AURA: AN ANY-FORMAT UNIFIED REPRESENTATION FOR SPATIAL AUDIO

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial audio is produced and consumed in a zoo of mutually incompatible formats—channel-based layouts such as 5.1, 7.1 and 7.1.4, and scene-based Ambisonics (FOA, HOA). Existing systems either hard-wire a single input/output pair (e.g. stereo-to-7.1.4 upmixing) or route everything through a *fixed*, order-limited Ambisonics basis whose angular resolution is capped a priori. We introduce AURA (an *Any-format Unified Representation for spatial Audio*), which maps *any* spatial-audio format into a single *learnable*, format-agnostic soundfield latent $\mathbf{z}_{\text{field}}$ and decodes it back into *any* target format. The latent is kept physically grounded by two inductive biases: *geometry conditioning*, which tags every channel with its physical direction via spherical-harmonic features and treats it as an element of an unordered set, making the encoder permutation-invariant and shared across formats of differing channel counts; and an *Ambisonics physics prior*, which regularizes $\mathbf{z}_{\text{field}}$ toward the deterministic HOA encoding of the input so the latent adapts to content without collapsing into an opaque channel code. Two capabilities follow: *any-to-any format conversion*, by encoding an input and decoding to the desired layout; and *stereo-to-soundfield generation*, where a flow-matching model generates $\mathbf{z}_{\text{field}}$ from stereo and renders it to arbitrary formats with mutually consistent outputs. On real corpora (STARSS22/23, Freesound 5.1, 3D-MARCo), the frozen front-end preserves spatial structure almost losslessly (sub-degree to 1.5° round-trip DOA error), and on the real 5.1 test set AURA reproduces the reference's inter-channel energy structure far more faithfully than a deterministic upmix (44.3 dB dynamic range vs. 91.0 dB, reference 57.9 dB).

1 INTRODUCTION

Immersive and spatial audio has become the default delivery target for music, film, gaming, and extended reality, yet the field is fragmented across incompatible representations. *Channel-based* formats prescribe a fixed loudspeaker layout (stereo, 5.1, 7.1, or the 12-channel 7.1.4 of Dolby Atmos–style delivery); *scene-based* formats instead encode the soundfield into Ambisonics coefficients—first-order (FOA, 4 channels) and higher-order (HOA, $(N{+}1)^2$ channels). Content authored in one representation routinely must be played back or re-mixed in another, and audio natively available in the richest formats (7.1.4, high-order HOA) is tiny next to the vast stereo catalogue. Two problems therefore recur: *format conversion* (render spatial content to a different layout) and *spatial upmixing* (synthesize an immersive mix from stereo).

Classical solutions use *fixed linear* machinery: Ambisonics gives a principled, rotation-equivariant basis, and any two loudspeaker formats bridge by encoding to HOA and decoding back with fixed matrices—but the order is chosen in advance, so the intermediate's angular resolution is *capped* regardless of content. On the generative side, learned upmixers such as IMMERSIVEFLOW (Liang et al., 2026) produce convincing stereo→7.1.4 upmixes via flow matching in a music-VAE latent, but hard-wire a *single* output layout: a new format means retraining with a new channel count.

We argue that both problems are symptoms of the same missing ingredient: a *single representation of the soundfield that is (a) learnable and content-adaptive, yet (b) format-agnostic*, so that it can be produced from any format and rendered to any format. We propose AURA, which learns exactly such a representation—a latent $\mathbf{z}_{\text{field}}$—and organizes both format conversion and stereo upmixing around it (Figure 1).

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.

The design question is how to make a *learned* latent format-agnostic without discarding the physics that makes spatial audio well-posed. AURA uses two inductive biases. First, **geometry conditioning**: rather than indexing channels by file position, we tag every channel with its true physical direction as real spherical-harmonic (SH) features in the ACN/SN3D convention of Ambisonics, and treat the channel axis as an *unordered set*. The encoder is thus permutation-invariant and the *same weights* process a 6-channel 5.1 and a 12-channel 7.1.4 buffer—reasoning about *where* energy arrives, not *which index* carries it. Second, an **Ambisonics physics prior**: an auxiliary head predicts the deterministic HOA encoding of the input from $\mathbf{z}_{\text{field}}$, regularizing the latent toward classical soundfield structure. Ambisonics is retained as a *prior and baseline*, not a hard bottleneck—$\mathbf{z}_{\text{field}}$ may exceed a fixed order when content warrants, but is kept from collapsing into an opaque channel-compression code.

Because a single decoder is conditioned entirely on the *target* format's directions, one trained AURA model emits 5.1, 7.1, 7.1.4, FOA, or HOA on demand, yielding any-to-any conversion directly. Attaching a flow-matching generator that samples $\mathbf{z}_{\text{field}}$ from stereo further gives a stereo upmixer whose renders to different formats are mutually consistent by construction (they decode the same latent). AURA thus subsumes fixed-target upmixing (Liang et al., 2026) as decoding to one layout.

Contributions.

- **A learnable, format-agnostic soundfield latent $\mathbf{z}_{\text{field}}$**, into which arbitrary channel- and scene-based formats encode and from which arbitrary formats decode, replacing fixed Ambisonics matrices with a content-adaptive map.
- **Two physics-grounded inductive biases**—geometry conditioning (per-channel SH direction features with set-invariant attention) and an Ambisonics auxiliary prior—that keep the latent physically meaningful and format-agnostic while trainable across heterogeneous, differently-supervised data.
- **A unified system** for both any-to-any format conversion and stereo→soundfield generation renderable to any format, generalizing prior fixed-target upmixing.
- **An objective evaluation protocol**—VAE front-end upper bound, energy preservation, spectral fidelity, downmix consistency—on real corpora (Freesound 5.1, 3D-MARCo paired arrays, FOA) against a fixed-HOA baseline.

2 Related Work

Spatial audio representations and Ambisonics. Ambisonics represents a soundfield as a truncated spherical-harmonic expansion, giving a compact, loudspeaker-independent, rotation-equivariant scene description (Gerzon, 1973; Daniel, 2000; Zotter & Frank, 2019); the underlying theory—spherical array processing (Rafaely, 2015), interior-field reconstruction (Ward & Abhayapala, 2001; Betlehem & Abhayapala, 2005), decoders for irregular layouts (Zotter & Frank, 2012)—is mature. Bridging layouts by encoding to Ambisonics and decoding back with (regularized) pseudo-inverse decoders is standard (Poletti, 2005; Zotter & Frank, 2019), as is amplitude panning (Pulkki, 1997). These pipelines are fixed and linear: the expansion order caps angular resolution independent of content, and object/scene standards like MPEG-H 3D Audio (Herre et al., 2015; Coleman et al., 2018) likewise assume a fixed rendering chain. AURA keeps this machinery as an explicit physics baseline and prior but replaces the fixed intermediate with a learned latent.

Spatial upmixing and format conversion. Upmixing stereo or mono to surround has a long signal-processing history—primary–ambient decomposition (Goodwin & Jot, 2007; Avendano & Jot, 2002), directional audio coding and spatial impulse-response rendering (Pulkki, 2007; Merimaa & Pulkki, 2005), parametric coherence coding (Faller, 2006), matrix upmixers (Irwan & Aarts, 2002)—but relies on hand-designed decompositions and fixed target layouts. Recent work casts spatial synthesis as *generative* modelling in a neural latent: ImmersiveFlow (Liang et al., 2026), our closest prior and direct comparison, generates 7.1.4 from stereo with flow matching over a music VAE latent but fixes output to a single 12-channel layout; others target binaural rendering from mono or video (Richard et al., 2021; Gebru et al., 2021; Leng et al., 2022) and mono-to-Ambisonics from 360° video (Morgado et al., 2018). AURA instead learns a *single format-agnostic* latent supporting any-to-any conversion *and* generation.

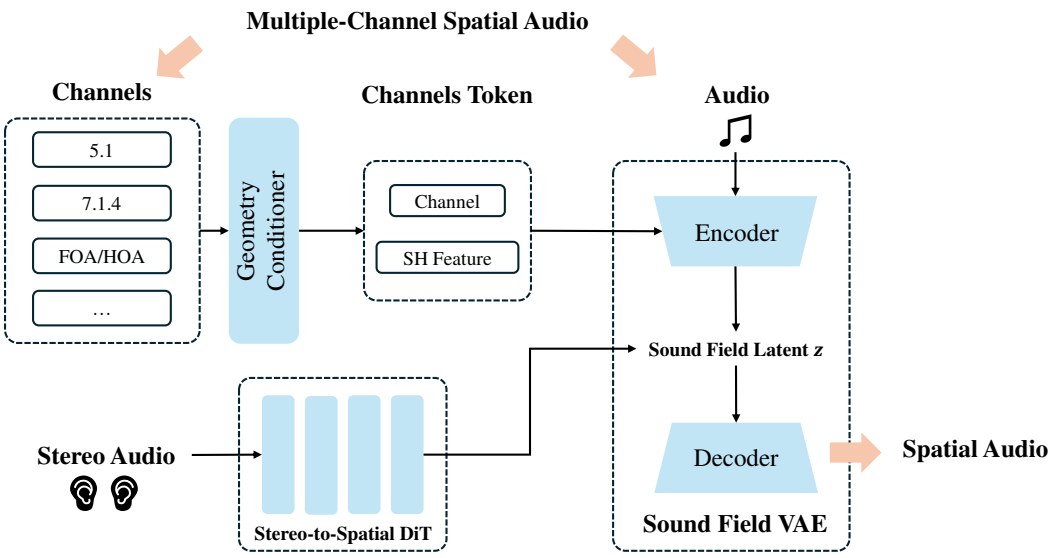

Figure 1: Overview of AURA. **(a)** Any spatial format is encoded into the format-agnostic latent $\mathbf{z}_{\text{field}}$ by a geometry-conditioned, permutation-invariant set encoder (channels carry spherical-harmonic direction features), and decoded back to any target format specified by its channel directions; an auxiliary HOA head supervises $\mathbf{z}_{\text{field}}$ with an Ambisonics physics prior. **(b)** A flow-matching DiT generates $\hat{\mathbf{z}}_{\text{field}}$ from a stereo VAE latent, which the frozen decoder from (a) renders to 7.1.4, 5.1, FOA, or HOA—all renders decode the same latent and are mutually consistent by construction.

Neural audio codecs and latent generative models. Modern audio generation operates in the latent of a neural autoencoder (Kingma & Welling, 2014)—a discrete codec (Zeghidour et al., 2022; Défossez et al., 2022; Kumar et al., 2023) or a continuous VAE (Evans et al., 2024)—underpinning recent language-model and diffusion approaches to audio generation (Dhariwal et al., 2020; Borsos et al., 2023; Agostinelli et al., 2023; Copet et al., 2023). Following latent diffusion (Rombach et al., 2022), we adopt a frozen music VAE (a stable-audio–style latent, as in IMMERSIVEFLOW) as a per-channel front-end, so AURA models *spatial* structure across channels while delegating waveform fidelity to the codec.

Flow matching and diffusion for audio. Diffusion and score-based models (Ho et al., 2020; Song et al., 2021; Karras et al., 2022) and, recently, flow matching along rectified or conditional-OT paths (Lipman et al., 2023; Liu et al., 2023; Albergo & Vanden-Eijnden, 2023; Tong et al., 2024a;b) give simple, simulation-free objectives applied widely to audio and speech (Kong et al., 2021; Le et al., 2023; Vyas et al., 2023). We use a conditional-OT flow matcher with a DiT backbone (Peebles & Xie, 2023; Dosovitskiy et al., 2021; Su et al., 2021; Zhang & Sennrich, 2019), reusing the IMMERSIVEFLOW (Liang et al., 2026) backbone but retargeting generation from a fixed channel tensor to the learned $\mathbf{z}_{\text{field}}$.

Permutation-invariant and geometry-aware set models. Treating inputs as unordered sets with permutation-invariant pooling or attention (Zaheer et al., 2017; Qi et al., 2017; Lee et al., 2019) lets one model handle variable-size inputs, and learnable query slots provide a fixed-size summary of such sets (Locatello et al., 2020). AURA applies these ideas to the *channel* axis of spatial audio, conditioning each set element on its physical direction via SH features so a single encoder/decoder generalizes across formats with different channel counts.

3 LEARNING A FORMAT-AGNOSTIC SOUNDFIELD REPRESENTATION

This section develops AURA around a single object, the latent $\mathbf{z}_{\text{field}}$: we fix notation and the latent domain (Sections 3.1–3.2), give the two components that make the representation format-agnostic yet physically grounded—geometry conditioning (Section 3.3) and the shared set-based

encoder/decoder (Sections 3.4–3.5)—then state the Ambisonics prior (Section 3.6), the training objective (Section 3.7), and how a flow-matching sampler turns the representation into a stereo upmixer (Section 3.8). A deterministic counterpart (Section 3.9) doubles as a data-prep tool and principal baseline.

3.1 Problem setup and notation

A spatial-audio signal in format f is a buffer $x^{(f)} \in \mathbb{R}^{C_f \times L}$ of C_f channels and L samples. Each format f has a known geometry: a loudspeaker layout assigns every channel c a direction (θ_c, ϕ_c) (azimuth, elevation), with an indicator $\ell_c \in \{0, 1\}$ marking non-directional low-frequency-effects (LFE) channels; Ambisonic formats are specified by their order N and the ACN/SN3D convention. Our goal is a single latent $\mathbf{z}_{\text{field}}$ such that *any* format encodes into it and it decodes into *any* format.

We operate entirely in the latent domain of a frozen music VAE (Section 3.2): each channel's waveform is independently mapped to a d-dimensional latent sequence, so a format-f clip becomes $X^{(f)} \in \mathbb{R}^{C_f \times d \times T}$ with T latent frames. All AURA modules act on these per-channel latents.

3.2 Frozen per-channel acoustic front-end

We use a pretrained stable-audio–style stereo music VAE ("LevoVAE") with latent dimension $d{=}64$ at $\approx$21.5 latent frames/s and a 44.1 kHz sample rate, following ImmersiveFlow (Liang et al., 2026). Each channel is encoded independently (a mono channel is duplicated to stereo and averaged back on decode). The VAE is *frozen* throughout: it supplies waveform fidelity while AURA learns only the cross-channel *spatial* structure, keeping the spatial model compact and cheap to train.

3.3 Geometry conditioning via spherical-harmonic direction features

The key to a format-agnostic model is to describe channels by *direction*, not by index. For a channel at (θ_c, ϕ_c) we compute real spherical-harmonic features up to order N_{sh} in the ACN/SN3D convention,

$$\mathbf{y}_c = \left[\, Y_l^m(\theta_c, \phi_c) \,\right]_{0 \le l \le N_{\text{sh}},\, -l \le m \le l} \in \mathbb{R}^{(N_{\text{sh}}+1)^2}, \tag{1}$$

the same basis used by classical Ambisonics encoding. A small MLP maps $\mathbf{y}_c$ to a direction embedding $\mathbf{e}_c \in \mathbb{R}^D$; a dedicated learnable token replaces $\mathbf{e}_c$ for LFE channels ($\ell_c{=}1$). Because the features are a smooth, rotation-friendly function of direction rather than a channel index, they transfer across formats and place channels of different layouts in a common geometric frame.

This is more than a re-parameterization. The SH basis is *shared with the physics*—the functions a first-principles Ambisonic encoder uses—so the geometry the model sees matches what the HOA prior (Section 3.6) supervises; and it is *continuous* in direction, so the model interpolates to layouts and elevations unseen in training, letting a decoder trained on standard layouts emit an arbitrary target geometry. The LFE token keeps non-directional low-frequency energy out of the directional field. We use the AmbiX convention (ACN/SN3D) throughout, so features, prior, and baseline share one coordinate language.

3.4 Format-agnostic encoder: any format $\to \mathbf{z}_{\text{field}}$

The encoder treats the C_f channels of a frame as an *unordered set*. We project each per-channel latent to width D, add its direction embedding $\mathbf{e}_c$, and apply a stack of *set-transformer* blocks (multi-head self-attention with no positional code on the channel axis, plus a feed-forward network) (Lee et al., 2019; Vaswani et al., 2017). Time is folded into the batch, so each frame is encoded from its own channel-token set. A set of S learnable query *slots* then pools the channel tokens by cross-attention and is projected to width d_{slot}, giving a fixed-size latent per frame that we flatten to

$$\mathbf{z}_{\text{field}} \in \mathbb{R}^{D_{\text{field}} \times T}, \qquad D_{\text{field}} = S \cdot d_{\text{slot}}. \tag{2}$$

Two properties follow by construction. *Permutation invariance*: with no channel-axis positional encoding and set attention, the output is invariant to channel reordering. *Format agnosticism*: the same weights accept any channel count C_f, because channels enter only through their latents and direction embeddings. We use $S{=}12$ slots of width $d_{\text{slot}}{=}64$ ($D_{\text{field}}{=}768$), $D{=}512$, and $N_{\text{sh}}{=}3$.

The slots act as a truncation order decoupled from angular resolution: rather than committing to $(N{+}1)^2$ coefficients, cross-attention specializes S slots to whatever directional structure the content exhibits, making $\mathbf{z}_{\text{field}}$ *content-adaptive* where a fixed-order basis spends the same coefficients regardless. Retaining the time axis lets the flow-matching model treat $\mathbf{z}_{\text{field}}$ as an ordinary temporal sequence. Cross-attention pooling is essential: mean-pooling is permutation-invariant but discards the directional distribution the slots capture.

3.5 Format-agnostic decoder: $\mathbf{z}_{\text{field}} \to$ any format

The decoder mirrors the encoder and is conditioned entirely on the *target* format's geometry, which is what lets one model emit any layout. The target channels' direction embeddings act as queries in a cross-attention that *reads* the $\mathbf{z}_{\text{field}}$ slots; a few self-attention blocks over the channel queries refine them, and a linear head projects each back to a per-channel VAE latent $\hat{X}^{(f)} \in \mathbb{R}^{C_f \times d \times T}$. Because the output channel set is defined solely by the directions supplied at decode time, a single trained decoder produces 5.1, 7.1, 7.1.4, FOA, or HOA without modification.

The encoder–decoder asymmetry is deliberate: the encoder *summarizes* a variable-size channel set into fixed slots; the decoder *expands* them against a variable-size set of target-direction queries, so the output channel count is simply the number of queries. Adding a height layer or a denser array means supplying more queries, with no change to weights or latent—a flexibility fixed upmixers lack (each new layout needs a new decoding matrix or retrained head, as in ImmersiveFlow). For Ambisonic targets we read the HOA prior head directly (Section 3.6).

3.6 Ambisonics physics prior

To keep $\mathbf{z}_{\text{field}}$ physically grounded we attach an auxiliary head that predicts, per frame, the deterministic HOA-coefficient latents of the input (order N_{hoa}, i.e. $(N_{\text{hoa}}{+}1)^2$ Ambisonic channels) from $\mathbf{z}_{\text{field}}$ with a $1{\times}1$ convolution over time. This head serves both to regularize $\mathbf{z}_{\text{field}}$ during training and, at inference, to render Ambisonic output formats directly (truncating/zero-padding by ACN order). It ties $\mathbf{z}_{\text{field}}$ to the classical soundfield structure without capping it at a fixed order.

The prior resolves a tension any learned soundfield latent faces. Without supervision toward a physical target, an encoder–decoder trained only to reconstruct channels can discover an *opaque* code that reconstructs well but bears no relation to where sound comes from—one that would not transfer across formats and would give the generator (Section 3.8) nothing physically meaningful to sample. Requiring $\mathbf{z}_{\text{field}}$ to also linearly predict the deterministic HOA encoding of the input anchors it to that soundfield. Crucially the prior is a *soft* regularizer (weight λ_h), not a hard bottleneck: the latent may carry detail beyond order N_{hoa} provided its low-order projection stays consistent with the physics. AURA thus treats Ambisonics as a prior belief about the soundfield, not the representation itself.

3.7 Training objective for the autoencoder

Let $E(\cdot, f)$ and $D(\cdot, f)$ be the encoder and decoder for format f, and $H(\cdot)$ the HOA head. Given a source clip $X^{(a)}$ in format a, an optional paired target $X^{(b)}$ in format b (when a common source exists, e.g. a deterministic downmix or a same-scene multi-array recording), and an optional HOA target B, we minimize

$$\mathcal{L}_{\text{recon}} = \left\| D(E(X^{(a)}, a), a) - X^{(a)} \right\|^2, \tag{3}$$

$$\mathcal{L}_{\text{cross}} = \left\| D(E(X^{(a)}, a), b) - X^{(b)} \right\|^2, \tag{4}$$

$$\mathcal{L}_{\text{latent}} = \left\| E\big(D(E(X^{(a)}, a), b), b\big) - E(X^{(a)}, a) \right\|^2, \tag{5}$$

$$\mathcal{L}_{\text{hoa}} = \left\| H(E(X^{(a)}, a)) - B \right\|^2, \tag{6}$$

with the total loss

$$\mathcal{L}_{\text{AE}} = \lambda_r \mathcal{L}_{\text{recon}} + \lambda_c \mathcal{L}_{\text{cross}} + \lambda_\ell \mathcal{L}_{\text{latent}} + \lambda_h \mathcal{L}_{\text{hoa}}. \tag{7}$$

$\mathcal{L}_{\text{recon}}$ is a self-reconstruction; $\mathcal{L}_{\text{cross}}$ supervises cross-format decoding when a paired target is available; $\mathcal{L}_{\text{latent}}$ is a cycle-consistency term requiring that re-encoding a cross-decoded output recovers

the same $\mathbf{z}_{\text{field}}$; and $\mathcal{L}_{\text{hoa}}$ is the physics prior. Terms whose inputs are absent for a given example (no pair, no HOA target) are simply skipped, so the same objective trains across heterogeneous datasets with differing supervision. We use $(\lambda_r, \lambda_c, \lambda_\ell, \lambda_h)=(1, 1, 0.1, 0.1)$. Together, geometry conditioning (in the architecture) and $\mathcal{L}_{\text{hoa}}$ (in the loss) prevent $\mathbf{z}_{\text{field}}$ from degenerating into an opaque channel code.

3.8 STEREO→SOUNDFIELD GENERATION

To generate immersive audio from stereo, we learn to sample $\mathbf{z}_{\text{field}}$ conditioned on a stereo input, then render it to any format with the frozen decoder. Given a trained, frozen AURA autoencoder, the generation target for a clip is $\mathbf{z}_{\text{field}} = E(X^{(\text{mc})}, \text{mc})$, the latent of its multi-channel ground truth (e.g. 7.1.4). We train a conditional flow-matching model with a DiT-style transformer backbone (the IMMERSIVEFLOW backbone: width 1024, 12 layers, 16 heads, RoPE, adaptive-modulation conditioning) (Liang et al., 2026; Peebles & Xie, 2023; Su et al., 2021). With a conditional-OT (rectified) path (Lipman et al., 2023; Tong et al., 2024a) and σ=0, we draw $t \sim \mathcal{U}[0, 1]$, $\mathbf{x}_0 \sim \mathcal{N}(0, I)$, form the straight-line interpolant $\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\,\mathbf{z}_{\text{field}}$ with target velocity $\mathbf{z}_{\text{field}} - \mathbf{x}_0$, and train the network $v_{\boldsymbol{\theta}}$ with the data-space objective

$$\mathcal{L}_{\text{FM}} = \left\| \big(\mathbf{x}_t + (1-t)\, v_{\boldsymbol{\theta}}(\mathbf{x}_t, t, \mathbf{c})\big) - \mathbf{z}_{\text{field}} \right\|^2, \tag{8}$$

where the conditioning $\mathbf{c}$ is the frozen stereo VAE latent, projected and added per frame into the DiT modulation. At inference we encode the stereo input with the VAE, integrate the flow ODE from Gaussian noise to a generated $\hat{\mathbf{z}_{\text{field}}}$, and decode it to every requested format. Because all outputs decode the *same* $\hat{\mathbf{z}_{\text{field}}}$, the 7.1.4, 5.1, and Ambisonic renders are mutually consistent by construction.

This decomposition—generate once, render many times—separates AURA from a fixed-target upmixer: *consistency for free* (the 5.1 and 7.1.4 outputs are two views of one latent) and *generation decoupled from delivery* (the ODE is integrated once; a new format costs only a decode). We reuse the IMMERSIVEFLOW backbone and recipe unchanged, so any difference on the 7.1.4 render is attributable to the generation target—learned $\mathbf{z}_{\text{field}}$ versus a fixed 12-channel tensor. Freezing the autoencoder gives a fixed target distribution and avoids encoder–generator collusion on a latent easy to generate but poor to render.

3.9 DETERMINISTIC PHYSICS BASELINE

For reference and for data preparation, we implement the classical route: a fixed SH encoding matrix maps any loudspeaker format to HOA, a regularized pseudo-inverse decodes HOA to any format, and format-to-format conversion routes through HOA at a default order N=3. This is the fixed, non-learned counterpart of AURA and serves as the **HOA baseline** in all conversion experiments.

4 EXPERIMENTAL EVALUATION

Our evaluation is designed to answer three questions in turn. *What is the representation's ceiling?*—we first isolate the frozen VAE front-end and measure its round-trip fidelity, since every reference is a VAE decoding (Section 4.2). *Does the learned representation help?*—we then evaluate AURA on any-to-any conversion and on stereo→soundfield generation, the two capabilities the representation is meant to unify (Sections 4.3–4.4). *Which design choices matter?*—we finally ablate the two inductive biases and the capacity of $\mathbf{z}_{\text{field}}$ (Section 4.5). Throughout, the fixed-HOA pipeline of Section 3.9 and IMMERSIVEFLOW (Liang et al., 2026) serve as the non-learned and fixed-target references, respectively. Generation is evaluated on the one format with a real multi-channel test set (5.1).

4.1 EXPERIMENTAL SETUP

Datasets. We use real spatial-audio corpora spanning both scene- and channel-based formats: **STARSS22/23** (Politis et al., 2022; Shimada et al., 2023) (real FOA scenes), **Freesound 5.1** (real consumer-produced 5.1, our only real multi-channel test set for generation), and **3D-MARCo** (Lee & Johnson, 2021) (concert-hall recordings in multiple channel-based 3D arrays and HOA), with

Table 1: Datasets and splits used in our experiments. #Clips are counted per split (train/test); for 3D-MARCo a "clip" is a scene captured simultaneously by four 9-channel arrays plus a 12-channel array. Hours are totals across splits. From each source we encode all available formats plus a deterministic stereo downmix (used as the generation condition); STARSS FOA additionally yields a deterministic 7.1.4 render and an HOA target.

Dataset	Native format(s)	#Clips (tr/te)	Hours	Use
STARSS22	FOA ($\rightarrow$7.1.4, HOA)	173/1	6.8	conv., eval
STARSS23	FOA ($\rightarrow$7.1.4, HOA)	184/63	10.9	conv., eval
Freesound 5.1	5.1 (real, consumer)	418/48	9.2	gen., eval
3D-MARCo	9-ch $\times 4$ arrays, 12-ch, HOA	11/2	0.3	conv. (paired), eval

first-order material from EigenScape and MRSAudio for pretraining (Green & Murphy, 2017; Huang et al., 2025; Adavanne et al., 2018). Paired cross-format supervision comes from same-scene multi-format recordings—most directly the 3D-MARCo arrays, which capture one performance simultaneously in several layouts—and, where a format is missing, from a deterministic downmix of a richer format. All audio is resampled to the VAE's native 44.1 kHz and encoded at ≈ 21.5 frames/s; clips are ≈ 10 s ($T \approx 215$). Splits and counts are in Table 1.

Metrics. We report three format-agnostic objective metrics, all computed by first encoding any output to HOA so numbers are comparable across formats—essential when no native format is shared by all outputs: (i) **DOA error** (degrees), the great-circle angle between the estimated source direction (Ambisonic active-intensity vector) and ground truth; (ii) **energy preservation** (dB), the output/reference total-energy ratio (0 dB ideal); and (iii) **downmix consistency** for generation—a *reference-free* SNR/correlation/energy match between a generated output's stereo downmix and the conditioning stereo, exposing generators that invent unrelated spatial content. Where reference channels exist we add *spectral* fidelity (STFT/mel-magnitude SNR, energy-aggregated) rather than waveform SNR, since the perceptual VAE does not reconstruct waveforms sample-for-sample (Section 4.2); for generation we also report Fréchet Audio Distance (Kilgour et al., 2019) and the reference-free PAM. These are necessary but not sufficient; a listening-test protocol is in the appendix.

Baselines. For conversion we compare against the **fixed HOA** pipeline (order-1 and -3 intermediates); for generation, against the deterministic upmix baseline (Section 3.9) on the real 5.1 test set. The comparison against IMMERSIVEFLOW (Liang et al., 2026) (fixed 7.1.4 target) is deferred until a *real* 7.1.4 corpus is available, as our 7.1.4 material is a deterministic render of FOA (Section 4.4).

Implementation. The autoencoder uses $D{=}512$, 4 encoder/decoder blocks, 8 heads, $S{=}12$ slots of width 64 ($D_{\text{field}}{=}768$), $N_{\text{sh}}{=}N_{\text{hoa}}{=}3$; the generator uses the DiT backbone of Section 3.8. Both train with AdamW (Kingma & Ba, 2015; Loshchilov & Hutter, 2019), lr 10^{-4}, 1000-step warmup, batch 16, EMA 0.999, up to $4{\times}10^5$ steps. Full details in the appendix.

4.2 REPRESENTATION FIDELITY: THE FRONT-END UPPER BOUND

Every reference in our evaluation is the VAE decoding of a real latent, so the frozen per-channel front-end sets an *upper bound* on end-to-end fidelity: no soundfield latent can recover detail the VAE has discarded. We measure it in isolation by round-tripping real multi-channel audio (encode$\rightarrow$decode, per channel) on the real 5.1 corpus and the 3D-MARCo 9-channel arrays (48 clips each), reporting spectral fidelity (STFT- and mel-magnitude SNR) and, crucially, *spatial* fidelity: the round-trip DOA shift and the change in the inter-channel correlation matrix (ICC-ℓ_1). Table 2 shows the front-end preserves spatial structure almost losslessly—DOA drifts by only 0.3° (5.1) to 1.5° (9-ch) and inter-channel correlation is essentially unchanged—while giving up moderate spectral detail (7–9 dB magnitude SNR), localizing the method's headroom in timbre rather than spatial placement and motivating latent-domain losses and perceptual metrics over waveform SNR.

Table 2: VAE front-end round-trip fidelity (representation upper bound), per channel-format. Spectral/mel SNR are magnitude-domain and energy-aggregated across channels (↑); energy preservation is $|\Delta\text{dB}|$ (↓, 0 ideal); DOA drift is the round-trip direction shift (↓); ICC-ℓ_1 is the mean absolute change in the inter-channel correlation matrix (↓).

Format	Spec. SNR (dB)	Mel SNR (dB)	$\lvert\Delta\text{E}\rvert$ (dB)	DOA drift (deg)	ICC-ℓ_1
5.1 (Freesound, real)	9.1	9.4	0.48	0.33	0.141
3D-MARCo 9-ch (real)	8.6	7.2	1.03	1.50	0.010

Table 3: Any-to-any conversion on real 3D-MARCo paired recordings (13 scenes). Latent-SNR (↑, dB) is against the true target latent; $|\Delta\text{E}|$ (↓, 0 ideal) is energy preservation after VAE decoding. DOA is omitted pending absolute-direction calibration (text). The cross-array row (hardest) converts between two arrays of the same scene.

Conversion (real target)	AURA lat-SNR (dB,↑)	AURA $\lvert\Delta\text{E}\rvert$ (dB,↓)	Fixed-HOA $\lvert\Delta\text{E}\rvert$ (dB,↓)
9-ch → 9-ch (self-recon)	**11.3**	1.62	0.01
9-ch → Hamasaki 12-ch	2.8	**1.17**	3.62
9-ch → HOA (N=3)	4.3	0.54	0.16
Hamasaki 12-ch → 9-ch	2.8	**3.59**	3.87
9-ch 2LCube → OCT3D (cross-array)	2.3	**1.56**	3.16

4.3 Any-to-any format conversion

Table 3 reports conversion on *real* paired data: 3D-MARCo captures each scene simultaneously in four 9-channel arrays, a 12-channel Hamasaki layout, and HOA, so a genuine target recording exists for each source→target pair (13 scenes). We evaluate two calibration-free axes: *latent-domain SNR* against the true target latent, and *energy preservation* after VAE decoding. We omit DOA here (Appendix A.6): the arrays and our approximate angles are not yet on a common absolute-direction frame. The pairs probe three regimes: *self-reconstruction* (9-ch→same); *cross-layout* to channel-based and Ambisonic targets (9-ch↔Hamasaki, 9-ch→HOA); and a *cross-array* case (9-ch 2LCube→true OCT3D)—the same scene via *different* arrays, testing whether $\mathbf{z}_{\text{field}}$ abstracts away pickup (the hardest). Fixed-HOA energy follows a different rendering path and is a reference point only.

Three findings stand out. *(i)* Self-reconstruction gives the highest latent-SNR (11.3 dB); fixed-HOA is energy-exact here (0.01 dB) as an identity map—a floor, not a competitor. *(ii)* On genuinely cross-format targets (Hamasaki, cross-array OCT3D) AURA preserves energy substantially better than fixed HOA (1.17 vs. 3.62 dB; 1.56 vs. 3.16 dB), because the learned decoder adapts to the target layout instead of forcing every signal through an order-limited bottleneck. *(iii)* Cross-format latent-SNR is modest (2–4 dB): reaching a *different* array's exact latent is intrinsically hard—recordings differ in pickup, not just playback geometry—so latent-SNR lower-bounds fidelity while energy shows perceptually relevant structure is retained.

4.4 Stereo→soundfield generation

Table 4 reports generation on the 5.1 render, the one format for which we hold a *real* (non-synthesized) test set (48 consumer clips; our 7.1.4 material is a deterministic FOA render, not honest ground truth). We compare AURA against the deterministic upmix (Section 3.9) on distribution realism (Fréchet Audio Distance (Kilgour et al., 2019) over CLAP, and reference-free PAM) and a *per-channel loudness* analysis, all on VAE decodings so the frozen VAE's distortion cancels. The picture is mixed. On distributional realism the upmix—which copies the true stereo verbatim into the front channels—attains lower FAD, with PAM within noise: a learned generator pays a fidelity cost a content-copying baseline does not. On *spatial energy structure*, however, AURA is markedly

Table 4: Stereo→soundfield generation on the real 5.1 test set (48 clips, 50 ODE steps). FAD (↓) / PAM (↑) in per-channel and mono-downmix aggregations; the loudness block compares each method's inter-channel energy to the reference. Best per row in **bold**.

Metric	AURA (ours)	Det. upmix	Reference
FAD, per-channel (↓)	0.306	**0.103**	—
FAD, downmix (↓)	0.290	**0.180**	—
PAM, per-channel (↑)	0.495	**0.501**	0.499
PAM, downmix (↑)	0.502	**0.509**	0.506
Per-channel loudness (5.1)			
Dynamic range (dB, →ref)	**44.3**	91.0	57.9
Mean $\lvert\Delta$dB$\rvert$ vs. ref (↓)	**17.4**	22.5	—
LFE vs. ref (dB, → 0)	$+16.8$	-34.4	—
Downmix consistency SNR (dB, ↑)	1.40	—	—

Table 5: Ablations on the autoencoder, trained 40k steps each (shortened schedule) and evaluated on 3D-MARCo pairs. Recon. Spec. SNR is 9-ch self-reconstruction (↑); cross columns are 9-ch→Hamasaki latent-SNR (↑) and $|\Delta\mathrm{E}|$ (↓). Best per column **bold**.

Variant	Recon Spec. SNR (dB)	Cross latent-SNR (dB)	Cross $\lvert\Delta$E$\rvert$ (dB)
Full AURA (S=12, N_{sh}=3)	**11.3**	1.66	1.75
− geometry cond. (index emb.)	11.1	1.60	**1.64**
− HOA prior (λ_h=0)	9.4	1.68	1.68
− cycle cons. (λ_ℓ=0)	11.2	1.66	1.76
S=8 slots	6.7	**1.70**	1.84
S=16 slots	9.7	1.66	1.74
N_{sh}=2	11.5	1.68	1.80

closer to the reference: its dynamic range (44.3 dB) tracks ref (57.9 dB) where the baseline inflates to 91.0 dB, and it populates the LFE where the deterministic route omits it.

4.5 Ablations

Table 5 isolates each inductive bias at a shortened 40k-step budget where the effects are already visible on reconstruction. The clearest signal is *latent capacity*: S=8 slots collapse reconstruction (spectral SNR 11.3 → 6.7 dB) while S=16 trades fidelity (9.7 dB) for no cross-format gain, so S=12 sits near the knee. Removing the HOA prior (λ_h=0) costs ∼2 dB, confirming it helps rather than merely constrains, whereas N_{sh}=2 and dropping cycle consistency (λ_ℓ=0) leave reconstruction unchanged. A per-index embedding barely dents it (11.1 dB): on a single array its cost is conceptual (forfeiting format agnosticism), not visible in reconstruction. The cross-format columns stay nearly flat—at 40k steps all variants sit near the difficulty floor of Section 4.3.

5 Conclusion

We presented AURA, a learnable, format-agnostic soundfield latent $\mathbf{z}_{\mathrm{field}}$ that encodes and decodes arbitrary spatial-audio formats—grounded by geometry conditioning and an Ambisonics prior—and supports any-to-any conversion and stereo→soundfield generation. It preserves spatial structure almost losslessly and reproduces reference inter-channel energy far better than a deterministic upmix; closing the residual spectral gap and extending to a real 7.1.4 corpus are the natural next steps.

Ethics statement

This work uses only publicly available spatial-audio corpora (STARSS22/23, Freesound 5.1, 3D-MARCo), each employed in accordance with its license; the Freesound material was filtered by

the quality-control procedure of Appendix A.5 to remove mislabeled or degenerate uploads. The research involves no human subjects, no personally identifiable information, and no user studies beyond the optional listening protocol of Appendix A.7, which would be conducted with informed consent and no collection of identifying data. We foresee no direct path to harmful application: the method converts and generates spatial-audio layouts and does not synthesize speech or identity. We declare no conflicts of interest, and we have complied with the ICLR Code of Ethics.

Reproducibility statement

The deterministic physics baseline, geometry/SH utilities, encoder/decoder, training objectives, and evaluation metrics are all implemented with self-contained smoke tests that run on synthetic data without downloads or a GPU. Appendix A.1 fixes the coordinate and Ambisonics conventions, Appendix A.2 the loudspeaker layouts, Appendix A.3 the front-end and network architecture, Appendix A.4 the full training objective and schedule with all loss weights, Appendix A.5 the dataset preparation and quality-control steps, and Appendix A.6 the exact metric definitions. Anonymized source code, configuration files, and checkpoints are included as supplementary material.

Statement on the use of AI

Large language models were used only as a general-purpose writing and coding assistant—polishing prose, formatting tables and figures, and refactoring utility code—under direct author supervision. They were not used to generate research ideas, design or run experiments, produce or analyze results, or write the technical claims of the paper. The authors take full responsibility for all content, including any text refined with such tools.

References

Sharath Adavanne, Archontis Politis, Joonas Nikunen, and Tuomas Virtanen. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 13(1):34–48, 2018.

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. MusicLM: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.

Michael S. Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *International Conference on Learning Representations (ICLR)*, 2023.

Carlos Avendano and Jean-Marc Jot. Frequency domain techniques for stereo to multichannel upmix. In *AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio*, 2002.

Terence Betlehem and Thushara D. Abhayapala. Theory and design of sound field reproduction in reverberant rooms. *The Journal of the Acoustical Society of America*, 117(4):2100–2111, 2005.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. AudioLM: A language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

Philip Coleman, Andreas Franck, Philip J. B. Jackson, Richard J. Hughes, Luca Remaggi, and Frank Melchior. An audio-visual system for object-based audio: From recording to listening. *IEEE Transactions on Multimedia*, 20(8):1919–1931, 2018.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Jérôme Daniel. *Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia*. PhD thesis, University of Paris VI, 2000.

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

Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Zach Evans, C. J. Carr, Josiah Taylor, Scott H. Hawley, and Jordi Pons. Fast timing-conditioned latent audio diffusion. In *International Conference on Machine Learning (ICML)*, 2024.

Christof Faller. Parametric multichannel audio coding: Synthesis of coherence cues. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):299–310, 2006.

Israel D. Gebru, Dejan Marković, Alexander Richard, Steven Krenn, Gladstone A. Butler, Fernando de la Torre, and Yaser Sheikh. Implicit HRTF modeling using temporal convolutional networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

Michael A. Gerzon. Periphony: With-height sound reproduction. *Journal of the Audio Engineering Society*, 21(1):2–10, 1973.

Michael M. Goodwin and Jean-Marc Jot. Primary-ambient signal decomposition and vector-based localization for spatial audio coding and enhancement. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2007.

Marc Ciufo Green and Damian Murphy. EigenScape: A database of spatial acoustic scene recordings. *Applied Sciences*, 7(11):1204, 2017.

Jürgen Herre, Johannes Hilpert, Achim Kuntz, and Jan Plogsties. MPEG-H 3d audio—the new standard for coding of immersive spatial audio. *IEEE Journal of Selected Topics in Signal Processing*, 9(5):770–779, 2015.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Rongjie Huang et al. MRSAudio: A large-scale multimodal recorded spatial audio dataset. In *arXiv preprint*, 2025.

Roy Irwan and Ronald M. Aarts. Two-to-five channel sound processing. *Journal of the Audio Engineering Society*, 50(11):914–926, 2002.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Interspeech*, 2019.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. DiffWave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations (ICLR)*, 2021.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved RVQGAN. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, et al. Voicebox: Text-guided multilingual universal speech generation at scale. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Hyunkook Lee and Dale Johnson. A dataset of higher-order ambisonics and multichannel microphone-array recordings for 3d audio (3D-MARCo). In *International Conference on Immersive and 3D Audio (I3DA)*, 2021.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning (ICML)*, 2019.

Yichong Leng, Zehua Chen, Junliang Guo, Haohe Liu, Jiawei Chen, Xu Tan, Danilo Mandic, Lei He, Xiangyang Li, Tao Qin, Sheng Zhao, and Tie-Yan Liu. BinauralGrad: A two-stage conditional diffusion probabilistic model for binaural audio synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Zining Liang, Runbang Wang, Yujia Xiao, Xuzhou Ye, and Qiuqiang Kong. ImmersiveFlow: Stereo-to-7.1.4 spatial audio generation with flow matching, 2026. `https://immersiveflow.github.io`.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Juha Merimaa and Ville Pulkki. Spatial impulse response rendering i: Analysis and synthesis. *Journal of the Audio Engineering Society*, 53(12):1115–1127, 2005.

Pedro Morgado, Nuno Vasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360 video. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

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

Mark A. Poletti. Three-dimensional surround sound systems based on spherical harmonics. *Journal of the Audio Engineering Society*, 53(11):1004–1025, 2005.

Archontis Politis, Kazuki Shimada, Parthasaarathy Sudarsanam, Sharath Adavanne, Daniel Krause, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, Yuki Mitsufuji, and Tuomas Virtanen. STARSS22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events. In *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2022.

Ville Pulkki. Virtual sound source positioning using vector base amplitude panning. *Journal of the Audio Engineering Society*, 45(6):456–466, 1997.

Ville Pulkki. Spatial sound reproduction with directional audio coding. *Journal of the Audio Engineering Society*, 55(6):503–516, 2007.

Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. PointNet: Deep learning on point sets for 3d classification and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Boaz Rafaely. *Fundamentals of Spherical Array Processing*. Springer, 2015.

Alexander Richard, Dejan Markovic, Israel D. Gebru, Steven Krenn, Gladstone A. Butler, Fernando de la Torre, and Yaser Sheikh. Neural synthesis of binaural speech from mono audio. In *International Conference on Learning Representations (ICLR)*, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

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. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research (TMLR)*, 2024a.

Alexander Tong, Nikolay Malkin, Kilian Fatras, Lazar Atanackovic, Yanlei Zhang, Guillaume Huguet, Guy Wolf, and Yoshua Bengio. Simulation-free Schrödinger bridges via score and flow matching. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024b.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Yossi Adi, Wei-Ning Hsu, et al. Audiobox: Unified audio generation with natural language prompts. *arXiv preprint arXiv:2312.15821*, 2023.

Darren B. Ward and Thushara D. Abhayapala. Reproduction of a plane-wave sound field using an array of loudspeakers. *IEEE Transactions on Speech and Audio Processing*, 9(6):697–707, 2001.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R. Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. SoundStream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2022.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Franz Zotter and Matthias Frank. All-round ambisonic panning and decoding. *Journal of the Audio Engineering Society*, 60(10):807–820, 2012.

Franz Zotter and Matthias Frank. *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*. Springer, 2019.

Table 6: Per-channel directions (azimuth°, elevation°) for each format, in file channel order. LFE channels (†) are non-directional.

Format	Channels: name (az°, el°)
stereo	L(30,0) R(−30,0)
5.1	L(30,0) R(−30,0) C(0,0) LFE† Ls(110,0) Rs(−110,0)
7.1	L(30,0) R(−30,0) C(0,0) LFE† Lss(90,0) Rss(−90,0) Lrs(135,0) Rrs(−135,0)
7.1.4	L(30,0) R(−30,0) C(0,0) LFE† Lss(90,0) Rss(−90,0) Lrs(135,0) Rrs(−135,0) Ltf(45,45) Rtf(−45,45) Ltb(135,45) Rtb(−135,45)
3D-MARCo 9-ch	FL(30,0) FR(−30,0) FC(0,0) RL(110,0) RR(−110,0) FLh(30,45) FRh(−30,45) RLh(110,45) RRh(−110,45)
Hamasaki 12-ch	FL(45,0) FR(−45,0) RL(135,0) RR(−135,0) FLh(45,30) FRh(−45,30) RLh(135,30) RRh(−135,30) FLh(45,60) FRh(−45,60) RLh(135,60) RRh(−135,60)

A Appendix

A.1 Coordinate and Ambisonics conventions

All directions use azimuth θ (counter-clockwise from front $+x$ when viewed from above) and elevation ϕ (up from the horizontal plane), in the Cartesian frame x front, y left, z up, so a unit direction is $(\cos\phi\cos\theta, \cos\phi\sin\theta, \sin\phi)$. Ambisonics follows the **AmbiX** convention: **ACN** channel ordering ($\mathrm{acn} = l(l{+}1) + m$) with **SN3D** normalization $N_l^m = \sqrt{(2{-}\delta_{m0})\,(l{-}|m|)!/(l{+}|m|)!}$. First-order channels are therefore ordered W, Y, Z, X (not the naive W, X, Y, Z). Real spherical harmonics are evaluated with the associated Legendre functions $P_l^{|m|}$ and the Condon–Shortley phase removed, so that the SN3D FOA identities $W{=}1,\ X{=}\cos\phi\cos\theta,\ Y{=}\cos\phi\sin\theta,\ Z{=}\sin\phi$ hold. The same basis is used for (i) the geometry-conditioning SH features (Section 3.3), (ii) the HOA prior head (Section 3.6), and (iii) the deterministic baseline (Section 3.9), so features, prior, and baseline share one coordinate language.

A.2 Loudspeaker layouts

Table 6 lists the per-channel directions of every format, in file channel order. Channel-based formats use standard angles; the 3D-MARCo playback layouts are approximate and treated as tunable hyperparameters (the four 9-channel arrays—2LCube, DeccaCuboid, OCT3D, PCMA3D—share one playback geometry, so $\mathcal{L}_{\mathrm{cross}}$ between them learns to abstract away pickup rather than layout). The Hamasaki 12-channel array (index-labelled "7.1.4" in the source release, which is incorrect: it has no centre or LFE) is modelled as four main speakers plus two height sub-layers at $+30°$ and $+60°$ elevation. LFE channels carry no direction and are routed through a dedicated learnable token.

A.3 Front-end and architecture details

The frozen front-end is a stable-audio–style stereo music VAE ("LevoVAE") with latent dimension $d{=}64$, downsampling ratio 2048 at 44.1 kHz (hence $\approx$21.5 latent frames/s), following ImmersiveFlow. Each channel is encoded independently; a mono channel is duplicated to stereo before encoding and averaged back on decode. Clips of $\approx$10 s give $T{\approx}215$ latent frames.

The autoencoder (Sections 3.4–3.5) uses model width $D{=}512$, 4 encoder and 4 decoder set-transformer blocks with 8 attention heads, $S{=}12$ query slots of width $d_{\mathrm{slot}}{=}64$ ($D_{\mathrm{field}}{=}768$), and SH/HOA orders $N_{\mathrm{sh}}{=}N_{\mathrm{hoa}}{=}3$. Direction embeddings come from a small MLP over the $(N_{\mathrm{sh}}{+}1)^2{=}16$-dim SH features; the HOA head is a $1{\times}1$ temporal convolution to the $(N_{\mathrm{hoa}}{+}1)^2{=}16$ Ambisonic latent channels. The generator (Section 3.8) is the ImmersiveFlow DiT backbone: input dimension 768, width 1024, 12 layers, 16 heads, RoPE with length 8192, and adaptive-modulation conditioning on the projected stereo VAE latent.

A.4 Full training objective and schedule

Beyond the four terms of Eq. (7), the final autoencoder adds a *perceptual loudness-alignment* regularizer $\mathcal{L}_{\text{energy}}$. A latent-domain mean-squared error correlates only weakly with perceived per-channel loudness (Pearson ≈ 0.33 on held-out data), which flattens the inter-channel energy distribution and under-/over-populates the LFE. We therefore pretrain a frozen *loudness read-out* head (a light MLP mapping a channel's latent to its log-RMS level in dB; held-out Pearson ≈ 0.97) and add a dB-domain MSE between the read-out of each reconstructed/cross-decoded channel and that of its target. The full loss is $\mathcal{L}_{\text{AE}} + \lambda_e \mathcal{L}_{\text{energy}}$ with $(\lambda_r, \lambda_c, \lambda_\ell, \lambda_h, \lambda_e) = (1, 1, 0.1, 0.1, 0.5)$.

Both the autoencoder and the generator train with AdamW (Kingma & Ba, 2015; Loshchilov & Hutter, 2019) at learning rate 10^{-4}, a 1000-step linear warm-up, EMA decay 0.999, and up to 2×10^5 steps, on a single 24 GB GPU. The generator is trained against the frozen autoencoder's encodings of real multi-channel ground truth for the target formats $\{7.1.4,\ 5.1,\ 7.1\}$; batches are format-homogeneous (Appendix A.5). Reported numbers use the EMA weights.

A.5 Data preparation and quality control

Latents are precomputed per channel and stored to HDF5 keyed by $\{\text{group}\}_\{\text{format}\}$, where *group* is a sanitized relative path so that the multiple formats of one recording (or the multiple 3D-MARCo arrays of one scene) share a group and never leak across the train/test split. Because the collate step stacks per-channel tensors, each batch must be *channel-count-homogeneous*; a format-bucketed batch sampler groups samples by (src, tgt) format so every batch is homogeneous while shuffling within and across buckets.

Paired cross-format supervision for $\mathcal{L}_{\text{cross}}$ comes from 3D-MARCo, which records each scene simultaneously with five arrays; we form ordered format pairs within a scene (including same-format cross-array pairs such as 2LCube↔OCT3D). Freesound 5.1/7.1 clips pass a heuristic quality filter (exactly six/eight channels, no all-silent or mono-duplicated channels, an LFE with a plausible low-frequency energy fraction, minimum duration) that only flags—never deletes—source files; 7.1 channels are permuted from the WAV order to our layout order. Where a format is missing, a deterministic downmix of a richer format supplies the target, and a deterministic stereo downmix is always produced as the generation condition.

A.6 Metric definitions and the DOA caveat

All objective metrics are computed after mapping each output into a common HOA frame (so numbers are comparable across formats with different channel counts), on the VAE decodings of method, baseline, and reference alike so the frozen VAE's distortion cancels. **Spectral SNR** is a magnitude-domain SNR of the STFT (and mel) spectrogram, energy-aggregated across channels; we prefer it to waveform SNR because the perceptual VAE does not reconstruct phase sample-for-sample (Section 4.2), so waveform SNR would be dominated by inaudible phase differences and near-silent channels. **Energy preservation** $|\Delta\text{E}|$ is the absolute total-energy ratio in dB (0 ideal). **Downmix consistency** is the SNR/correlation/energy match between a generated output's stereo downmix and its conditioning stereo, and is reference-free.

We omit **DOA error** from the conversion and ablation tables. Our 3D-MARCo playback angles are approximate (Appendix A.2) and the arrays are not yet registered to a common absolute-direction frame, so a degree-valued cross-array DOA error would report calibration mismatch rather than model quality. DOA *is* reported for the front-end upper bound (Table 2), where round-trip direction drift is measured against a signal in its own frame and is therefore well-defined. Absolute-direction calibration of the paired arrays is left to future work.

A.7 Listening-test protocol

Objective metrics are treated as necessary but not sufficient. We outline a MUSHRA-style subjective protocol for a future evaluation: for each stereo clip, listeners rate the multi-channel renders of AURA, the deterministic upmix, a hidden reference (where a real target exists), and an anchor (mono or low-order downmix), over a calibrated multi-channel monitoring setup. Attributes are overall quality, spatial fidelity (localization and envelopment), and timbral fidelity. The same pro-

tocol applies to any-to-any conversion by presenting the converted render against the true target format. Anonymized source code, including the deterministic baseline, geometry/SH utilities, encoder/decoder, objectives, and evaluation metrics with self-contained synthetic-data smoke tests, is provided as supplementary material.