+ SCM	6.08	5.14	0.05	0.88
+ Cepstral	6.15	5.09	0.06	0.87
+ SI-SDR	6.25	5.08	0.05	0.87
+ Grouped KL + $\mathcal{L}_W$ (full)	**6.39**	**4.96**	**0.04**	**0.90**

4.3 GENERATION AND EDITING QUALITY

We examine whether executable plans support content fidelity across music, sound, and speech through comparisons with domain baselines (Table 2). AMBIT leads paired CLAP, classifier KL, and FAD in music and sound, while AudioX-Turbo retains lower FD-PaSST/FD-OpenL3. For speech, it achieves lower WER/CER but lower UTMOS than Qwen3-TTS. These results are consistent with source descriptions and literal transcripts constraining event content and spoken words, improving semantic and lexical fidelity without uniformly improving perceptual or distributional quality.

Editing adds an extra requirement, retaining the source's acoustic context. We compare editing baselines (Table 3). AMBIT improves all six metrics over SwanWeave, the strongest baseline. Table 5 evaluates unchanged-source preservation. Higher paired CLAP and lower LSD accompany gains across the distributional metrics. The results are consistent with complementary inputs: AMBIT CLAP grounds the requested edit, while reference latents supply acoustic context beyond the plan.

4.4 SPATIAL AUDIO EVALUATION

Spatial control depends on translating plan attributes into audio. We compare scene and object metrics against cascaded baselines. BEWO + AS and TangoFlux + AS use audio spatialization (AS) with Qwen3.5-4B SCENEPLANs and Pyroomacoustics (Scheibler et al., 2018); Qwen3.5-4B + DiT uses our renderer (Appendix D.2). Object metrics use intervals with one annotated active source (Appendix D.3); planner validity appears in Appendix D.5. Qwen3.5-4B + DiT improves GCC, CRW, and wCS over AS pipelines but has higher scene and object direction errors; binaural gains thus need not imply accurate source placement. AMBIT leads both groups (Table 4). Its gains over Qwen3.5-4B + DiT with the same renderer support more effective spatial controls from the learned planner, extending directional improvements to individual source intervals.

Preserving unedited sources is a separate requirement from matching target directions. We compare scene quality, object directions, and preservation against SwanWeave and Qwen3-Omni + DiT (Xu et al., 2025) (Table 5). The latter uses our renderer and improves scene and object metrics and W

Table 2: Generation content quality. Bold and underlined values indicate the best and second-best results within each evaluated domain; displayed ties receive the same marking.

Domain	System	Paired CLAP ↑	FAD ↓	KL-PANNs ↓	KL-PaSST ↓	FD-PaSST ↓	FD-OpenL3 ↓
Music	AudioX-Turbo (Tian et al., 2026a)	0.36	1.07	0.68	1.58	**234.65**	**37.40**
	AudioX (Tian et al., 2026b)	0.35	1.30	0.74	1.56	243.68	46.67
	Stable Audio Open (Evans et al., 2025)	0.18	2.08	0.90	1.81	410.64	68.72
	TangoFlux (Hung et al., 2026)	0.36	1.01	0.59	1.43	319.37	92.09
	BEWO (Sun et al., 2025)	0.26	2.05	0.56	1.40	566.57	57.30
	AMBIT	**0.46**	**0.71**	**0.51**	**0.95**	261.45	44.59
Sound	AudioX-Turbo (Tian et al., 2026a)	0.34	0.79	0.63	2.36	**213.60**	**40.56**
	AudioX (Tian et al., 2026b)	0.32	1.08	0.68	2.35	228.62	52.17
	Stable Audio Open (Evans et al., 2025)	0.17	1.58	0.79	2.70	334.40	61.81
	TangoFlux (Hung et al., 2026)	0.36	0.99	0.54	2.18	243.97	85.85
	BEWO (Sun et al., 2025)	0.21	2.40	0.55	2.45	517.14	66.27
	AMBIT	**0.41**	**0.75**	**0.53**	**1.78**	235.49	49.99

Domain	System	WER ↓	CER ↓	UTMOS ↑
Speech	UniAudio2 (Yang et al., 2026)	0.08	0.06	4.11
	Qwen3-TTS 1.7B (Hu et al., 2026)	0.09	0.08	**4.18**
	Parler-TTS Large (Lyth & King, 2024)	0.11	0.06	3.40
	AMBIT	**0.07**	**0.05**	3.52

Table 3: Editing content quality and spectral reconstruction. The KL here is KL-PANNs. Bold and underlined values indicate the best and second-best results. Spatial metrics are reported in Table 5.

System	Paired CLAP ↑	FD-CLAP ↓	FAD ↓	FD-PANNs ↓	KL ↓	LSD ↓
Audio-Omni (Tian et al., 2026c)	0.46	0.23	0.91	23.75	2.04	2.09
MMEdit (Tao et al., 2025)	0.48	0.31	1.15	22.73	1.37	2.45
SwanWeave (Lei et al., 2026a)	0.68	0.15	0.57	11.40	0.92	1.65
SmartDJ (Lan et al., 2025)	0.63	0.17	0.70	19.30	1.16	1.89
ZETA (Manor & Michaeli, 2024)	0.41	0.38	1.82	33.56	1.63	2.86
AudioEditor–BEWO (Jia et al., 2025)	0.44	0.36	1.87	25.12	1.39	2.43
SDEdit (Meng et al., 2021)	0.32	0.48	2.13	41.45	1.95	3.36
AMBIT	**0.84**	**0.05**	**0.24**	**6.01**	**0.46**	**1.27**

Table 4: Spatial generation at scene and object levels. Object metrics use isolated source intervals, while scene direction errors use mixtures. Angular errors are in radians. Bold marks the best score.

System	Scene spatial quality							Object direction		
	GCC ↓	CRW ↓	FSAD ↓	DOA ↓	Az. MAE ↓	Ang. traj. RMSE ↓	wCS ↑	Spherical MAE ↓	Az. MAE ↓	Ang. traj. RMSE ↓
BEWO + AS	34.71	73.22	0.03	1.04	1.00	1.13	0.11	1.001	0.949	1.066
TangoFlux + AS	36.42	67.88	0.04	1.03	0.98	1.12	0.09	0.994	0.942	1.055
Qwen3.5-4B + DiT	31.07	64.36	0.03	1.07	1.02	1.18	0.27	1.030	0.973	1.113
AMBIT	**12.59**	**8.75**	**0.02**	**0.51**	**0.40**	**0.63**	**0.56**	**0.531**	**0.415**	**0.629**

NMSE over SwanWeave, but increases unchanged-source direction and level errors. This motivates measuring preservation against original source audio alongside reconstruction against the target scene. AMBIT leads all three metric groups, improving target direction accuracy alongside retention of the original direction, waveform, and level on eligible unchanged-source intervals. These results are consistent with source grounding and compiled controls specifying the desired changes, while full reference latents supply the acoustic context needed to preserve unedited sources.

4.5 POST-TRAINING WITH OPSD

Post-training should improve content alongside spatial execution. We compare OPSD configurations with continued SFT, our ReFL adaptation (Xu et al., 2023), and two branch GRPO (Shao et al., 2024; Liu et al., 2025b) under identical evaluation settings (Appendix E.5). Joint OPSD leads all reported

Table 5: Spatial editing: scene metrics (upper), preservation and object metrics on isolated intervals (lower). Angular errors are in radians; level changes in dB. Bold marks the best score.

System	Scene spatial quality						
	GCC ↓	CRW ↓	FSAD ↓	DOA ↓	Az. MAE ↓	Ang. traj. RMSE ↓	wCS ↑
SwanWeave	30.77	41.11	0.04	1.11	1.06	1.25	0.27
Qwen3-Omni + DiT	20.45	26.82	0.02	0.86	0.85	1.01	0.36
AMBIT	**14.81**	**11.95**	**0.01**	**0.39**	**0.36**	**0.52**	**0.71**

System	Unedited source preservation			Object direction		
	Direction MAE ↓	W NMSE ↓	Abs. level change ↓	Spherical MAE ↓	Azimuth MAE ↓	Angular traj. RMSE ↓
SwanWeave	0.90	1.49	8.40	1.053	1.016	1.157
Qwen3-Omni + DiT	1.42	0.58	20.94	0.734	0.711	0.797
AMBIT	**0.17**	**0.44**	**1.65**	**0.200**	**0.154**	**0.270**

metrics (Table 6). DiT OPSD improves every metric over SFT, whereas AR OPSD improves only LSD and FSAD. These gains support jointly refining planning and rendering for editing.

A further question is whether planning refinement benefits rendering under student plans. We contrast joint OPSD with branch supervision variants. AR OPSD retains GT plans for flow supervision; DiT OPSD distills rendering at states reached under student plans while retaining GT token supervision. Joint OPSD adds planning distillation on student prefixes; all configurations update the shared Transformer. Joint OPSD further improves classifier KL and CRW over DiT OPSD. These gains support refining both decisions and execution on student contexts: student plans determine DiT conditions and supervised states, including the effects of planner imperfections. Rendering gradients affect subsequent planning through shared parameters, while updated plans change later rendering trajectories, coupling scene decisions and acoustic realization across fresh student rollouts.

Table 6: Editing quality after post-training. AR OPSD and DiT OPSD retain GT supervision in the other branch. Bold and underlined values indicate the best and second-best scores per metric.

System	Content quality					Reconstruction and binaural quality			
	Paired CLAP ↑	FD-CLAP ↓	FAD ↓	FD-PANNs ↓	KL ↓	LSD ↓	GCC ↓	CRW ↓	FSAD ↓
AMBIT	0.839	0.047	0.242	6.012	0.455	1.268	14.805	11.947	0.0142
AMBIT + SFT	0.855	0.042	0.232	5.398	0.402	1.110	11.897	7.518	0.0119
AMBIT + ReFL	0.853	0.044	0.233	5.505	0.391	1.146	12.053	7.845	<u>0.0116</u>
AMBIT + GRPO	0.851	0.043	0.228	5.553	0.417	1.149	12.553	8.923	0.0117
AMBIT + AR OPSD	0.852	0.042	0.237	5.472	0.413	1.109	11.903	8.193	0.0117
AMBIT + DiT OPSD	<u>0.860</u>	<u>0.041</u>	<u>0.223</u>	<u>5.314</u>	<u>0.385</u>	<u>1.091</u>	<u>11.606</u>	<u>6.766</u>	0.0117
AMBIT + OPSD	**0.867**	**0.040**	**0.211**	**5.282**	**0.351**	**1.076**	**11.130**	**5.854**	**0.0115**

5 CONCLUSION AND LIMITATIONS

We presented AMBIT, a unified Transfusion model for native FOA generation and editing. Executable SCENEPLANs translate user intent into explicit source content, timing, and motion controls. Contrastive features ground requested edits in sources, while full reference latents supply acoustic details absent from the plan. Our continuous AMBIT VAE pairs omnidirectional reconstruction with inter-channel supervision to preserve content and spatial geometry. Joint Transfusion OPSD supervises planning on student prefixes and rendering under the resulting student plans, exposing the renderer to the conditions under which it must render at inference. Shared updates couple their refinement across fresh student rollouts. The results support treating source decisions and acoustic execution as a coupled learning problem for spatial audio authoring. Our evaluation covers short FOA scenes; longer scenes and acoustic consistency across repeated edits require further evaluation.

AI USE STATEMENT

We used generative AI tools to assist with writing and polishing the manuscript and to identify potentially relevant prior work. We also used Qwen3-Omni during dataset construction to generate captions for real recordings and source-level ScenePlan pseudo-annotations where annotations were unavailable, as described in Appendix A. The authors reviewed all the materials and take responsibility for the final text, citations, methods, experiments, and results.

ETHICS STATEMENT

AMBIT enables controllable spatial audio generation and editing for applications such as immersive media and sound design. These capabilities could also be misused to fabricate speech, alter acoustic events, or misrepresent their spatial context. We encourage clear disclosure of generated or edited audio and discourage deceptive or non-consensual applications. Our data combine recordings from existing datasets with synthetically rendered spatial scenes; their sources and processing are documented in Appendix A. Use and redistribution of third-party data remain subject to their original licenses and applicable privacy and consent requirements. Our subjective listening study involved 20 student participants, all of whom provided informed consent before participation, including consent to record their audio preferences and ratings for research purposes (Appendix D.4).

REPRODUCIBILITY STATEMENT

Our model architecture and objectives are described in Section 3, with the executable SCENEPLAN representation, FOA autoencoder, reference encoder, and shared planning and rendering backbone detailed in Appendix B. Appendix A documents the data sources, synthetic scene construction, annotation procedures, and dataset partitions. Training and inference settings are provided in Appendix C, while teacher construction, on-policy distillation objectives, supervision variants, and post-training controls are specified in Appendix E. Section 4 and Appendix D describe the baselines, evaluation protocols, metric definitions, and subjective listening study. To support independent verification and further research, we will publicly release our training and inference code, model checkpoints, and evaluation scripts upon publication.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. Musiclm: Generating music from text, 2023. URL https://arxiv.org/abs/2301.11325.

Evelina Bakhturina, Vitaly Lavrukhin, Boris Ginsburg, and Yang Zhang. Hi-fi multi-speaker english tts dataset. *arXiv preprint arXiv:2104.01497*, 2021.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725. IEEE, 2020.

William Chen, Prem Seetharaman, Rithesh Kumar, Oriol Nieto, Shinji Watanabe, Justin Salamon, and Zeyu Jin. Audiochat: Unified audio storytelling, editing, and understanding with transfusion forcing. *arXiv preprint arXiv:2602.17097*, 2026.

Ziyang Chen, David F Fouhey, and Andrew Owens. Sound localization by self-supervised time delay estimation. In *European Conference on Computer Vision*, pp. 489–508. Springer, 2022.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

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

Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Linfeng Feng, Lei Zhao, Boyu Zhu, Xiao-Lei Zhang, and Xuelong Li. Audiospa: Spatializing sound events with text. *arXiv preprint arXiv:2502.11219*, 2025.

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. FSD50K: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852, 2022.

Sanchit Gandhi, Patrick Von Platen, and Alexander M Rush. Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling. *arXiv preprint arXiv:2311.00430*, 2023.

Jort F Gemmeke, Daniel PW 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)*, pp. 776–780. 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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, et al. Qwen3-tts technical report. *arXiv preprint arXiv:2601.15621*, 2026.

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.

Chia-Yu Hung, Navonil Majumder, Zhifeng Kong, Ambuj Mehrish, Amir Zadeh, Chuan Li, Rafael Valle, Bryan Catanzaro, and Soujanya Poria. Tangoflux: Super fast and faithful text to audio generation with flow matching and clap-ranked preference optimization. In *International Conference on Learning Representations*, volume 2026, pp. 150793–150816, 2026.

Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, Ziang Zhang, Xiaoda Yang, Rongjie Huang, Yidi Jiang, Qian Chen, Siqi Zheng, and Zhou Zhao. WavTokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 93809–93826, 2025. URL `https://proceedings.iclr.cc/paper_files/paper/2025/file/ea1f5f0878d43ff4fb8bf64ef4a2326c-Paper-Conference.pdf`.

Yuhang Jia, Yang Chen, Jinghua Zhao, Shiwan Zhao, Wenjia Zeng, Yong Chen, and Yong Qin. Audioeditor: A training-free diffusion-based audio editing framework. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Dengyang Jiang, Xin Jin, Dongyang Liu, Zanyi Wang, Mingzhe Zheng, Ruoyi Du, Xiangpeng Yang, Qilong Wu, Zhen Li, Peng Gao, et al. D-opsd: On-policy self-distillation for continuously tuning step-distilled diffusion models. *arXiv preprint arXiv:2605.05204*, 2026a.

Yuxuan Jiang, Zehua Chen, Zeqian Ju, Chang Li, Weibei Dou, and Jun Zhu. Freeaudio: Training-free timing planning for controllable long-form text-to-audio generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 9871–9880, 2025.

Yuxuan Jiang, Zehua Chen, Zeqian Ju, Yusheng Dai, Weibei Dou, and Jun Zhu. Controlaudio: Tackling text-guided, timing-indicated and intelligible audio generation via progressive diffusion modeling. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1394–1413, 2026b.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 119–132, 2019.

Jaeyeon Kim, Heeseung Yun, and Gunhee Kim. Visage: Video-to-spatial audio generation. In *International Conference on Learning Representations*, volume 2025, pp. 46404–46424, 2025.

Jaeyeon Kim, Heeseung Yun, and Gunhee Kim. Towards scene-aware video-to-spatial audio generation. *International Journal of Computer Vision*, 134(4):184, 2026.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved rvqgan. *Advances in neural information processing systems*, 36:27980–27993, 2023.

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.

Zitong Lan, Yiduo Hao, and Mingmin Zhao. Guiding audio editing with audio language model. *arXiv preprint arXiv:2509.21625*, 2025.

Ryan Langman, Xuesong Yang, Paarth Neekhara, Shehzeen Hussain, Edresson Casanova, Evelina Bakhturina, and Jason Li. Hifitts-2: A large-scale high bandwidth speech dataset. *arXiv preprint arXiv:2506.04152*, 2025.

Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey. Sdr–half-baked or well done? In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 626–630. IEEE, 2019.

Ke Lei, Chenyuhao Wen, Yu Zhang, Wenxiang Guo, Changhao Pan, Sashuai Zhou, Yongshi Li, Ruiqi Li, Ruofan Hu, Haorui Xu, et al. One-stage multi-task instruction-guided 3d spatial audio editing. *arXiv preprint arXiv:2609.04975*, 2026a. URL https://arxiv.org/abs/2609.04975v2.

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*, 2026b.

Xiquan Li, Junxi Liu, Wenxi Chen, Haina Zhu, Ziyang Ma, and Xie Chen. Resonate: Reinforcing text-to-audio generation via online feedback from large audio language models. *arXiv preprint arXiv:2603.11661*, 2026.

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. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2871–2883, 2024.

Hongyu Liu, Chun Wang, Feng Gao, Xuanhua He, Yue Ma, Ziyu Wan, Yong Zhang, Xiaoming Wei, and Qifeng Chen. Opsd-v: On-policy self-distillation for post-training few-step autoregressive video generators. *arXiv preprint arXiv:2607.08766*, 2026.

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 *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 39060–39084. PMLR, 2025a. URL https://proceedings.mlr.press/v267/liu25as.html.

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. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 40783–40818. Curran Associates, Inc., 2025b. doi: 10.52202/085713-1362. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/3a10c46572628d58cb44fb705f25cbbf-Paper-Conference.pdf.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Xubo Liu, Zhongkai Zhu, Haohe Liu, Yi Yuan, Qiushi Huang, Meng Cui, Jinhua Liang, Yin Cao, Qiuqiang Kong, Mark D Plumbley, et al. Wavjourney: Compositional audio creation with large language models. *IEEE Transactions on Audio, Speech and Language Processing*, 33:2830–2844, 2025c.

Dan Lyth and Simon King. Natural language guidance of high-fidelity text-to-speech with synthetic annotations. *arXiv preprint arXiv:2402.01912*, 2024.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7739–7751. 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. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 564–572, 2024.

Hila Manor and Tomer Michaeli. Zero-shot unsupervised and text-based audio editing using ddpm inversion. *arXiv preprint arXiv:2402.10009*, 2024.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

Tianrui Pan, Jie Liu, Zewen Huang, Jie Tang, and Gangshan Wu. In-the-wild audio spatialization with flexible text-guided localization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1989–2001, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Ville Pulkki. Directional audio coding in spatial sound reproduction and stereo upmixing. In *Proc. of the AES 28th Int. Conf, Pitea, Sweden*, 2006.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *arXiv preprint arXiv:2204.02152*, 2022.

Miguel Sarabia, Elena Menyaylenko, Alessandro Toso, Skyler Seto, Zakaria Aldeneh, Shadi Pirhosseinloo, Luca Zappella, Barry-John Theobald, Nicholas Apostoloff, and Jonathan Sheaffer. Spatial librispeech: An augmented dataset for spatial audio learning. *arXiv preprint arXiv:2308.09514*, 2023.

Robin Scheibler, Eric Bezzam, and Ivan Dokmanić. Pyroomacoustics: A python package for audio room simulation and array processing algorithms. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 351–355. IEEE, 2018.

Kentaro Seki, Yuki Okamoto, Kouei Yamaoka, Yuki Saito, Shinnosuke Takamichi, and Hiroshi Saruwatari. Spatial-clap: Learning spatially-aware audio–text embeddings for multi-source conditions. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 14742–14746. IEEE, 2026.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Kazuki Shimada, Archontis Politis, Parthasaarathy Sudarsanam, Daniel A. Krause, Kengo Uchida, Sharath Adavanne, Aapo Hakala, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, Tuomas Virtanen, and Yuki Mitsufuji. STARSS23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events. In *Advances in Neural Information Processing Systems*, volume 36, pp. 72931–72957, 2023. doi: 10.52202/075280-3189. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/e6c9671ed3b3106b71cafda3ba225c1a-Abstract-Datasets_and_Benchmarks.html.

Parthasaarathy Sudarsanam, Sebastian Braun, and Hannes Gamper. Foa tokenizer: Low-bitrate neural codec for first order ambisonics with spatial consistency loss. *arXiv preprint arXiv:2510.22241*, 2025.

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.

Ye Tao, Wen Wu, Chao Zhang, Mengyue Wu, Shuai Wang, and Xuenan Xu. Mmedit: A unified framework for multi-type audio editing via audio language model. *arXiv preprint arXiv:2512.20339*, 2025.

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, Ruibin Yuan, Jiahao Pan, Qifeng Liu, Xu Tan, Qifeng Chen, Wei Xue, and Yike Guo. Vidmuse: A simple video-to-music generation framework with long-short-term modeling. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18782–18793, 2025.

Zeyue Tian, Lei Ke, Zhaoyang Liu, Ruibin Yuan, Liumeng Xue, Yujiu Yang, Weijia Chen, Xu Tan, Qifeng Chen, Wei Xue, et al. Audiox-turbo: A unified framework for efficient anything-to-audio generation. *arXiv preprint arXiv:2606.12555*, 2026a.

Zeyue Tian, Zhaoyang Liu, Yizhu Jin, Ruibin Yuan, Liumeng Xue, Xu Tan, Qifeng Chen, Wei Xue, and Yike Guo. AudioX: A unified framework for anything-to-audio generation. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 110625–110651, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/b36dc39b319ba6ba2a0fd7601951efb4-Paper-Conference.pdf.

Zeyue Tian, Binxin Yang, Zhaoyang Liu, Jiexuan Zhang, Ruibin Yuan, Hubery Yin, Qifeng Chen, Chen Li, Jing Lyu, Wei Xue, et al. Audio-omni: Extending multi-modal understanding to versatile audio generation and editing. *arXiv preprint arXiv:2604.10708*, 2026c.

Yuancheng Wang, Zeqian Ju, Xu Tan, Lei He, Zhizheng Wu, Jiang Bian, et al. Audit: Audio editing by following instructions with latent diffusion models. *Advances in Neural Information Processing Systems*, 36:71340–71357, 2023.

Yuyue Wang, Xihua Wang, Xin Cheng, Yijing Chen, and Ruihua Song. Unified synthesis of compositional speech and sound from free-form text prompts. *arXiv preprint arXiv:2605.28063*, 2026.

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.

Zeyu Xie, Xuenan Xu, Zhizheng Wu, and Mengyue Wu. Picoaudio: Enabling precise timestamp and frequency controllability of audio events in text-to-audio generation. *arXiv preprint arXiv:2407.02869*, 2024.

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. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

Dongchao Yang, Yuanyuan Wang, Dading Chong, Songxiang Liu, Xixin Wu, and Helen Meng. Uniaudio 2.0: A unified audio language model with text-aligned factorized audio tokenization. *arXiv preprint arXiv:2602.04683*, 2026.

Lingfeng Yao, Chenpei Huang, Xingke Yang, Ziye Geng, Changqing Luo, Hao Wang, Jiang Liu, and Miao Pan. Physwave: Physics-guided latent diffusion models for controllable spatial audio generation. *arXiv preprint arXiv:2608.29549*, 2026.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*, 2019.

Shiyi Zhang, Mushui Liu, Yunze Tong, Wanggui He, Siyu Zou, Jinlong Liu, Yunlong Yu, Jian Song, Hao Jiang, Pipei Huang, et al. Self-opd: On-policy distillation for flow matching models without teacher. *arXiv preprint arXiv:2608.26872*, 2026.

Lei Zhao, Sizhou Chen, Linfeng Feng, Jichao Zhang, Xiao-Lei Zhang, Chi Zhang, and Xuelong Li. Dualspec: Text-to-spatial-audio generation via dual-spectrogram guided diffusion model. *IEEE Transactions on Multimedia*, 2026a.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026b.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *International Conference on Learning Representations*, volume 2025, pp. 6446–6469, 2025.

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*, 2026.

A DATASET DETAILS

Table 7 lists the SCENEPLAN field ranges for scene and source attributes (Section 3.1). Synthetic scenes are rendered at 44.1 kHz using ACN $[W, Y, Z, X]$ channel order and SN3D normalization. Coordinates point forward ($+X$), left ($+Y$), and up ($+Z$); positive azimuth turns left.

Table 7: SCENEPLAN field ranges. Times, angles, and distances use seconds, degrees, and metres.

Field	Representation and constraints
Duration	At most 648 latent frames at a 1,024-sample hop (approximately 15.05 s)
Room	Dry, moderate, reverberant, or outdoor
Sources	One to four plan local source IDs.
Content	Source kind, description, and verbatim transcript for speech
Activity	Interval $[o_i, f_i)$ with $0 \le o_i < f_i \le d$
Trajectory	Static pose (θ, ψ, ρ) or linear motion with 2 to 8 keyframes
Gain	Recorded in dB; the planning codec decodes a constant 0 dB

Real recordings and captioning. The generation dataset combines synthetic scenes with a collection of 80,000 FOA clips from Sphere360 (Liu et al., 2025a) and 3,809 clips from STARSS23 (Shimada et al., 2023). We segment STARSS23 recordings into consecutive 10 s clips, discarding shorter tails and cropping available videos to the corresponding intervals. Qwen3-Omni (Xu et al., 2025) receives synchronized video and the omnidirectional W channel as mono audio to produce concise captions; clips without video use audio alone. The full FOA signals are retained in their assigned data partitions. Available STARSS23 event and spatial annotations are converted into SCENEPLAN fields. For Sphere360 and unannotated STARSS23 clips, a second Qwen3-Omni pass expands each caption into source descriptions, activity intervals, and trajectories under the SCENEPLAN schema. It retains explicit cues and infers missing times, directions, and distances as pseudo annotations. Training plans supervise AR prediction and provide compiled conditions for DiT rendering.

Synthetic assets and rendering. Speech recordings come from LibriTTS (Zen et al., 2019), Hi-Fi TTS (Bakhturina et al., 2021), and HiFiTTS-2 (Langman et al., 2025); music and sound assets come from AudioCaps (Kim et al., 2019), AudioSet (Gemmeke et al., 2017), FSD50K (Fonseca et al., 2022), PicoAudio (Xie et al., 2024), VGGSound (Chen et al., 2020), and MusicCaps (Agostinelli et al., 2023). Pyroomacoustics (Scheibler et al., 2018) supplies four-channel room responses at trajectory keyframes. Source signals are placed at their onsets, convolved with these responses, and combined by interpolating adjacent keyframes. Static sources use one response. Scene gain targets -23 dBFS W channel RMS, subject to a -1 dBFS true peak ceiling.

For synthetic scene construction, the listener is horizontally centred at height $\min(1.7, H/2)$ m. Initial azimuth, elevation, and distance are sampled from $[-180, 180)^\circ$, $[-22, 22]^\circ$, and $[0.9, 2.2]$ m, with a 0.3 m wall margin. Table 8 gives the room distributions used for scene rendering.

Table 8: Room sampling ranges for synthetic scenes. Outdoor scenes use no reflections.

Class	Dimensions $L \times W \times H$ (m)	RT60 (s)	Image order
Dry	$[6, 8] \times [5, 7] \times [3, 4]$	$[0.16, 0.28]$	5 to 7
Moderate	$[8, 12] \times [7, 10] \times [3.2, 4.8]$	$[0.38, 0.62]$	6 to 9
Reverberant	$[12, 18] \times [10, 15] \times [4.5, 7]$	$[0.78, 1.20]$	6 to 9
Outdoor	$[10, 16] \times [10, 16] \times [5, 8]$	—	0

Editing and partitions. Requests express source content and qualitative spatial or temporal constraints, retaining literal speech transcripts. Each edit applies SPATIAL/TIME RELOCATE, STATIC→LINEAR, LINEAR→STATIC, ADD, or REMOVE to one source. Dataset annotations retain source correspondence and assets across an editing pair; serialized plans use the local indexing described in Appendix B.5. Room and scene duration remain fixed, and source and target share calibration and master gain apart from peak ceiling attenuation. Synthetic generation partitions are disjoint in source assets; editing inherits these partitions and draws addition assets from the same partition. Sphere360 recordings are used only for generation training and validation. The 8,000-example generation test set includes 50 FOA clips of 10 s each from the annotated STARSS23

dev-test split. These clips are included in the reported generation results. Splits are assigned by original recording or video, with all derived clips kept in the same partition. The original recordings underlying the 50 test clips, including all other clips derived from those recordings, are excluded from training and validation. Real recordings outside these held-out recording groups are used for generation training and validation, following the same training-to-validation ratio as synthetic scenes. Table 9 summarizes the overall partitions for both tasks.

Table 9: Dataset partitions. Generation training combines synthetic scenes and real recordings. The 8k generation test examples include 50 FOA clips of 10 s each from the annotated STARSS23 dev-test split.

Partition	Train	Validation	Test
Generation examples	1.6M	32k	8k
Editing pairs	1M	20k	5k

AMBIT VAE training data. The VAE corpus contains 1M training clips, 20k validation clips, and 2,000 test clips, combining synthetic spatial audio with FOA clips from Spatial LibriSpeech (Sarabia et al., 2023). For music and sound, mono assets are rendered with Pyroomacoustics (Scheibler et al., 2018) under varied room acoustics, source positions, and motion, then mixed into FOA scenes. Speech recordings are spatialized into FOA scenes. For clips already provided in FOA format, we retain their original channels.

AMBIT CLAP training data. We train AMBIT CLAP on 1M source–target audio pairs from the editing training split. Each recording is a synthetic FOA scene: music and sound assets are rendered with Pyroomacoustics (Scheibler et al., 2018) and mixed under a shared room and duration, while speech recordings are spatialized into FOA. From the paired SCENEPLANs, we construct a semantic caption of source content and a scene caption that binds each source to its activity, trajectory, and gain. Relocate and motion edits keep the semantic caption fixed and change the scene caption; additions and removals change both. We construct up to four negative captions per recording by swapping trajectories or activity intervals between distinguishable sources.

B MODEL DETAILS

B.1 SCENEPLAN CONDITIONING

Frozen Qwen3.5-0.8B features of width 1,024 are projected to 768-dimensional cross-attention tokens for semantic clauses. The compiler rejects text over 512 tokens to preserve speech. Each slot concatenates 32 activity and 32 pose embedding channels, giving 256 channels across four sources. A frame is active whenever its sample interval overlaps the source activity interval.

At latent frame midpoints clipped to source activity, elevation and distance interpolate linearly between adjacent keyframes; azimuth uses the shortest signed increment in $[-\pi, \pi)$. Endpoint poses extend beyond the keyframe range. Only then do we compute $(\sin\theta, \cos\theta, \sin\psi, \cos\psi, \log(1+\rho))$ and apply a $5 \to 64 \to 32$ MLP with SiLU activations. Inactive source blocks are zero; dropped conditions use learned unknown embeddings. Text and gain are excluded from spatial controls.

B.2 AMBIT VARIATIONAL AUTOENCODER

Architecture and KL reduction. The Oobleck based VAE uses base width 128, channel multipliers $(1, 2, 4, 8, 16)$, strides $(2, 2, 4, 8, 8)$, and Snake activations. Its 64 latent channels have a hop of 1,024 samples, with 40 channels assigned to z^W and 24 to z^R. For posterior scale $\sigma = \mathrm{softplus}(s) + 10^{-4}$, the group penalties in Equation 2 are computed as follows:

$$K_g = \frac{1}{BT} \sum_{b,t} \sum_{c \in g} \left(\mu_{bct}^2 + \sigma_{bct}^2 - \log \sigma_{bct}^2 - 1 \right). \tag{11}$$

Here B, T, and g denote batch size, latent sequence length, and channel group. KL terms are summed over channels and averaged over examples and time steps, with the conventional $1/2$ absorbed into their coefficients. The first four ablation configurations use $10^{-4} K_{\mathrm{all}}$ over all 64 channels. The

full configuration replaces this uniform penalty with $\beta_W K_W + \beta_R K_R$, using $\beta_W = 5 \times 10^{-5}$ and $\beta_R = 2 \times 10^{-4}$.

Reconstruction settings. STFT reconstruction uses equally weighted spectral convergence and log magnitude terms with perceptual K weighting and an overall coefficient of 1. Hann windows of 32 to 2,048 samples in powers of two use quarter window hops. The multi scale STFT discriminator uses adversarial and feature matching coefficients of 0.1 and 5.0; waveform L_1 is disabled. Phase supervision combines instantaneous frequency, group delay, and normalized complex distance with unit internal weights and target weight 0.08. The full configuration uses cepstral, SI-SDR, and high frequency overshoot coefficients of 0.15, 0.003, and 0.05. Auxiliary W recovery reuses the decoder with z^R zeroed, the configured spectral objective weighted by 0.3, and SI-SDR by 0.003.

Base configuration and cumulative additions. Phase (base) denotes the reconstruction baseline with spectral phase supervision. It includes the multiresolution STFT, phase, high frequency overshoot, adversarial, and feature matching terms, together with uniform KL regularization of 10^{-4} across all 64 latent channels. Its phase objective combines instantaneous frequency, group delay, and normalized complex spectral differences within each channel. The next configuration adds SCM, which supervises covariance relationships between channels; its stabilized phase alignment is distinct from the base spectral phase objective. Cepstral and SI-SDR losses are added in the following two configurations. The full configuration replaces uniform KL with grouped KL and adds auxiliary W reconstruction, while retaining all preceding loss terms for FOA reconstruction.

The table label Phase (base) names this initial ablation configuration. In Equation 4, $\mathcal{L}_{\text{base}}$ collects the reconstruction and adversarial terms active in the given configuration; the full model includes the added cepstral and SI-SDR terms in this collection.

Spatial covariance matching. SCM supervises relationships between FOA channels by matching statistics derived from their local spatial covariance matrices. For four channel STFT vectors $X(f,\tau) \in \mathbb{C}^4$, we compute $C = \langle XX^{\mathsf{H}} \rangle_5$ over five neighboring frames, where H denotes conjugate transpose. We use Hann windows, FFT sizes 512 and 2,048, quarter window hops, and no STFT centering. Temporal averaging excludes padded positions at sequence boundaries.

Let $e_{brf\tau} = \operatorname{tr} C_{br}(f,\tau)$ denote reference trace energy for example b and resolution r, and let N_r denote the number of time-frequency bins. The shared stabilization floor is

$$\delta_{br} = \max\left(10^{-12}, \frac{10^{-3}}{N_r} \sum_{f,\tau} e_{brf\tau}\right). \tag{12}$$

It is constant across bins within each example and resolution and is used for both prediction and reference. Suppressing these indices, the statistics in Equation 3 are

$$p_i = \frac{C_{ii}}{\operatorname{tr} C + \delta}, \qquad \gamma_{ij} = \frac{|C_{ij}|}{\sqrt{(C_{ii}+\delta)(C_{jj}+\delta)}}, \qquad U_{ij} = \frac{C_{ij}}{|C_{ij}| + \delta}. \tag{13}$$

Hatted quantities use reconstructed covariance and the same reference floor. The overline in Equation 3 denotes complex conjugation. Since $|U_{ij}| < 1$, the final term is stabilized phase alignment with magnitude dependence; it need not vanish for identical covariance. This term belongs to SCM and is separate from the spectral phase objective retained in every cumulative ablation configuration.

With $\epsilon = 10^{-8}$, reference weights are

$$\begin{aligned} w_{brf\tau} &= \frac{e_{brf\tau}}{\sum_{f',\tau'} e_{brf'\tau'} + \epsilon}, \\ v_{brijf\tau} &= w_{brf\tau}\ \operatorname{clip}(\gamma_{brijf\tau}, 0, 1)\left(0.1 + \frac{0.9}{1 + (f/1500\,\text{Hz})^2}\right). \end{aligned} \tag{14}$$

Here f is physical frequency. These weights are determined by the reference; the coherence factor is detached. For per bin errors A and pairwise errors B, the reductions are

$$\langle A \rangle_w = \frac{\sum_{b,r,f,\tau} w_{brf\tau} A_{brf\tau}}{\max(\sum_{b,r,f,\tau} w_{brf\tau}, \epsilon)}, \qquad \langle B \rangle_v = \frac{\sum_{b,r,f,\tau} \sum_{i<j} v_{brijf\tau} B_{brijf\tau}}{\max(\sum_{b,r,f,\tau} \sum_{i<j} v_{brijf\tau}, \epsilon)}. \tag{15}$$

The energy and coherence errors sum over four channels and six unordered channel pairs, respectively. Phase alignment instead uses the weighted mean over all six pairs and all examples, resolutions, and bins. Its weights emphasize coherent observations and lower frequencies. All three internal loss coefficients are 1, and the outer coefficient is $\lambda_{\mathrm{SCM}} = 0.015$.

B.3 AMBIT CLAP REFERENCE FEATURES

VAE latents pooled with stride 4, together with mean/RMS features and positional encodings, enter a Transformer with six layers, width 384, and six attention heads. It produces temporal features and normalized semantic and scene embeddings of dimensions 512 and 256, respectively, aligned with separate projections of frozen text features. For head $h \in \{\mathrm{sem}, \mathrm{scene}\}$, let a_i^h and t_j^h denote normalized audio and text embeddings. Scaled cosine scores $s_{ij}^h = \alpha_h (a_i^h)^\top t_j^h$, with logit scale α_h, give the audio to text loss

$$\ell_h^{A \to T} = -\frac{1}{B} \sum_i \frac{1}{|\mathcal{P}_h(i)|} \sum_{j \in \mathcal{P}_h(i)} \log \frac{\exp(s_{ij}^h)}{\sum_{k \in \mathcal{V}_h(i)} \exp(s_{ik}^h)}. \tag{16}$$

Here B is the batch size, $\mathcal{P}_h(i)$ contains positive text matches for audio i, and $\mathcal{V}_h(i)$ contains eligible text candidates, excluding ambiguous partial content overlaps.

Semantic positives match source content and counts. Scene positives additionally match each source's spatial and temporal attributes and room properties. Spatial edits that preserve content are semantic positives and scene negatives. Swapped trajectories or activity intervals provide hard negatives for source attribute binding.

We average the audio to text and text to audio losses for each head to obtain $\mathcal{L}_{\mathrm{sem}}$ and $\mathcal{L}_{\mathrm{scene}}$. The complete objective supplements Equation 5 with binding, editing pair, and source attribute terms:

$$\mathcal{L}_{\text{AMBIT-CLAP}} = \mathcal{L}_{\mathrm{align}} + \lambda_{\mathrm{bind}} \mathcal{L}_{\mathrm{bind}} + \lambda_{\mathrm{pair}} \mathcal{L}_{\mathrm{pair}} + \lambda_{\mathrm{event}} \mathcal{L}_{\mathrm{event}}. \tag{17}$$

The selected configuration uses $\lambda_{\mathrm{scene}} = 1$, $\lambda_{\mathrm{bind}} = \lambda_{\mathrm{pair}} = 5$, and $\lambda_{\mathrm{event}} = 0.1$. Binding and pair terms rank matched descriptions above attribute swapped or crossed descriptions; four learned source queries supervise content and attributes under a shared source assignment. After pretraining, the encoder is frozen and supplies global and temporal features $F(z_s)$ to the editing planner. It is separate from the audio CLAP model used for content quality evaluation.

B.4 GENERATION PLANNING AND RENDERING

The renderer has 15 Transformer blocks, width 1,024, and 16 attention heads. Its input concatenates 64 noisy latent channels, 256 spatial condition channels, and 64 reference latent channels. Generation sets the reference channels to zero, retaining the same 384 channel input interface used for editing. Speech active frames receive weight $w_t = 3$, compared with $w_t = 1$ elsewhere, in the normalized residual loss for rendering

$$\|r\|_{m,w}^2 = \frac{\sum_{c,t} m_t w_t r_{ct}^2}{64 \sum_t m_t w_t}. \tag{18}$$

For sound-only clips, a loss with weight 0.05 compares latent differences at lags 1, 2, and 4:

$$\mathcal{L}_\Delta = \frac{1}{3} \sum_{\ell \in \{1,2,4\}} \mathrm{SmoothL1}(\Delta_\ell \widehat{z}_0, \Delta_\ell z_0), \qquad \widehat{z}_0 = z_t - t v_\theta. \tag{19}$$

Only coordinate pairs valid at both times contribute to $(\Delta_\ell z)_{:,t} = z_{:,t+\ell} - z_{:,t}$. Semantic and spatial conditions are independently dropped with probability 0.15 for each training example.

Autoregressive plan completion. The AR planner translates raw user requests into complete, executable SCENEPLANs for speech, music, environmental sounds, and their combinations. It determines source inventory and content, specifies spatial positions and motion, and completes onset and offset times, turning natural language descriptions into structured scene specifications for rendering. Inventory, span selection, and qualitative heads support this autoregressive process, with completed fields conditioning subsequent plan tokens. Table 10 summarizes how these source and scene attributes are produced.

The codec has a vocabulary of 4,096 tokens, including 1,536 text pieces, 649 frame values, 1° angular bins, and 256 logarithmic distance bins over 0.1 to 50 m. Onsets are rounded down and offsets are rounded up to the latent-frame grid. Grammar constrained greedy decoding serializes scene and source fields. Planning LoRA has rank 8 and $\alpha = 8$ and is active only in AR mode. Within this planner, head outputs determine their corresponding fields, numerical rules complete executable values, and remaining legal choices use AR logits.

Table 10: Source and scene fields completed by the AR planner.

Plan fields	AR mechanism	Prediction and completion
Room and duration	AR logits	Four room classes and 1 to 648 duration frames.
Source inventory	Inventory head	Predict source count and sound, music, or speech categories.
Description and transcript	Learned span selectors	Select literal request spans and encode them with the text codec.
Motion and azimuth	Qualitative head	Static or linear; eight compass categories for start and end directions.
Onset and offset	Qualitative head	Five relative time phases; select a feasible phase pair and complete integer frame values.
Radial motion	Qualitative head	Free, approaching, or receding; complete endpoint distances.
Structural tokens and gain	Deterministic rules	Grammar supplies required delimiters; decoded gain is fixed at 0 dB.

Spatial and temporal completion. For generation, deterministic rules instantiate learned qualitative controls using seed 42 and keys for each request, source, and field. The numerical defaults differ from the data sampling ranges in Appendix A. Eight compass centres start at front (0°) and advance leftward in 45° steps. Azimuths lie within 8° of the selected centre before codec quantization. Initial distance lies in $[2.2, 3.25)$ m. Approaching and receding linear sources decrease or increase it by $[0.75, 1)$ m; free radial motion keeps distance fixed. Static sources use one position selected by summed endpoint scores. Coincident linear endpoints receive a 6° azimuth offset.

Beginning, early, middle, late, and ending correspond to possible duration fractions like $[0, 0.2]$, $[0.1, 0.4]$, $[0.3, 0.7]$, $[0.6, 0.9]$, and $[0.8, 1]$, respectively. The highest scoring feasible onset and offset phase pair is selected, and integer-valued onset and offset frames are then assigned with onset before offset and at least ten frames when feasible. Speech uses values within two frames of the feasible outer boundaries. These values fix numerical tokens in the AR stream; editing prediction is detailed in Appendix B.5.

B.5 Editing planner and renderer

Editing fills the 64 reference channels with source audio latents, using the same input projection and Transformer blocks as generation. The projection columns associated with source latents are initialized to zero and learned during training. The AR prefix combines native latents with global and temporal AMBIT CLAP features:

$$h_{s,\ell} = h_{s,\ell}^{\text{base}} + P_g[g_{\text{sem}}; g_{\text{scene}}] + [\text{Up}_4(P_{\text{temp}}H)]_\ell. \tag{20}$$

These projections start at zero; Up_4 repeats features over four latent frames. A two-layer decoder with source slots and a NEW slot supplies AR memory through gated residual attention.

Autoregressive edit completion. For editing, the same planner translates raw edit instructions, together with source audio, into complete target SCENEPLANs. It grounds the requested change in the source scene and completes the resulting source content, spatial configuration, and onset and offset times. The output jointly describes retained and modified sources, so a local instruction produces a complete scene specification for rendering. Source attributes and edit decisions supply memory that guides AR decoding of the updated plan. Duration follows the source audio, and decoded gain remains 0 dB.

Canonical target order. Targets are stably sorted by codec projected onset, source kind, and compact JSON of source content and attributes with sorted keys. Exact ties retain input order and have identical projected content. This ordering preserves acoustic fields and aligns AR targets,

paired rendering conditions, and edit annotations. Paired source correspondences survive reordering, keeping supervision attached to the same source. Inference uses the learned order.

Shared source supervision. Equation 7 uses $\lambda = 0.1$. Paired annotations link source attributes before and after editing. We minimize the field cost over injective assignments from occupied source records to predicted slots, including the kind cost of unused slots. Cost ties use the first assignment in lexicographic slot order. The same assignment supervises recognition, edit binding, preservation, modification, and target presence, with no independent rematching after editing.

Removals keep their assigned slot with REMOVE and an empty target; additions use the NEW slot and enter canonical order. Field masks distinguish preserved attributes from requested changes.

C TRAINING AND INFERENCE SETTINGS

AMBIT supports FOA generation and editing within a single Transfusion model, sharing Transformer blocks across autoregressive planning and rectified flow rendering. The complete student contains 332.6M trainable parameters, including the shared backbone, native task interfaces, and structured heads. Generation uses language instructions, while editing additionally conditions on source audio. Table 11 summarizes the learning rates for VAE pretraining and joint Transfusion optimization.

AMBIT VAE pretraining. AMBIT VAE is trained on approximately 1M FOA clips, including synthetic speech and sound scenes and Spatial LibriSpeech (Sarabia et al., 2023), using 4 s crops. All cumulative VAE ablations in Table 1 are trained from scratch for matched optimization steps on the same data split.

Joint Transfusion optimization. Generation and editing use the same Transfusion model, with shared Transformer blocks for AR planning and DiT rendering. Generation training uses construction plans for synthetic scenes and converted or caption derived SCENEPLAN annotations for real recordings, as described in Appendix A. During generation planner warm-up, the pretrained renderer remains frozen while planning adapters, heads, and LoRA (Hu et al., 2021) parameters are trained; LoRA is active only during planning. Subsequent joint optimization on editing pairs uses the paired objective in Section 3.3, updating the shared Transformer, native adapters, and structured heads. Paired rendering uses compiled ground truth target plans, whereas inference uses compiled student predictions. For editing, source latents condition both modes, while AMBIT CLAP features additionally condition the planner. The renderer's source latent input projection columns and CLAP feature projections are initialized to zero. The pretrained text backbones, AMBIT CLAP, and AMBIT VAE remain frozen and are excluded from the trainable parameter count.

Table 11: Learning rates for VAE pretraining and the joint optimization stage. All updates use AdamW.

Component	Optimization objective	Learning rates
AMBIT VAE	FOA reconstruction	VAE 1.5×10^{-4}; discriminator 3×10^{-4}
Transfusion	Joint AR and DiT training	Adapters and heads 10^{-4}; shared blocks 2×10^{-6}

Flow time sampling and inference. For generation training, flow times follow $t = 1 - (u - 0.075)/0.925$, where $u = \mathrm{sigmoid}(\xi)$ and $\xi \sim \mathcal{N}(0, 1)$, conditioned on $u \geq 0.075$. Paired editing training samples t uniformly from $[0, 1]$.

The unified model supports both tasks through AR planning, deterministic plan compilation, DiT rendering with Euler integration, and VAE decoding. Generation omits source audio conditioning. Editing uses source latents in both planning and rendering, with additional AMBIT CLAP features for planning (Section 3.3). Generation uses 100 Euler steps, guidance scale 3.0, and guidance rescaling 0.4; editing uses 20 Euler steps and guidance scale 1.0.

Within each task, planner comparisons hold the VAE and renderer weights, sampler settings, and noise seeds fixed. All compared planners produce plans for the same deterministic compiler and DiT rendering interface. OPSD further refines the shared student through joint planning and rendering supervision; its optimization settings and post-training controls appear in Appendix E.

D EVALUATION DETAILS AND ADDITIONAL RESULTS

D.1 EVALUATION PROTOCOL AND BASELINE INTERFACES

Generation and editing results use the 8k generation and 5k editing benchmarks, respectively. The AMBIT VAE is evaluated on 2,000 FOA clips. Music, Sound, and Speech categories can overlap; pooled spatial results count each scene once. Audio comparisons use executable plans and metric valid observations from the relevant benchmark; missing measurements are excluded rather than assigned zero error. External generators receive semantic descriptions, while AMBIT receives qualitative requests. Editing baselines use source audio and an instruction or target description according to their native interfaces. Qwen3-Omni is a planner for our DiT, distinct from the Audio-Omni editor: its source observation is the W waveform, whereas AMBIT uses native FOA features.

D.2 EXTERNAL AUDIO SPATIALIZATION

For BEWO + AS and TangoFlux + AS in Table 4, Qwen3.5-4B supplies the same executable plans to both systems. Each audio backend synthesizes source signals from the plan's semantic fields; these signals are downmixed to mono and placed within their predicted activity intervals. Pyroomacoustics renders the predicted positions and room classes in ACN/SN3D FOA, using fixed inference geometry rather than ground truth room dimensions. Moving sources crossfade room responses at 250ms intervals. Qwen3.5-4B + DiT instead passes its plan through our compiler and neural renderer. These comparisons assess complete planning and rendering systems with a common target waveform.

D.3 METRICS

The generation, editing, and VAE test sets contain 8,000 requests, 5,000 requests, and 2,000 clips, respectively. All systems use the same evaluation examples for each task and metric. Directional errors use a common set of frames with valid direction estimates in the reference and every compared output. Object metrics additionally use isolated source intervals.

Content and speech. Paired CLAP measures similarity between output and reference audio. FAD and feature based Fréchet distances compare embedding distributions; PANNs/PaSST KL compares predicted acoustic classes. Distributional scores use the evaluated features directly. Generation uses native W or arithmetic mean mono, peak normalized to -1 dBFS before resampling; editing uses mono downmixes of binauralized FOA without gain normalization. CLAP, PANNs, and VGGish operate at 48, 32, and 16 kHz. For clips longer than 10 s, CLAP window embeddings are normalized, averaged, and renormalized. Distil-Whisper large-v3 supplies transcriptions; generation WER/CER aggregate edit counts over reference lengths, while UTMOS averages individual clip scores.

Reconstruction and binaural quality. FOA is decoded with fixed MIT KEMAR filters using eight virtual speakers at azimuths $45, 135, 225, 315^\circ$ and elevations $\pm 40^\circ$; native stereo is unchanged. LSD is the framewise RMS difference in $\log_{10}$ power at 24 kHz, with a 1,024 sample Hann window and 240 sample hop. GCC is $1,000$ times the squared difference of clip mean interaural delays in milliseconds, estimated by GCC-PHAT at 16 kHz with a ± 1 ms search. CRW compares StereoCRW delay predictions; FSAD compares the distributions of its spatial audio features.

Scene direction and angular trajectories. The scene columns in Tables 4 and 5 compare output and GT waveform directions. Active intensity directions use 1,024 sample frames at 44.1 kHz. Valid frames require coherence at least 0.1, a nonzero direction, and FOA energy at least $\max(10^{-12}, 10^{-4} e_{\max})$ within each waveform. We intersect valid frames across GT and compared outputs; editing additionally restricts frames to target plan activity. For spherical angular difference d_t, wrapped azimuth difference $\Delta\theta_t \in [-\pi, \pi)$, and normalized GT energy weights a_t,

$$\mathrm{DOA} = \sum_t a_t d_t, \qquad \mathrm{AzMAE} = \sum_t a_t |\Delta\theta_t|, \qquad \mathrm{AngRMSE} = \sqrt{\sum_t a_t d_t^2}. \tag{21}$$

Here d_t is the arccosine of the clipped inner product between output and GT unit directions. Azimuth excludes undefined horizontal directions and renormalizes the weights. Overlapping sources con-

tribute mixture directions. Errors are in radians and scenes are averaged equally. Angular trajectory RMSE measures directional deviations over time without evaluating source distance.

Object direction and angular trajectories. The object columns use scenes with one to four GT sources. For each object, eligible frames contain exactly one active GT source; its interval mask is eroded by two frames at each boundary. Intersecting this mask with common valid direction frames gives $\mathcal{J}_{so}$ for object o in scene s. With GT energy weights $a_{so,t}$ normalized over this interval,

$$\begin{aligned} \mathrm{SphMAE}_{so} &= \sum_{t\in\mathcal{J}_{so}} a_{so,t} d_t, \qquad \mathrm{AzMAE}_{so} = \sum_{t\in\mathcal{J}_{so}} a_{so,t} |\Delta\theta_t|, \\ \mathrm{AngRMSE}_{so} &= \sqrt{\sum_{t\in\mathcal{J}_{so}} a_{so,t} d_t^2}. \end{aligned} \tag{22}$$

Azimuth excludes frames whose horizontal unit direction norm is at most 10^{-6} in either waveform and renormalizes the remaining weights. Object metrics compare actual output and GT FOA directions; they do not compare output directions with the predicted plan. For any object metric m_{so}, the reported average is

$$\overline{m} = \frac{1}{|\mathcal{V}|} \sum_{s\in\mathcal{V}} \frac{1}{|\mathcal{O}_s|} \sum_{o\in\mathcal{O}_s} m_{so}, \tag{23}$$

where $\mathcal{O}_s$ and $\mathcal{V}$ denote valid objects and scenes for each metric. Frames use GT energy weights. We weight valid objects equally within each scene and then average equally over all eligible scenes with one to four sources. Object RMSE is computed per object before this averaging. We apply no source separation, temporal alignment, or angular reassignment. The exclusive activity mask can retain reverberant tails despite its boundary guard.

Spatial feature similarity. We use PSELDNets activity coupled direction features at 24 kHz. For vectors a_{tc} and b_{tc} from output and reference,

$$\mathrm{wCS} = \frac{\sum_{t,c} a_{tc}^{\mathsf{T}} b_{tc}}{\sum_{t,c} \|a_{tc}\|_2 \|b_{tc}\|_2}. \tag{24}$$

This magnitude weighted cosine uses all 170 classes without thresholding or class selection. Padding frames are excluded; a zero denominator is treated as an undefined measurement.

Unedited source preservation. The preservation block of Table 5 compares output with source audio where the same unchanged source is the only annotated active source in both source and target plans. The mask excludes the union of edited source activity in both plans, then removes two frames from each retained interval boundary. For time relocation, both original and shifted intervals are excluded; for removal, the original interval remains excluded. Direction MAE uses common valid direction frames and source energy weighting. W NMSE divides squared waveform error by source energy over eligible samples; absolute level change is $|10\log_{10}(P_{\mathrm{out}}/P_{\mathrm{src}})|$, where P denotes mean W power on those samples. These masks establish exclusivity from activity annotations, without acoustic source separation; reverberant tails may remain despite the boundary guard.

VAE comparison. Table 1 reports codec reconstruction and cumulative objective ablations. The ablations follow Appendices B.2 and C, with a 24 kHz evaluation frontend. W SI-SDR and paired CLAP assess omnidirectional reconstruction; direction and intensity coherence errors assess spatial reconstruction. Spatial scoring uses FOA signals resampled to 24 kHz after a single delay estimated from W by GCC-PHAT is applied to all channels, retaining only their common temporal overlap.

Intensity coherence error. The development of this metric is inspired by FOA tokenizer (Sudarsanam et al., 2025) and DirAC (Pulkki, 2006). For each complete 40 ms frame with a 20 ms hop, let $\mathbf{d} = (Y, Z, X)^{\mathsf{T}}$ and let $\langle\cdot\rangle_k$ denote the uniform sample mean in frame k, without subtracting the

channel means. Define $e_k^W = \langle W^2 \rangle_k$, $e_k^D = \langle \|\mathbf{d}\|_2^2 \rangle_k$, and $\mathbf{i}_k = \langle W\mathbf{d} \rangle_k$. Coherence and its error are

$$r_k = \mathrm{clip}_{[0,1]} \left(\frac{\|\mathbf{i}_k\|_2}{\sqrt{\max(e_k^W e_k^D, 10^{-16})}} \right),$$
$$E_{\mathrm{IC}}(x, \widehat{x}) = \frac{\sum_k a_k E_k \, |\widehat{r}_k - r_k|}{\max(\sum_k a_k E_k, 10^{-12})}. \tag{25}$$

Here $E_k = e_k^W + e_k^D$ is the total reference channel energy. A frame is active when $E_k > \max(10^{-3} \max_l E_l, 10^{-10})$; a_k is its binary activity mask. Hatted statistics are computed independently from the reconstruction; the mask and aggregation weights use only the reference. We average E_{IC} equally across clips. This time domain statistic measures the magnitude of the normalized intensity vector, whereas SCM uses pairwise STFT covariance statistics. Direction error uses the angle between reference and reconstructed intensity vectors in degrees, weighted by $\|\mathbf{i}_k\|_2$ over active reference frames with $r_k \geq 0.1$. The additional threshold applies only to direction error; intensity coherence error uses all active reference frames, including frames of low coherence.

D.4 Subjective listening study

Participants and consent. We recruited 20 student participants from our university. All participants provided informed consent before the study, including consent to record their audio preferences and ratings for research purposes.

Cases and assignment. We prepared separate pools of 1,000 matched FOA cases for generation and editing, and randomly partitioned each pool into 50 non-overlapping sets of 20 cases. Each participant selected one available set per task without previewing its contents; no set was reused. Each case contained three system outputs for the same request. The generation panel compared AMBIT, BEWO + AS, and TangoFlux + AS; the editing panel compared AMBIT, SwanWeave, and Qwen3-Omni + DiT. The study covered 400 distinct cases per task and 2,400 system-output presentations in total, with one participant rating all three outputs for each case.

Table 12: Subjective evaluation of FOA generation and editing. All dimensions use a five-point scale; higher is better. Entries report mean ratings over valid responses. The highest mean in each column is highlighted in bold.

(a) Generation.

System	Quality MOS	Instruction adherence	Spatial adherence	Spatial naturalness
BEWO + AS	2.93	2.96	2.83	2.89
TangoFlux + AS	2.77	2.84	2.63	2.76
AMBIT	**3.11**	**3.25**	**3.37**	**3.03**

(b) Editing.

System	Quality MOS	Edit success	Unedited content preservation
SwanWeave	1.89	1.96	1.93
Qwen3-Omni + DiT	3.51	3.26	3.44
AMBIT	**3.64**	**3.57**	**3.71**

Listening procedure. Participants used headphones after a brief familiarization session. System identities were concealed, and presentation order was randomized within each case. All FOA outputs used the same fixed MIT KEMAR binaural decoder described in Appendix D.3, with a shared playback gain within each case. Generation trials presented the original request; editing trials additionally provided the original audio and editing instruction. Participants could replay audio and take breaks.

Ratings and reporting. Participants rated generation on overall audio quality, instruction adherence, spatial adherence, and spatial naturalness. They rated editing on overall audio quality, edit success,

and preservation of unedited content. Each dimension used a five-point scale from 1 (very poor) to 5 (excellent). Table 12 reports the mean of valid ratings for each system and dimension. The same participant rated all three outputs for each case under the same request and playback setup, enabling within-case comparison.

Results. AMBIT outperforms both baselines in mean listener ratings across all seven measures. For generation, it achieves scores of 3.11 in overall quality, 3.25 in instruction adherence, 3.37 in spatial adherence, and 3.03 in spatial naturalness. Its largest advantage is in spatial adherence, where it exceeds the strongest baseline by 0.54 points, followed by a 0.29-point gain in instruction adherence. For editing, AMBIT achieves 3.64 in overall quality, 3.57 in edit success, and 3.71 in preservation of unedited content, exceeding Qwen3-Omni + DiT by 0.13, 0.31, and 0.27 points, respectively. These results highlight AMBIT's strengths in following spatial instructions and carrying out requested edits while preserving the surrounding audio content.

D.5 PLANNER VALIDITY

Table 13 separates structural planning accuracy from rendered audio quality. Executability checks schema and compiler compatibility; source count accuracy checks the requested inventory. Qwen uses its first response without repair or retries, whereas AMBIT uses grammar-constrained decoding. All requests in each evaluation panel remain in the denominator, including failed outputs. Neither measure establishes full instruction satisfaction or correct source attribute binding.

Table 13: Planner executability and source count accuracy (%).

Planner	Executable ↑	Source count accuracy ↑
Generation		
AMBIT AR	100.00	100.00
Qwen3.5-0.8B	22.88	22.69
Qwen3.5-2B	20.19	20.38
Qwen3.5-4B	98.08	98.06
Qwen3.5-9B	87.81	87.83
Qwen3.5-27B	100.00	100.00
Editing		
AMBIT AR	100.00	86.10
Qwen3-Omni	97.00	46.50

E TRANSFUSION OPSD

This section specifies the teacher interfaces, loss reductions, shared updates, and supervision variants in Section 3.4. Joint OPSD denotes distillation in both modes; AR OPSD and DiT OPSD replace distillation in the other mode with paired GT supervision.

E.1 TEACHER CONSTRUCTION AND REFERENCE TOKENS

Planning and rendering teachers start from the same supervised model, using target plan memory and target audio reference tokens, respectively. After preparation, both are frozen; the student retains its initial weights. A separate frozen copy provides reference distributions under ordinary conditions.

Planning teacher. The AR teacher inserts complete target plan memory $M(\mathcal{S}^\star)$ between source and student tokens. Source queries read source tokens only; memory queries read source and memory tokens; student queries additionally read their causal prefix. Source latents, source AMBIT CLAP features, instructions, and ordinary rotary positions are preserved. Preparation combines ordinary and target memory token supervision with paired rendering:

$$\mathcal{L}^{\mathrm{AR}}_{\mathrm{prep}} = a\mathcal{L}_{\mathrm{ord}} + b\mathcal{L}_{\mathrm{mem}} + c\mathcal{L}^{\star}_{\mathrm{RF}}. \tag{26}$$

The first two terms supervise GT tokens on GT prefixes; the rendering term uses GT target plans. Each token loss combines equal weights on the mean token CE and the mean of present field means:

source count, kind, text, activity, motion, and geometry. During distillation, teacher and student instead receive the same student-generated prefix under their respective input conditions.

Rendering teacher. Clean target latents supply temporally aligned reference tokens:

$$H_y[\tau] = W_{\mathrm{ref}} z_y[\tau] + r_{\mathrm{ref}}, \qquad z_y = E_V(x_y), \tag{27}$$

where W_{ref} projects 64 channels to width 1,024 and r_{ref} identifies the reference role. All valid frames are inserted after native prepended tokens, with rotary times aligned to state frames. Reference queries read reference tokens only; ordinary queries can additionally read them. Reference outputs are removed before native post-processing. The preparation objective is

$$\begin{aligned} \mathcal{L}^{\mathrm{DiT}}_{\mathrm{prep}} &= \mathbb{E}\Big[\big\| v_{\phi_{\mathrm{DiT}}}(\zeta_t, t; \mathcal{S}^\star, z_s, H_y) - (\epsilon - z_y)\big\|^2_m\Big], \\ \zeta_t &= (1-t) z_y + t\epsilon. \end{aligned} \tag{28}$$

This updates native teacher parameters and the reference projection and role embedding. Student refinement instead queries actual student states using the generated plan and unchanged source conditioning; the target reference tokens remain exclusive to the rendering teacher.

Preparation configuration. Table 14 specifies preparation budgets of 1,000 AR updates and 3,000 DiT updates, excluding the shared supervised initialization. Both use AdamW with a global batch of 128 requests from the editing training partition. The pretrained text backbones, AMBIT CLAP, and VAE remain frozen during teacher preparation and subsequent student refinement.

Table 14: Teacher preparation budgets and settings. Batches are global.

Teacher	Update budget	Batch	Learning rates	Loss weights
AR	1,000	128	5×10^{-6} / 3.75×10^{-7}	$(a, b, c) = (0.1, 0.3, 0.5)$
DiT	3,000	128	3×10^{-6} / 10^{-4}	RF: 1

AR learning rates apply to adapters and structured heads, then shared blocks and rendering modules; DiT rates apply to native parameters, then the reference projection and role embedding. DiT uses AdamW betas $(0.9, 0.999)$, weight decay 10^{-3}, and gradient clipping at norm 1. AR uses its terminal checkpoint. DiT evaluates checkpoints every 250 updates on 500 fixed validation requests and selects the checkpoint with the lowest paired RF error under target audio conditioning.

E.2 Planning targets and reference constraints

Each request generates one grammar constrained greedy plan under ordinary source audio and instruction conditioning. For planning distillation, distributions use temperature 1 on the same legal support, excluding singleton decisions. Partition non-forced decisions into present families $\mathcal{F}$ with indices $\mathcal{I}_f$ and detached eligibility b_i:

$$\eta_i = \frac{b_i}{|\mathcal{F}|\,|\mathcal{I}_{f(i)}|}. \tag{29}$$

Families distinguish count, content, activity, motion, geometry, room, and other decisions. Room uses ordinary reference supervision. For unambiguously aligned facts $a_i^\star$, exclude sites where the fact is illegal or $\max_a q_i(a) - q_i(a_i^\star) > \delta_{\mathrm{fact}}$, with $\delta_{\mathrm{fact}} = 0.05$. Labels select sites without altering teacher probabilities. A frozen supervised copy supplies ordinary condition distributions q_i^{ref} on the same student prefixes for room, event kind, text, and activity:

$$\ell_{\mathrm{ref}} = \frac{1}{|\mathcal{F}_{\mathrm{ref}}|} \sum_{f \in \mathcal{F}_{\mathrm{ref}}} \frac{1}{|\mathcal{I}_f|} \sum_{i \in \mathcal{I}_f} b_i^{\mathrm{ref}} D_{\mathrm{KL}}(q_i^{\mathrm{ref}} \| p_{\theta,i}). \tag{30}$$

Here $\mathcal{F}_{\mathrm{ref}}$ contains the present reference families. Reference eligibility uses the same checks; rejected sites retain the original denominators. Source indices alone do not establish identity.

E.3 Student states and rendering targets

Joint OPSD and DiT OPSD use the same rendering distillation procedure. The AR student generates a plan $\widehat{\mathcal{S}}$, whose compiled conditions guide ordinary DiT sampling alongside source latents z_s. Each

request uses two noise trajectories, 20 Euler steps, and guidance 1.0. Calls occur at $t_j = 1 - j/20$, $j = 0, \ldots, 19$. Uniformly querying one call from each of four equal strata per trajectory gives eight states with $\omega_j = 1/8$.

At identical states, source latents, and student plans, let v_j^S, v_j^T denote student velocities and target-audio-conditioned teacher velocities:

$$\begin{aligned} \widehat{z}_{0,j}^S &= z_j - t_j v_j^S, \qquad \widehat{z}_{0,j}^T = z_j - t_j v_j^T, \\ \|\widehat{z}_{0,j}^S - \mathrm{sg}(\widehat{z}_{0,j}^T)\|_m^2 &= t_j^2 \|v_j^S - \mathrm{sg}(v_j^T)\|_m^2. \end{aligned} \tag{31}$$

Equation 9 thus matches clean latent predictions, averaged over valid coordinates and sampling calls. The target is the teacher prediction at the collected state; target audio supplies its separate reference tokens. Sampled states, plans, and teacher predictions are detached, while student predictions and trainable conditions are recomputed with gradients under ordinary student conditioning.

E.4 Shared gradients and supervision variants

Joint OPSD. All three losses in Equation 10 reach shared parameters; task specific modules receive gradients along their active paths. Joint OPSD adds no paired GT token or flow loss during student refinement and does not add the structured objective $\mathcal{L}_{\mathrm{struct}}$. GT plans provide teacher memory and eligibility facts, while GT audio provides teacher reference tokens. Teacher preparation uses the paired supervision specified in Appendix E.1.

Branch supervision variants. The two ablations retain the shared student and replace one distillation term with its GT counterpart:

$$\begin{aligned} \mathcal{L}_{\mathrm{AR\ OPSD}} &= \mathbb{E}_{\mathcal{D}}\left[\lambda_{\mathrm{AR}}\ell_{\mathrm{AR}} + \lambda_{\mathrm{DiT}}\ell_{\mathrm{RF}}^{\mathrm{GT}} + \lambda_{\mathrm{ref}}\ell_{\mathrm{ref}}\right], \\ \mathcal{L}_{\mathrm{DiT\ OPSD}} &= \mathbb{E}_{\mathcal{D}}\left[\lambda_{\mathrm{AR}}\ell_{\mathrm{CE}}^{\mathrm{GT}} + \lambda_{\mathrm{DiT}}\ell_{\mathrm{DiT}} + \lambda_{\mathrm{ref}}\ell_{\mathrm{ref}}\right]. \end{aligned} \tag{32}$$

AR OPSD retains planning distillation on student prefixes and replaces rendering distillation with paired flow matching. The term $\ell_{\mathrm{RF}}^{\mathrm{GT}}$ uses the compiled GT target plan $\mathcal{S}^\star$, source latent z_s, and states $\zeta_t = (1-t)z_y + t\epsilon$, with target velocity $\epsilon - z_y$ as in Equation 6. It does not use student sampling states as paired flow targets or require the audio reference teacher.

DiT OPSD replaces planning distillation with ordinary next-token cross-entropy $\ell_{\mathrm{CE}}^{\mathrm{GT}}$, averaged over valid GT tokens with GT prefixes and ordinary editing conditions. Its rendering branch is identical to joint OPSD: ordinary student plans condition the trajectories, and the same prepared teacher receives independent GT audio reference tokens at the queried student states. GT prefixes are used only for the supervised AR loss; DiT rollouts use the student's generated plan, and the AR target memory teacher is unused.

Both variants retain the reference model, field selection, eligibility rules, and λ_{ref} of joint OPSD; neither adds $\mathcal{L}_{\mathrm{struct}}$. Both modes remain trainable: their losses update the shared Transformer and respective interfaces before one optimizer step. These are supervision ablations within a coupled model, where each loss can affect both modes through shared parameters across updates.

Optimization. The three OPSD configurations use the same student initialization, native parameter groups, and global batch of 64 requests, with $\lambda_{\mathrm{AR}} = 0.1$, $\lambda_{\mathrm{DiT}} = 1$, and $\lambda_{\mathrm{ref}} = 0.1$. Learning rates are 5×10^{-6} for AR adapters and structured heads and approximately 3.75×10^{-7} for shared and rendering parameters. Shared weights occur once in the optimizer. Request losses accumulate before global averaging, gradient clipping at 1, and one AdamW step. Unavailable rollout losses contribute zero without changing the batch denominator; GT terms remain defined from the paired record. Each update uses fresh student rollouts. Teachers, the reference model, pretrained text backbones, AMBIT CLAP, and AMBIT VAE remain frozen.

E.5 Post-training controls and comparison protocol

Table 6 compares six post-training configurations: continued SFT, ReFL, GRPO, AR OPSD, DiT OPSD, and joint OPSD. All retain the native shared and task specific student parameter groups; pretrained text backbones, AMBIT CLAP, and AMBIT VAE are frozen. The OPSD variants are defined in Appendix E.4. Below, $\mathcal{L}_{\mathrm{paired}}$ denotes the native editing objective in Equation 7, including its structured supervision.

Comparison protocol. Evaluation uses identical test requests, noise seeds, and sampling settings. Paired batch sizes, reward rollouts, trajectory reuse, and teacher preparation differ as specified for each recipe. These comparisons evaluate complete refinement recipes.

SFT. SFT continues $\mathcal{L}_{\text{paired}}$ with 192 paired examples per update and no reward objective.

ReFL. Our adaptation adds a semantic-similarity reward computed by frozen AMBIT CLAP between the generated terminal latent $\hat{z}_0$ and the raw edit instruction e:

$$\mathcal{L}_{\text{ReFL}} = \mathcal{L}_{\text{paired}} - \lambda_{\text{ReFL}}\, \mathbb{E}[\cos(f_{\text{FOA}}(\widehat{z}_0), f_{\text{text}}(e))]\,, \tag{33}$$

with $\lambda_{\text{ReFL}} = 1$, 144 paired examples, and three reward requests per update. Plans are sampled without gradients; reward backpropagation traverses the final four of 20 Euler steps with guidance 1.0. It reaches rendering and shared parameters, while paired supervision also updates AR. A detached sampled plan therefore does not imply a frozen planner function.

Reward scope. ReFL and GRPO use the semantic alignment head, whose positives include spatial edits that preserve content (Section 3.3). This reward measures semantic compatibility with the instruction; it does not explicitly evaluate the requested direction, timing, or preservation of unedited sources. A relative edit instruction also need not describe the complete target scene. The controls retain paired supervision for spatial and preservation objectives, but their reward does not provide the target plan and audio information available to OPSD teachers.

GRPO. The two-branch control uses categorical policy optimization (Shao et al., 2024) and Flow-GRPO transitions (Liu et al., 2025b). Each request samples $K = 6$ candidate plans and trajectories. Valid candidates receive the same semantic reward as ReFL; undecodable plans receive -1 and no DiT trajectory. For reward r_j, the detached advantage is

$$A_j = \frac{r_j - \bar{r}}{\sqrt{K^{-1}\sum_{k=1}^{K}(r_k - \bar{r})^2 + \varepsilon}}, \qquad \bar{r} = \frac{1}{K}\sum_{k=1}^{K} r_k, \tag{34}$$

where $\varepsilon = 10^{-6}$. AR uses legal-token sampling at temperature 0.8; DiT uses 20 Gaussian SDE transitions with exploration scale 0.3, starting from noise at $t = 1$.

For branch $b \in \{\text{AR}, \text{DiT}\}$, define $\rho^b_{ju} = \exp(s^b_{\theta,ju} - \text{sg}(s^b_{\text{old},ju}))$. AR scores are categorical log probabilities; DiT scores are Gaussian transition log densities averaged over valid latent coordinates. Thus the DiT ratio is a dimension-normalized surrogate, rather than the likelihood ratio of the full latent vector. The clipped objective is

$$\mathcal{L}^b_{\text{clip}} = -\frac{1}{K}\sum_{j=1}^{K} \text{avg}_{u\in\mathcal{U}^b_j} \min\big(\rho^b_{ju}A_j, \text{clip}(\rho^b_{ju}, 1-\delta, 1+\delta)A_j\big)\,, \tag{35}$$

with $\delta = 0.2$. Each candidate uniformly queries up to eight non-forced AR decisions and two DiT transitions; an empty DiT set contributes zero. Each behavior group is reused for two updates.

GRPO retains paired supervision and frozen reference penalties:

$$\mathcal{L}_{\text{GRPO}} = \mathcal{L}_{\text{paired}} + \sum_{b\in\{\text{AR},\text{DiT}\}} \big[\lambda_{\text{G},b}\mathcal{L}^b_{\text{clip}} + \lambda_{\text{KL},b}\mathcal{K}_b\big]\,. \tag{36}$$

Here $\mathcal{K}_{\text{AR}}$ averages $D_{\text{KL}}(p_\theta \| p_{\text{ref}})$ on legal support; $\mathcal{K}_{\text{DiT}}$ averages equal variance Gaussian KL, $\|\mu_\theta - \mu_{\text{ref}}\|^2_m/(2\sigma^2)$. Both branches use $\lambda_{\text{G},b} = 1$ and $\lambda_{\text{KL},b} = 0.02$, alongside 144 paired examples per update. Action scores are recomputed with gradients; sampled actions, states, rewards, advantages, and reference predictions are detached.