AUDIOSPHERE: Towards Self-Supervised Spatial Audio Representation Models

Goksenin Yuksel
Donders Institute, Radboud University
Nijmegen, the Netherlands
goksenin.yuksel@donders.ru.nl

Marcel van Gerven
Donders Institute, Radboud University
Nijmegen, the Netherlands
marcel.vangerven@donders.ru.nl

Kiki van der Heijden
Donders Institute, Radboud University, Nijmegen, the Netherlands
Mortimer B Zuckerman Institute, Columbia University, New York, United States
kiki.vanderheijden@donders.ru.nl

Abstract

Recent audio embedding models learn representations that transfer across clip- and frame-level tasks, yet they operate almost exclusively on single-channel signals and are blind to spatial structure. Spatial audio models capture that structure but rely on clip-level supervision or contrastive objectives, yielding embeddings unsuitable for frame-level tasks like sound event detection and localization (SELD). We argue this dichotomy is not fundamental: patch-level spatial self-supervision can jointly learn acoustic and spatial features at both temporal granularities. We introduce AudioSphere, a multi-channel masked autoencoder trained on Ambisonics recordings that reconstructs both spectral content and directional cues via a joint spatiotemporal masking strategy that masks both time-frequency patches and intensity vectors. AudioSphere supports frame-level predictions while producing strong clip-level representations via aggregation. Alongside the model, we release RealSELD, a standardized benchmark for evaluating frozen backbones on real-world multi-channel SELD. AudioSphere enables, for the first time, SELD with a frozen self-supervised backbone, while matching or exceeding prior self-supervised audio embedding models on clip-level classification tasks spanning environmental sound, speech, and music, and on frame-level sound event detection.[1]

1 Introduction

Despite the complexity and diversity of everyday sound scenes, human listeners effortlessly perceive both *what* they hear and *where* it comes from. Audio embedding models aspire to this breadth [Turian et al., 2022, Yang et al., 2021], and recent self-supervised learning (SSL) approaches achieve strong performance on conventional benchmarks with no fine-tuning [Chen et al., 2023, Baevski et al., 2020, Yadav et al., 2024]. Yet, state-of-the-art models are trained almost exclusively on single-channel corpora such as AudioSet [Gemmeke et al., 2017] and LibriSpeech [Panayotov et al., 2015]. This renders them essentially "spatially blind." This omission is consequential: in humans, spatial cues are central to auditory scene analysis, aiding the separation of overlapping sound waves into distinct auditory objects [Bregman, 1984, van der Heijden et al., 2019]. Consequently, the absence of spatial information in learned embeddings not only precludes sound localization but also limits their ability to parse complex acoustic scenes.

[1] All code and data are available at https://github.com/labhamlet/AudioSphere and https://huggingface.co/labhamlet.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

While SSL has produced embeddings that transfer across a wide range of monaural audio tasks, a fundamental divergence persists in current architectures. Monaural models achieve remarkable performance by scaling to vast unlabeled corpora, but ignore the spatial dimension [Chen et al., 2023, Baevski et al., 2020, Yadav et al., 2024]. Conversely, spatial models like BAT [Zheng et al., 2024] and ELSA [Devnani et al., 2024] capture directional cues but remain limited to clip level supervised labels or contrastive spatial metadata. Labeled spatial supervision has three consequences. It restricts training to annotated real-world scenes, which is scarce, or to simulated scenes with known geometry. Unless such models are trained at very large scale or the resulting representations are optimized for a particular task or dataset and transfer poorly. And their clip-level objectives are not designed to learn the frame-resolved representations required for tracking moving sources.

We address this limitation by pre-training an audio representation model on multi-channel audio recordings using spatial-acoustic patch-level self-supervision. In this work, we adopt the Ambisonics spatial audio format [Zotter and Frank, 2019] for learning spatial audio embeddings. Ambisonics provides a compact, geometry-independent encoding of the spatial sound field, decoupling the representation from the recording hardware [Zotter and Frank, 2019]. Moreover, any multi-channel recording can be projected into this canonical format via Ambisonics Signal Matching (ASM) [McCormack et al., 2022], a signal-processing technique that maps signals from an arbitrary microphone array to Ambisonics. This projection preserves the spatial cues of the original sound field while yielding a fixed-dimensional input, independent of the number or arrangement of microphones.

To this end, we introduce AUDIOSPHERE, a multi-channel masked autoencoder that learns joint acoustic and spatial representations. AUDIOSPHERE reconstructs both log-mel spectrograms and active intensity vectors through a tube-masking strategy that jointly masks spectral and spatial patches, forcing the encoder to internalize spatial information alongside the acoustic features required for clip-level classification and frame-level event detection. To train AUDIOSPHERE at scale, we develop a simulation pipeline that generates 70,000 high-fidelity Ambisonic room impulse responses (ARIRs) from Matterport3D [Chang et al., 2017] environments, paired with an in-batch spatial sampling strategy that increases the effective batch size 16-fold while amortizing computational costs.

The main contributions of our work are as follows. (1) AUDIOSPHERE learns joint spatial and acoustic embeddings through self-supervised training, unlocking real-world SELD with a single model. With only shallow MLP probing on a frozen backbone, it outperforms fully supervised baselines and prior spatial models on static scenes and transfers competitively to dynamic environments. The same embeddings also serve as strong acoustic representations, transferring to environmental-sound, speech, and music classification, as well as frame-level sound event detection. (2) To support reproducible evaluation of spatial audio representations, we release REALSELD, the first evaluation kit for real-world SELD, which adapts five established real-world SELD datasets into a frozen evaluation format. (3) By using Ambisonics as a canonical representation, AUDIOSPHERE processes tetrahedral, planar, and spherical microphone arrays without retraining; on real-world recordings from a planar array (RSL2019-dev), the ASM + AUDIOSPHERE pipeline achieves a median direction-of-arrival error of 1.9°, compared with 5.6° for MUSIC and 13.1° for the Binaural Signal Matching (BSM) + Spatial-AST (BAT) pipeline.

2 Related Work

Self-supervised audio representation learning. The dominant paradigm for learning transferable audio representations is masked spectrogram modeling with Transformer backbones. Such methods include SSAST [Gong et al., 2022], MSM-MAE [Niizumi et al., 2022], MaskSpec [Chong et al., 2023], MAE-AST [Baade et al., 2022], and MW-MAE [Yadav et al., 2024], which achieves strong performance on the HEAR benchmark through multi-window local-global attention in the decoder. SSAM [Yadav and Tan, 2024] replaces the Transformer encoder with a Mamba backbone [Gu and Dao, 2023] and attains comparable results. Recently, Dasheng [Dinkel et al., 2024] demonstrated the impact of massive data scaling, achieving state-of-the-art performance on the HEAR benchmark by increasing the pre-training volume from the standard $\sim$5.8k hours to 272k hours. Beyond reconstruction-based objectives, other methods operate on latent representations. BEATs [Chen et al., 2023] masks latent embeddings produced by an acoustic tokenizer, while ATST-Frame and ATST-Clip [Li et al., 2024] utilize a student-teacher self-distillation paradigm to reconstruct frame-level latent embeddings from masked spectrogram views. Contrastive approaches such as COLA [Saeed et al., 2021] dispense with masking entirely. A parallel line of work targets speech specifically; models

such as Wav2Vec2 [Baevski et al., 2020], HuBERT [Hsu et al., 2021], and WavLM [Chen et al., 2021] achieve strong speech performance but transfer less reliably to non-speech tasks [Turian et al., 2022]. Critically, all of the above operate on single-channel audio and learn no spatial representation of the sound scene.

Spatial audio representation learning. While spatial audio has recently gained traction, existing models are typically trained in a supervised manner, relying on synthetic labels from simulated sound scenes. Supervised approaches, such as BAT [Zheng et al., 2024] and ELSA [Devnani et al., 2024], rely on labels or natural-language metadata, thereby limiting their training to fully synthetic datasets. While models optimized exclusively for regressing to synthetic spatial labels [Sarabia et al., 2023] achieve localization, they do not jointly learn the acoustic classes and thus fail to capture the underlying acoustic properties of the audio signal. Furthermore, because these frameworks rely on static, clip-level supervision, their objectives are not designed to learn the frame-resolved representations required for moving-source SELD. To mitigate these supervision and temporal granularity bottlenecks, w2v-SELD [dos Santos et al., 2023] recently explored learning representations from unlabeled Ambisonics data. w2v-SELD operates entirely in the waveform domain using a wav2vec2-based objective strictly tailored for SELD. Consequently, it does not address transfer to tasks beyond SELD, and its generalization capacity is severely bottlenecked by a pre-training corpus limited to 200 hours. Recent audio-LLM frameworks, such as Sci-Phi [Jiang et al., 2025] and Whisper-based variants [Tang et al., 2024], offer natural-language interfaces for static scenes on clip-level tasks, but their audio embeddings have not been shown to transfer to spatial and acoustic downstream tasks. They effectively yield a natural language interface for spatial sound scene understanding rather than a spatial-acoustic representation model.

AUDIOSPHERE differs from the state-of-the-art in three key dimensions: (i) It learns joint spatial-acoustic features via a frame-level reconstruction objective, requiring zero ground-truth labels; (ii) Unlike task-specific encoders, AUDIOSPHERE produces embeddings that excel at both acoustic and spatial downstream tasks through a shallow MLP probe on frozen features; (iii) As a self-contained front-end, it provides a "plug-and-play" spatial-acoustic layer for any downstream architecture, including existing audio-LLM pipelines.

3 Learning Self-Supervised Spatial Audio Representations

3.1 First-order Ambisonics

We assume a far-field acoustic scenario in which the sound field is composed of a distribution of plane waves. The spatio-temporal amplitude of the sound field, denoted as $s(t, \Omega)$, depends on time t and the direction of arrival $\Omega = (\theta, \phi)$, where $\theta \in [0, 2\pi)$ denotes azimuth and $\phi \in [-\pi/2, \pi/2]$ denotes elevation.

Ambisonics relies on the decomposition of the spherical sound field into orthogonal spatial basis functions known as Spherical Harmonics (SH). The sound field can be reconstructed from its Ambisonic coefficients via:

$$s(t, \Omega) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} B_{lm}(t) Y_{lm}(\Omega) \tag{1}$$

where $B_{lm}(t)$ are the Ambisonic coefficients of order l and degree m. In practice, the summation is truncated at a maximum order L, resulting in $K = (L+1)^2$ channels. The basis functions $Y_{lm}(\theta, \phi)$ are Real Spherical Harmonics defined using associated Legendre functions $P_l^{|m|}(\sin \phi)$. Here, we utilize the N3D normalization and ACN channel ordering (see Appendix A for full definitions). Ambisonics, therefore, serves as a canonical representation of the 3D sound field, providing a reusable format for encoding spatial audio from any microphone geometry.

Ambisonics Signal Matching (ASM). To enable AUDIOSPHERE to process diverse microphone arrays, we adopt Ambisonics Signal Matching (ASM) [McCormack et al., 2022, Bastine et al., 2022], which generalizes traditional Ambisonic encoding to arbitrary configurations via Tikhonov regularization [Golub et al., 1999]. While parallel work explores deep neural networks for this mapping [Heikkinen et al., 2026], we opt for the established ASM framework given its established effectiveness. Our pipeline is modular; neural format conversion techniques can be integrated as they advance. We use the `SHroom` [Gayer, 2026] for all ASM computations. ASM is a fixed linear front

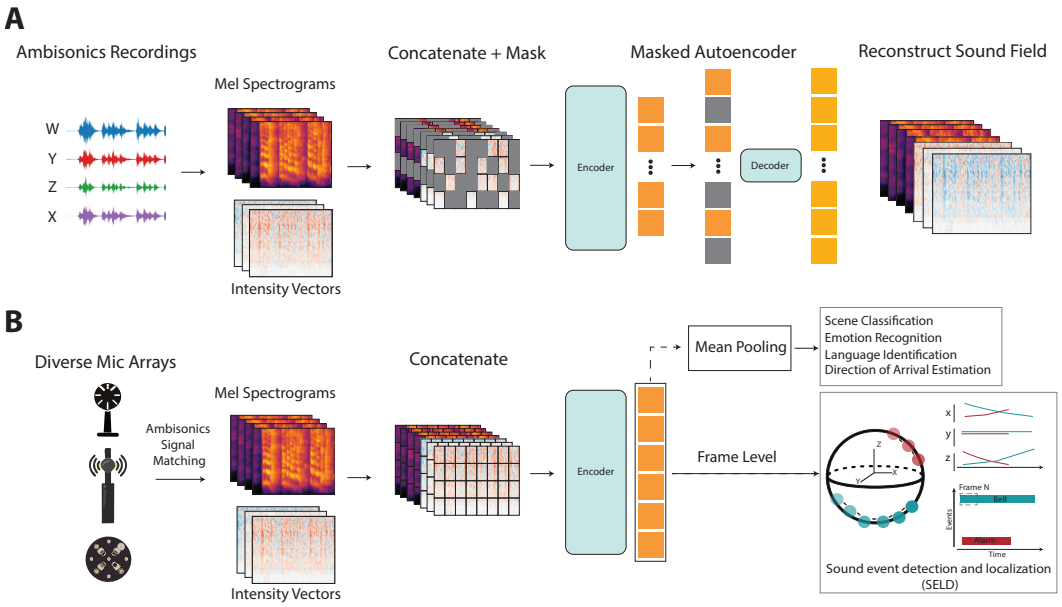

Figure 1: **(A) Pre-Training:** We employ the Masked Autoencoder (MAE) approach to learn spatial audio representations. Ambisonics recordings are converted into mel spectrograms and Intensity Vectors, concatenated, and masked. The Encoder processes unmasked patches and the Decoder reconstructs the original full sound field. **(B) Inference:** The pre-trained Encoder can be flexibly applied to diverse microphone arrays. Diverse input signals are standardized via Ambisonics Signal Matching and passed through the Encoder to extract frame-level and scene-level Ambisonics embeddings.

end with its filters precomputed from the array geometry and contains no trainable parameters, but the mapping is differentiable with respect to its input.

3.2 Dataset Generation

We generate training data in three stages: (i) high-quality simulation of room acoustics via Ambisonic room impulse responses (ARIRs), (ii) stochastic scene composition, and (iii) online mixing with AudioSet sound clips.

Simulating acoustic scenes. Large-scale, natively recorded Ambisonics datasets remain scarce: the largest available corpus contains 246 hours of audio [Kim et al., 2025, Morgado et al., 2020], compared to the standard $\sim$5.8K hours of single-channel sound clips in AudioSet [Gemmeke et al., 2017]. Here, we used the $\sim$4.8 K hours ($\sim$1.74M clips) currently available from AudioSet. To bridge this gap, we leverage high-resolution 3D meshes from Matterport3D [Chang et al., 2017] and the Monte Carlo ray-tracing simulator from SoundSpaces 2.0 [Chen et al., 2022] to generate Ambisonic room impulse responses (ARIRs) that encode both room acoustics and first-order spatial cues in realistic scenarios.

Scene composition. We randomly selected 70 MatterPort3D houses with a sufficient mesh quality. For each house we simulate 1,000 scenes, each comprising a randomly sampled listener position, a sound source (distance $\in [1.5, 5]$ m; azimuth $\in [0, 2\pi)$; elevation $\in [-\pi/2, \pi/2]$, and one or more noise sources. Half of the scenes use localized noise at a single position; the other half use diffuse noise from three to five randomly placed sources. For each configuration we render $\mathrm{ARIR}(s, r, \theta)$ from source to listener and $\mathrm{ARIR}(\phi_i, r, \theta)$ from each noise source to listener, yielding 70,000 ARIR sets (see Appendix K for full parameters). We restrict our pre-training simulations to static sources, as moving sources require sampling hundreds of RIRs per trajectory and thus scale poorly for self-supervised pre-training. We note, however, that because AUDIOSPHERE learns spatial cues at the patch level, it can localize moving sources during inference (empirically validated in Section 5.2).

Online mixing. During training, each clip from the AudioSet training set [Gemmeke et al., 2017] ($\sim$1.74M clips, $\sim$4.8K hours) is randomly paired with a WHAM! noise clip [Maciejewski et al., 2020] and trimmed to 10 s with a 200 ms linear fade-in/fade-out. The target signal T is obtained by

convolving the AudioSet clip with $\mathrm{ARIR}(s, r, \theta)$, and each noise clip produces N_i via convolution with $\mathrm{ARIR}(\phi_i, r, \theta)$. For diffuse scenes the noise field is $D = \sum_{i=1}^{M} N_i$. The final scene is:

$$S = \begin{cases} T + bN & \text{(localized noise)} \\ T + bD & \text{(diffuse noise)} \end{cases} \tag{2}$$

where b sets the signal-to-noise ratio, sampled uniformly from $[+5, +40]$ dB. Although AudioSet clips may already contain reverberation and background noise, we prioritize acoustic diversity following established practice [Zheng et al., 2024, Devnani et al., 2024, Politis, 2022].

3.3 Model Training

Input features. We transformed the channels of each sound scene (i.e., the waveforms) into log-scale mel spectrograms using 128 mel filters in the frequency range of 50-16000 Hz with a 25 ms Hanning window and 10 ms hop length, resulting in spectrograms of dimension 1001×128. We added zero padding to achieve a dimension of 1024×128. For AUDIOSPHERE, following Devnani et al. [2024], Wang et al. [2022], we extracted intensity vectors utilizing the equation below:

$$\boldsymbol{I}_{\text{active}}(t, f) = \Re \left[A^*_{0,0}(t, f) \begin{pmatrix} A_{1,\text{-}1}(t, f) \\ A_{1,0}(t, f) \\ A_{1,1}(t, f) \end{pmatrix} \right], \tag{3}$$

where $A_{n,m}$ are the nth and m^{th} order and mode of the ambisonics signal corresponding to its omnidirectional (W) and three dipole (Y, Z, X) components, and $(\cdot)^*$ denotes complex conjugation. IVs are scaled to unit norm. These two targets address acoustic and spatial aspects of the Ambisonics sound field. The log-mel spectrograms of the W, X, Y, Z channels encode the acoustic content of the scene, while the active intensity vectors, the real part of the cross-spectrum between W and the directional components, integrate the relative phase and magnitude relations between channels into a physically interpretable directional vector. We concatenated mel spectrograms and intensity vectors, resulting in input $x = [x_{\text{mel}}, IVs]$.

Patching and masking. A single convolutional layer partitions each multi-channel spectrogram into N non-overlapping patches $\{P_i\}_{i=1}^{N}$, $P_i \in \mathbb{R}^{7\times8\times16}$, and projects each into a d-dimensional embedding $E_i \in \mathbb{R}^d$ with d=768. Fixed 2D sinusoidal positional embeddings are added. We employ *tube masking*: 80% of time-frequency patch locations are selected uniformly at random, and the corresponding patches are masked across all input channels simultaneously. Masked embeddings are replaced with a shared learnable mask token during decoding.

Tube masking couples spatial and acoustic learning. Because the Ambisonics and intensity-vector channels share one time–frequency grid, masking a location in only some channels would let the model copy it from the unmasked channels at the same location, which impedes learning joint acoustic-spatial representations. Masking all channels at once forces the encoder to infer the directional channels and intensity vectors from the surrounding spatial context. Appendix F (Table 9) compares tube masking with per-channel masking and cross-channel reconstruction targets, and varies masking pattern and ratio. The results suggest that tube masking outperforms per-channel masking on both an acoustic probe and a spatial one, whereas a random masking pattern with a high ratio is beneficial, so we use random tube masking at 80%.

Encoder. The unmasked embeddings are processed by a 12-layer ViT-Base encoder [Dosovitskiy et al., 2021] (hidden dimension = 768, 12 attention heads, MLP dimension = 3072), producing representations $\{O_i\}_{i=1}^{N'}$ where N' is the number of visible patches.

Decoder. The decoder takes as input the unmasked encoder outputs $\{O_i\}$ and the learnable mask tokens $\{M_i\}$, with $M_i \in \mathbb{R}^{768}$, augmented with fixed sinusoidal positional embeddings. The decoder consists of eight Transformer layers with local-global attention [Yadav et al., 2024], using window sizes $[2, 5, 10, 25, 50, 100, 0, 0]$ where 0 denotes global (full) attention. A linear projection head reconstructs the four-channel spectrogram and three IVs from the decoder output ($768 \rightarrow 7 \times 8 \times 16$).

In-batch sampling. As the online mixing of naturalistic acoustic scenes is computationally expensive, we introduced a random in-batch sampling procedure to increase the effective batch size in a computationally efficient manner. We randomly sampled 16 partially overlapping 2-second segments, yielding 16 samples of dimension 200 × 128. This increases the nominal batch size of 96 to an effective batch size of 1536.

Table 1: **The REALSELD datasets**. *Impulse response*: real room impulse responses convolved with isolated sound events (Real) or direct recordings (–).

Name	Motion	Impulse Response	Max Overlap	Noise
TUT-2018 Real (ov1, 2, 3) [Adavanne et al., 2019a]	Static	Real	1, 2, 3	Ambient 30 dB
TAU-2019 [Adavanne et al., 2019b]	Static	Real	2	Ambient 30 dB
TAU-2020 [Politis et al., 2020a]	Dynamic	Real	2	Ambient [30-6 dB]
TAU-2021 [Politis et al., 2021]	Dynamic	Real	3	Dynamic + Ambient
STARSS23 [Shimada et al., 2023]	Dynamic	–	6	Dynamic + Ambient

4 Downstream Evaluation

We aim to answer four questions using the pre-trained, frozen AUDIOSPHERE encoder. (i) *Does spatial pre-training transfer to monaural audio understanding?* We test this with HEAR [Turian et al., 2022], a suite of classification and detection tasks consisting of single-channel audio. (ii) *Do the learned embeddings carry frame-resolved spatial information that supports real-world SELD, in static and in dynamic scenes?* We test this with REALSELD, our evaluation kit of real-world multi-channel data (Section 5.2). (iii) *Does all of this hold across microphone geometries?* We test this on tetrahedral and planar arrays. (iv) *Does the encoder rely on spatial cues to understand spatial scenes?* We test this by removing or corrupting spatial information at pre-training or inference time.

Evaluating the transfer of spatial-acoustic audio representations to monaural tasks. We use the subset of HEAR from Yadav et al. [2024]: *acoustic events and scenes*, i.e., tagging (FSD50K), classifying environmental sounds (ESC-50), and counting speakers (LibriCount); *speech*, i.e., emotion recognition (CREMA-D), spoken-language identification (VoxLingua107), and keyword spotting (Speech Commands); and *music*, i.e., pitch classification (NSynth) and percussion instrument, stroke, and tonic classification (Beijing Opera, Mridangam). Because this subset contains no timestamp-level task, we add the HEAR sound event detection task DCASE-2016 Task 2 [Mesaros et al., 2018]. Appendix G lists all tasks. As a monaural reference we include AUDIOSPHERE-Clean, trained on AudioSet with the MW-MAE framework plus our in-batch sampling, and we ablate the ratio of clean (single-channel) to simulated (four-channel) pre-training data. Single-channel audio is duplicated across channels when mixed with four-channel data, following Devnani et al. [2024].

Evaluating the transfer to SELD tasks in real-world scenes. To test in a standardized manner how well audio representation models perform SELD in real-world scenes, we introduce REALSELD. REALSELD embeds five established real-world SELD datasets (seven evaluation tasks) in the HEAR evaluation format. The datasets differ in difficulty and realism (Table 1). TUT-2018 Real [Adavanne et al., 2019a] and TAU-2019 [Adavanne et al., 2019b] contain *static* scenes with one to three simultaneous sources and ambient noise at 30 dB SNR. TAU-2020 and TAU-2021 introduce *moving* sources and more complex background noise. STARSS23 contains real recordings of natural indoor scenes with up to six simultaneous moving sources and complex background noise. The TUT and TAU datasets are constructed by convolving isolated sound events with real room impulse responses recorded in diverse and challenging environments, including underground bomb shelters, gym spaces, classrooms, glass-walled meeting rooms, and lecture halls, while STARSS23 is recorded directly.

The SELD task is formulated using the ACCDOA framework [Shimada et al., 2021] in which the probe predicts, per frame and per class, a vector whose direction is the direction-of-arrival (DoA) and whose length indicates whether the class is active. We report location-dependent metrics for dynamic sound scenes: Error rate ER_{20° and F-score F_{20°, which count a detection as correct only if the class is right and the predicted direction is within 20° of the reference; class-aware localization: localization error LE_{CD} (angular distance of correct-class predictions) and localization recall LR_{CD} (how often a direction is produced for a reference event); and location-independent metrics for static sources following the DCASE convention (Appendix I). Note that to resolve the mismatch between the 80 ms temporal resolution of the AUDIOSPHERE embeddings and the 100 ms label resolution of the dynamic RealSELD datasets, we interpolated the embeddings to the 100 ms label resolution using average pooling.

Evaluating benefit of pre-training for dataset-specific fine-tuning. Although a pre-trained audio representation model may perform suboptimally on a task data distribution that diverges strongly from the training distribution in comparison with a supervised model fine-tuned extensively on

this particular distribution, the characteristics of the learned representations may nevertheless aid the model's transfer to such a specific task data distribution by minimizing the need for extensive fine-tuning. To assess the transferability of AUDIOSPHERE's embeddings, we compared three training scenarios. *Scratch*: training the AUDIOSPHERE architecture on the target dataset only, instead of pre-training on our training dataset. *Frozen probing*: evaluating the frozen pre-trained model with a trainable downstream head on the target dataset, fitting only the head, which is either the shallow MLP probe used throughout or a lightweight temporal head from the SELDnet baseline. *Fine-tuning*: a two-step approach in which the model is first pre-trained on our training dataset, followed by joint optimization of the model and the downstream head on the target dataset. On STARSS23 [Shimada et al., 2023] (we use the extended version of STARSS23 dataset as the DCASE 2023 Baseline) we compared training AUDIOSPHERE from scratch versus fine-tuning the pretrained AUDIOSPHERE: to train from scratch, we tested four different learning rates $\{10^{-3}, 5{\times}10^{-4}, 2{\times}10^{-4}, 10^{-4}\}$, and for fine-tuning we used a learning rate of 10^{-4}.

Evaluating the generalization across microphone array geometries. We evaluate real-world recordings from various arrays: the tetrahedral arrays of TAU-2019, TAU-2020, TAU-2021 and STARSS23 and the planar ReSpeaker 2.0 array of RSL2019-dev [Sheelvant et al., 2019], all encoded to first-order Ambisonics with ASM. RSL2019-dev is converted to the HEAR format with stratified train/validation/test splits (50/25/25%) by source location; we re-implement MUSIC [Schmidt, 1986] on the held-out split and evaluate the BAT (Spatial-AST) encoder on the same split after mapping the recordings to its binaural input with Binaural Signal Matching (BSM) [Gayer et al., 2025] (Appendix C).

5 Results

5.1 AUDIOSPHERE Transfers to Clip-Level Classification and Frame-Level Detection on HEAR

Table 2 summarizes the HEAR performance of recent audio representation models and AUDIOSPHERE (per-task results in Appendix Table 14). AUDIOSPHERE (α=0.75) reaches an aggregate of 83.2, the highest among the compared models, including Dasheng, which was trained on roughly 56 times more source audio ($\sim$272k hours versus the $\sim$4.8k hours; Appendix Table 15 summarizes the pre-training data and protocol of every compared model). Dasheng's corpus contains substantially more speech, and it keeps a clear lead on speech tasks. Speech-only models (Wav2Vec 2.0, HuBERT, WavLM: 55.7–68.7) and the supervised PaSST and Spatial-AST lag in aggregate, reflecting narrower transfer.

Table 2: Aggregate HEAR score (mean over the 11 tasks of Appendix Table 13; per-task results in Appendix Table 14), $\pm$ SD over three downstream seeds. Grey: supervised. **AS**: AudioSet.

Model	Corpus	HEAR
BEATs	AS	76.6
SSAM	AS	79.6
ATST-Clip	AS	79.8
MW-MAE	AS	80.0
ATST-Frame	AS	81.9
Dasheng	Mix (272k h)	82.8
PaSST	AS	68.1
Spatial-AST	AS	74.6
AudioSphere (α=0.0)	AS	81.1
AudioSphere (α=0.75)	AS	**83.2** $\pm$ 0.1
AudioSphere-Clean	AS	83.0 $\pm$ 0.2

The clean data ratio study effectively reveals that the best transfer to the monaural HEAR tasks emerges from a hybrid training distribution, consistent with Devnani et al. [2024]. This demonstrates that incorporating a proportion of clean, single-channel data during pre-training is critical for learning representations that perform well on single-channel tasks. A hybrid training distribution lifts the HEAR aggregate from 81.1 (pure simulation) to 83.2, on par with the clean-trained model (83.0 ± 0.2 over three probe seeds), while simultaneously equipping the encoder with spatial sensitivity that transfers to real-world SELD (Table 10). Crucially, the spatial pipeline matches the monaural baseline (across three downstream evaluation seeds, 83.2 ± 0.1 for AUDIOSPHERE α=0.75 versus 83.0 ± 0.2 for AUDIOSPHERE-Clean) while additionally enabling frozen-probe SELD, demonstrating that spatial pre-training comes at no cost to transfer on monaural clip-level classification and frame-level event detection.

Table 3: **Static real-world SELD on TUT-2018 REAL**. Sound event detection and direction-of-arrival estimation by number of overlapping sources. We report the mean over three downstream seeds ± the standard deviation. **Bold** = best per column within each block; "–" = not reported. DOAnet's LE covers only its detected frames (LR 46.5/11.5/2.9 vs. 83.2/48.0/26.1 for AUDIOSPHERE), i.e., fewer frames. †: frozen self-supervised backbone; other rows are supervised.

Method	Overlap 1	Overlap 2	Overlap 3
SED	ER↓ / F↑	ER↓ / F↑	ER↓ / F↑
SELDNet [Adavanne et al., 2019a]	0.40 / 60.3	0.49 / 53.1	0.53 / 51.1
MSEDnet [Adavanne et al., 2017]	0.35 / 66.2	**0.38** / 61.6	**0.41** / 59.5
SEDnet [Adavanne et al., 2017]	0.38 / 64.6	0.42 / 61.5	0.43 / 57.2
AUDIOSPHERE†	**0.29**±0.01 / **82.1**±0.8	**0.38**±0.00 / **74.7**±0.3	0.42±0.01 / **71.1**±0.2
DOA	LE° ↓ / LR ↑	LE° ↓ / LR ↑	LE° ↓ / LR ↑
SELDNet [Adavanne et al., 2019a]	26.6 / 64.9	33.7 / 41.5	36.1 / 24.6
DOAnet [Adavanne et al., 2018]	**6.3** / 46.5	**20.1** / 11.5	25.8 / 2.9
Spatial LibriSpeech [Sarabia et al., 2023]	12.4 / –	–	–
ELSA† [Devnani et al., 2024]	15.0 / –	–	–
AUDIOSPHERE†	13.4±1.7 / **83.2**±1.0	20.8±1.9 / **48.0**±1.7	**25.3**±2.0 / **26.1**±2.0

5.2 AUDIOSPHERE Transfers to Static and Dynamic Real-World Scenes

Static Scenes. Tables 3 and 4 show that AUDIOSPHERE transfers to static SELD datasets TUT-2018 Real [Adavanne et al., 2019a] and TAU-2019 [Adavanne et al., 2019b]. On TUT-2018 Real, the frozen model achieves the highest F-score and localization recall at every polyphony level and the lowest error rate on overlap 1. For localization error, DOAnet retains a clear lead on overlap 1 ($6.3°$ vs. $13.4° \pm 1.7°$); on overlaps 2 and 3 its point estimates fall within one standard deviation of AUDIOSPHERE's seed means. On TAU-2019, AUDIOSPHERE improves on the in-domain supervised baseline SELDNet in detection and localization while SELDNet retains higher LR, and it scores well above the reported w2v-SELD results and the evaluated BSM + Spatial-AST (BAT) pipeline, suggesting that patch-level self-supervision yields frame-resolved spatial features that clip-level supervised classification does not.

Table 4: **Static real-world SELD on TAU-2019**. Same protocol, metrics and notation as Table 3. Baselines: SELDNet [Adavanne et al., 2019a], BAT [Zheng et al., 2024], w2v-SELD [dos Santos et al., 2023].

	SED	DOA
Method	ER↓ / F↑	LE° ↓ / LR ↑
Supervised		
SELDNet	0.34 / 79.9	28.5 / **85.4**
Self-supervised (frozen)		
BAT	0.92 / 19.3	57.5 / 38.0
w2v-SELD	0.56 / 65.2	80.6 / 58.4
AUDIOSPHERE	**0.28** / **84.6**	**14.1** / 78.5
± SD	0.05 / 2.1	1.3 / 0.6

Dynamic scenes. Table 5 demonstrates AUDIOSPHERE's out-of-domain generalization on dynamic scenes. With a frozen backbone and shallow MLP probe, AUDIOSPHERE surpasses the supervised baseline on the simpler TAU-2020 ($ER_{20°}$ 0.57 vs. 0.72; $F_{20°}$ 46.6 vs. 37.4). However, without the temporal continuity needed to track moving sources, the MLP probe lags on the highly polyphonic TAU-2021. Attaching a lightweight temporal head (TM) to the frozen backbone recovers this performance. On TAU-2021, $F_{20°}$ jumps from 21.0 to 37.2 and LE_{CD} drops from $36.9°$ to $24.3°$, suggesting that AUDIOSPHERE's frame-level embeddings paired with a temporal head tracks and localizes highly polyphonic sources. Furthermore, fine-tuning yields the strongest overall results both with MLP and the TM.

Transfer to STARSS23. Figure 2 illustrates the benefit of our learned representations compared to a randomly initialized model when transferred to the STARSS23 dataset, which consists of challenging real-world dynamic recordings. Comparing validation curves confirms that AUDIOSPHERE provides an exceptionally strong initialization: fine-tuning converges faster, achieves lower localization error, and attains higher localization recall than training from scratch. Under frozen shallow-MLP probing, the overall localization error on STARSS23 is $47.1°$; however, a class-level analysis (Appendix J) reveals that this figure is heavily inflated by undetected brief sound events, which yield the maximum possible error under direct transfer. Fitting the temporal head to the frozen backbone recovers these classes, raising $F_{20°}$ to 37.4 and reducing LE_{CD} from $47.1°$ to $22.8°$ (Appendix J), and additionally Figure 2 shows that fine-tuning eliminates this degradation.

Table 5: **Dynamic real-world SELD.** Frozen probing and fine-tuning on TAU-2020/2021 with the shallow MLP probe and with the lightweight temporal head (TM) of Section 4. **Bold** = best per column.

	TAU-2020				TAU-2021			
Method	$\mathbf{ER}_{20^\circ}\downarrow$	$\mathbf{F}_{20^\circ}\uparrow$	$\mathbf{LE}_{\mathbf{CD}}\downarrow$	$\mathbf{LR}_{\mathbf{CD}}\uparrow$	$\mathbf{ER}_{20^\circ}\downarrow$	$\mathbf{F}_{20^\circ}\uparrow$	$\mathbf{LE}_{\mathbf{CD}}\downarrow$	$\mathbf{LR}_{\mathbf{CD}}\uparrow$
Supervised baseline								
SELDNet	0.72	37.4	22.8	60.7	0.73	30.7	24.5	40.5
Probing (frozen backbone)								
Frozen + MLP	0.57	46.6	23.5	63.2	0.75	21.0	36.9	49.1
Frozen + TM	0.41	63.6	15.0	75.0	0.63	37.2	24.3	61.3
Fine-tuning								
Fine-tune + MLP	0.42	65.8	13.4	**78.2**	0.58	47.9	18.2	66.3
Fine-tune + TM	**0.40**	**68.4**	**11.1**	77.8	**0.51**	**57.7**	**14.6**	**67.4**

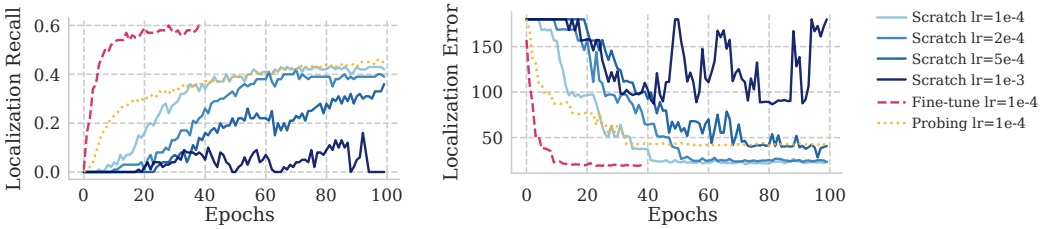

Figure 2: **Training dynamics on STARSS23.** Validation localization recall and localization error across epochs with a shallow MLP head attached to the backbone.

5.3 AUDIOSPHERE Works on Diverse Array Geometries

Table 6a shows that AUDIOSPHERE transfers to tetrahedral microphone recordings on TAU-2019 without retraining the encoder. This table demonstrates that the detection is unchanged (ER 0.28 vs. 0.25) and localization degrades moderately (LE 21.2° vs. 14.1°). The same holds on the dynamic TAU-2020, TAU-2021 and STARSS23 datasets (Appendix D.1, Table 7). The residual gap to native FOA is due to spatial sampling and the ASM conversion. A 4-channel tetrahedral array samples the first-order field at lower fidelity than the 32-channel Eigenmike. This observation is evidenced by SELDNet degrading by a similar amount on ASM-FOA input, and Appendix E quantifying the cue degradation directly.

On real-world recordings of spatial sound scenes from a planar array (Table 6b), the ASM + AUDIOSPHERE pipeline achieves the lowest median DoA error of 1.9°, compared with 5.6° for MUSIC and 13.1° for the evaluated BSM + Spatial-AST pipeline. Evaluating the frozen AUDIOSPHERE encoder on the same binaural renderings used for Spatial-AST (Appendix D.2) yields a median error of 3.4°, indicating that the test-time input representation alone does not explain the difference. As front–back confusion is a common limitation of binaural input, we also report front–back folded predictions: folding reduces the Spatial-AST error from 13.1° to about 4.3°, so front–back confusion explains a substantial but not complete part of the gap (azimuth confusion matrices in Appendix M).

5.4 AUDIOSPHERE Utilizes Spatial Cues to Improve Scene Analysis

We further isolate the contribution of spatial cues by constructing three reduced-input variants. In the first variant, AUDIOSPHERE (W duplicated), we take the Ambisonics-pretrained model but replace the X, Y, and Z channels with copies of W at inference time, effectively stripping directional cues. In the second variant, AUDIOSPHERE-W is pre-trained exclusively on the omnidirectional W channel, removing all spatial information from the learned representations while retaining reverberation and noise characteristics. The last variant is AUDIOSPHERE-Clean, which is pre-trained solely on the AudioSet. All variants were pre-trained for 125k steps.

Table 12 (Appendix F.1) shows that the full Ambisonics input outperforms all reduced variants across detection and localization metrics. Localization error drops from approximately 73° for all

Table 6: **Array geometry.** (a) Native Ambisonics vs. ASM-encoded tetrahedral input on TAU2019 for SELDNet and frozen AUDIOSPHERE. (b) DoA error on RSL2019-dev (median $\pm$ IQR). **Bold** = best per column.

Input	ER↓	F↑	LE↓	LR↑
SELDNet (supervised)				
Ambisonics	0.34	79.9	28.5	85.4
Mic. array	0.35	80.0	30.8	84.0
AUDIOSPHERE (self-supervised, frozen)				
Ambisonics	0.28$\pm$0.05	84.6$\pm$2.1	14.1$\pm$1.3	78.5$\pm$0.6
Mic. array	0.25	84.8	21.2	73.6

Method	DoA error (°)↓
Baselines	
BSM + Spatial-AST (BAT)	13.1 ± 29.0
MUSIC	5.6 ± 11.6
Ours	
ASM + AUDIOSPHERE	$\mathbf{1.9} \pm 6.6$

three reduced variants to $14.9°$ for the full AUDIOSPHERE model, confirming that inter-channel spatial structure is indispensable for accurate direction-of-arrival estimation. On the detection side, AUDIOSPHERE-Clean and AUDIOSPHERE-W perform comparably (ER 0.29 vs. 0.29; F 81.9 vs. 82.2). This suggests that the simulated reverberation and noise added during pre-training have a negligible impact on event detection, and that potential double-reverberation artifacts do not degrade performance. In contrast, AUDIOSPHERE (W duplicated) yields noticeably worse detection (ER 0.33, F 79.9), likely because the model expects informative spatial variation across channels but instead receives identical copies of W, producing a distribution mismatch that degrades its learned representations. The full AUDIOSPHERE model achieves the best detection performance overall (ER 0.24, F 85.7). Altogether, these results confirm that multi-channel Ambisonics structure provides complementary information beyond what is available in the mono signal alone, benefiting both detection and localization performance.

6 Discussion & Conclusion

We presented AUDIOSPHERE, a self-supervised spatial audio embedding model based on a multi-channel masked autoencoder that learns joint spatial-acoustic audio representations. It processes first-order Ambisonics natively and, through ASM, generalizes to diverse microphone geometries without retraining. Frozen AUDIOSPHERE transfers to clip-level classification of environmental sounds, speech and music, to frame-level sound event detection, and to real-world SELD. It is the first self-supervised model to acquire spatial understanding while retaining this monaural transfer. We additionally release REALSELD, the first probing benchmark for SELD, comprising seven real-world tasks drawn from five established datasets designed to evaluate pre-trained audio embeddings for SELD.

Limitations. First, while ASM enables hardware flexibility, the quality of the Ambisonics encoding is bounded by the spatial sampling capacity of the physical hardware. For example, localization errors on TAU-2020, TAU-2021, and STARSS23 increase when replacing the 32-channel Eigenmike signal with a 4-channel tetrahedral decomposition. Native, geometry-agnostic microphone handling directly within the backbone could bypass this. Second, AUDIOSPHERE is pre-trained to reconstruct independent, frame-level spatial snapshots. Consequently, its frozen representations lack the intrinsic temporal-continuity priors necessary to resolve high polyphony and moving sources. A temporal probing head or fine-tuning bridges this gap, but a shallow MLP probe on a fully frozen backbone struggles to track moving mixtures. Future work could explore temporal SSL objectives, such as multi-source reconstruction over longer time windows, to learn motion priors directly from dynamic scenes. Finally, our spatial evaluation covers sound event detection and DoA estimation only; it does not evaluate distance estimation, source separation, dereverberation, or room-acoustic inference.

Future Research. Three avenues appear most promising. First, scaling training data using large-scale spatial corpora, such as $360°$ video platforms, could reduce reliance on simulation, further narrow the sim-to-real gap, and expose the model to natural motion dynamics. Second, integrating AUDIOSPHERE with large language models could enable natural-language reasoning of spatial sound scenes. Third, extending AUDIOSPHERE to distance estimation would enable complete 3D source localization, moving beyond the current directional-only representation.

Acknowledgments and Disclosure of Funding

This project received funding from the NWO Talent Program (VI.Veni.202.184; KH). This work used the Dutch national e-infrastructure with the support of the SURF Cooperative, using grant no. EINF-12218. We would like to thank Robert Jan Schlimbach from the Snellius team for helpful discussions and their help with high-performance cluster utilization. We thank Yhonatan Gayer for the support with the SHroom framework for Ambisonics Signal Matching and for the feedback on our description of the spatial-sampling requirements of Ambisonics encoding. The authors declare no competing financial interests.

References

Sharath Adavanne, Pasi Pertilä, and Tuomas Virtanen. Sound event detection using spatial features and convolutional recurrent neural network. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 771–775, 2017. doi: 10.1109/ICASSP.2017.7952260.

Sharath Adavanne, Archontis Politis, and Tuomas Virtanen. Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 1462–1466, 2018. doi: 10.23919/EUSIPCO.2018.8553182.

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, 2019a. doi: 10.1109/JSTSP.2018.2885636.

Sharath Adavanne, Archontis Politis, and Tuomas Virtanen. A multi-room reverberant dataset for sound event localization and detection. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, pages 10–14, New York University, NY, USA, October 2019b. URL `https://dcase.community/workshop2019/proceedings`.

Akshay Anantapadmanabhan, Ashwin Bellur, and Hema A Murthy. Modal analysis and transcription of strokes of the mridangam using non-negative matrix factorization. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 181–185, 2013. doi: 10.1109/ICASSP.2013.6637633.

Alan Baade, Puyuan Peng, and David Harwath. MAE-AST: Masked autoencoding audio spectrogram transformer. In *Interspeech 2022*, pages 2438–2442, 2022. doi: 10.21437/Interspeech.2022-10961.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Amy Bastine, Lachlan Birnie, Thushara D. Abhayapala, Prasanga Samarasinghe, and Vladimir Tourbabin. Ambisonics capture using microphones on head-worn device of arbitrary geometry. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 309–313, 2022. doi: 10.23919/EUSIPCO55093.2022.9909803.

Albert S Bregman. Auditory scene analysis. In *Proceedings of the 7th International Conference on Pattern Recognition*, pages 168–175. Citeseer, 1984.

Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Trans Affect Comput*, 5 (4):377–390, October 2014.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017.

Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip W Robinson, and Kristen Grauman. Soundspaces 2.0: A simulation platform for visual-acoustic learning. In *NeurIPS 2022 Datasets and Benchmarks Track*, 2022.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, and Furu Wei. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *CoRR*, abs/2110.13900, 2021. URL http://dblp.uni-trier.de/db/journals/corr/corr2110.html#abs-2110-13900.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. BEATs: Audio pre-training with acoustic tokenizers. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 5178–5193. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/chen23ag.html.

Dading Chong, Helin Wang, Peilin Zhou, and Qingcheng Zeng. Masked spectrogram prediction for self-supervised audio pre-training. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10095691.

Bhavika Devnani, Skyler Seto, Zakaria Aldeneh, Alessandro Toso, Elena Menyaylenko, Barry-John Theobald, Jonathan Sheaffer, and Miguel Sarabia. Learning spatially-aware language and audio embeddings. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 33505–33537. Curran Associates, Inc., 2024. doi: 10.52202/079017-1056.

Heinrich Dinkel, Zhiyong Yan, Yongqing Wang, Junbo Zhang, Yujun Wang, and Bin Wang. Scaling up masked audio encoder learning for general audio classification. In *Interspeech 2024*, 2024.

Orlem Lima dos Santos, Karen Rosero, and Roberto de Alencar Lotufo. w2v-SELD: A sound event localization and detection framework for self-supervised spatial audio pre-training, 2023. URL https://arxiv.org/abs/2312.06907.

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. *ICLR*, 2021.

Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Douglas Eck, Karen Simonyan, and Mohammad Norouzi. Neural audio synthesis of musical notes with wavenet autoencoders, 2017.

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. doi: 10.1109/TASLP.2021.3133208.

Yhonatan Gayer. SHroom: A python framework for ambisonics room acoustics simulation and binaural rendering, 2026. URL https://arxiv.org/abs/2603.27342.

Yhonatan Gayer, Vladimir Tourbabin, Zamir Ben-Hur, David Alon, and Boaz Rafaely. Ambisonics encoder for wearable array with improved binaural reproduction, 2025. URL https://arxiv.org/abs/2507.04108.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio Set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017. doi: 10.1109/ICASSP.2017.7952261.

Gene H. Golub, Per Christian Hansen, and Dianne P. O'Leary. Tikhonov regularization and total least squares. *SIAM Journal on Matrix Analysis and Applications*, 21(1):185–194, 1999. doi: 10.1137/S0895479897326432. URL https://doi.org/10.1137/S0895479897326432.

Yuan Gong, Cheng-I Lai, Yu-An Chung, and James Glass. SSAST: Self-supervised audio spectrogram transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10699–10709, 2022.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Mikko Heikkinen, Archontis Politis, Konstantinos Drossos, and Tuomas Virtanen. Beyond omnidirectional: Neural ambisonics encoding for arbitrary microphone directivity patterns using cross-attention, 2026. URL https://arxiv.org/abs/2601.23196.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:3451–3460, October 2021. ISSN 2329-9290. doi: 10.1109/TASLP.2021.3122291. URL https://doi.org/10.1109/TASLP.2021.3122291.

Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. In *NeurIPS*, 2022.

Xilin Jiang, Hannes Gamper, and Sebastian Braun. Sci-phi: A large language model spatial audio descriptor, 2025. URL https://arxiv.org/abs/2510.05542.

Jaeyeon Kim, Heeseung Yun, and Gunhee Kim. ViSAGe: Video-to-spatial audio generation. In *ICLR*, 2025.

Xian Li, Nian Shao, and Xiaofei Li. Self-supervised audio teacher-student transformer for both clip-level and frame-level tasks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:1336–1351, 2024. doi: 10.1109/TASLP.2024.3352248.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Matthew Maciejewski, Gordon Wichern, and Jonathan Le Roux. WHAMR!: Noisy and reverberant single-channel speech separation. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020.

Leo McCormack, Archontis Politis, Raimundo Gonzalez, Tapio Lokki, and Ville Pulkki. Parametric ambisonic encoding of arbitrary microphone arrays. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2062–2075, 2022. doi: 10.1109/TASLP.2022.3182857.

A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley. Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(2):379–393, Feb 2018. ISSN 2329-9290. doi: 10.1109/TASLP.2017.2778423.

Pedro Morgado, Yi Li, and Nuno Nvasconcelos. Learning representations from audio-visual spatial alignment. *Advances in Neural Information Processing Systems*, 33, 2020.

Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. Masked spectrogram modeling using masked autoencoders for learning general-purpose audio representation. In Joseph Turian, Björn W. Schuller, Dorien Herremans, Katrin Kirchoff, Paola Garcia Perera, and Philippe Esling, editors, *HEAR: Holistic Evaluation of Audio Representations (NeurIPS 2021 Competition)*, volume 166 of *Proceedings of Machine Learning Research*, pages 1–24. PMLR, 13–14 Dec 2022. URL https://proceedings.mlr.press/v166/niizumi22a.html.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964.

Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pages 1015–1018. ACM Press. ISBN 978-1-4503-3459-4. doi: 10.1145/2733373.2806390. URL http://dl.acm.org/citation.cfm?doid=2733373.2806390.

Archontis Politis. [dcase2022 task 3] synthetic seld mixtures for baseline training, April 2022. URL https://doi.org/10.5281/zenodo.6406873.

Archontis Politis, Sharath Adavanne, and Tuomas Virtanen. A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, pages 165–169, Tokyo, Japan, November 2020a. URL https://dcase.community/workshop2020/proceedings.

Archontis Politis, Annamaria Mesaros, Sharath Adavanne, Toni Heittola, and Tuomas Virtanen. Overview and evaluation of sound event localization and detection in DCASE 2019. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:684–698, 2020b. URL https://ieeexplore.ieee.org/abstract/document/9306885.

Archontis Politis, Sharath Adavanne, Daniel Krause, Antoine Deleforge, Prerak Srivastava, and Tuomas Virtanen. A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection. In *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, pages 125–129, Barcelona, Spain, November 2021. ISBN 978-84-09-36072-7. URL https://dcase.community/workshop2021/proceedings.

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, 2022. URL https://arxiv.org/abs/2206.01948.

Aaqib Saeed, David Grangier, and Neil Zeghidour. Contrastive learning of general-purpose audio representations. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3875–3879, 2021. doi: 10.1109/ICASSP39728.2021.9413528.

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. In *Proc. Interspeech*, pages 3724–3728, 2023. doi: 10.21437/Interspeech.2023-2117.

R. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34(3):276–280, 1986. doi: 10.1109/TAP.1986.1143830.

Rohan Sheelvant, Bidisha Sharma, Maulik Madhavi, Rohan Kumar Das, S. R. M. Prasanna, and Haizhou Li. RSL2019: A realistic speech localization corpus. In *2019 22nd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6, 2019. doi: 10.1109/O-COCOSDA46868.2019.9060842.

Kazuki Shimada, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, and Yuki Mitsufuji. ACCDOA: Activity-coupled cartesian direction of arrival representation for sound event localization and detection, 2021. URL https://arxiv.org/abs/2010.15306.

Kazuki Shimada, Archontis Politis, Parthasaarathy Sudarsanam, Daniel 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 *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Fabian-Robert Stöter, Soumitro Chakrabarty, Emanuël Habets, and Bernd Edler. Libricount, a dataset for speaker count estimation, April 2018. URL https://doi.org/10.5281/zenodo.1216072.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Jun Zhang, Lu Lu, Zejun Ma, Yuxuan Wang, and Chao Zhang. Can large language models understand spatial audio?, 2024. URL https://arxiv.org/abs/2406.07914.

Mi Tian, Ajay Srinivasamurthy, Mark Sandler, and Xavier Serra. A study of instrument-wise onset detection in Beijing Opera percussion ensembles. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2159–2163, 2014. doi: 10.1109/ICASSP.2014.6853981.

Joseph Turian, Jordie Shier, Humair Raj Khan, Bhiksha Raj, Björn W. Schuller, Christian J. Steinmetz, Colin Malloy, George Tzanetakis, Gissel Velarde, Kirk McNally, Max Henry, Nicolas Pinto, Camille Noufi, Christian Clough, Dorien Herremans, Eduardo Fonseca, Jesse Engel, Justin Salamon, Philippe Esling, Pranay Manocha, Shinji Watanabe, Zeyu Jin, and Yonatan Bisk. HEAR: Holistic Evaluation of Audio Representations. In Douwe Kiela, Marco Ciccone, and Barbara Caputo, editors, *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*, volume 176 of *Proceedings of Machine Learning Research*, pages 125–145. PMLR, 06–14 Dec 2022. URL https://proceedings.mlr.press/v176/turian22a.html.

Jörgen Valk and Tanel Alumäe. Voxlingua107: A dataset for spoken language recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 652–658, 2021. doi: 10.1109/SLT48900.2021.9383459.

Kiki van der Heijden, Josef P Rauschecker, Beatrice de Gelder, and Elia Formisano. Cortical mechanisms of spatial hearing. *Nature Reviews Neuroscience*, 20(10):609–623, 2019.

Shanshan Wang, Archontis Politis, Annamaria Mesaros, and Tuomas Virtanen. Self-supervised learning of audio representations from audio-visual data using spatial alignment. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1467–1479, 2022. doi: 10.1109/JSTSP.2022.3180592.

Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition, 2018. URL https://arxiv.org/abs/1804.03209.

Sarthak Yadav and Zheng-Hua Tan. Audio Mamba: Selective state spaces for self-supervised audio representations. In *Interspeech 2024*, pages 552–556, 2024. doi: 10.21437/Interspeech.2024-1274.

Sarthak Yadav, Sergios Theodoridis, Lars Kai Hansen, and Zheng-Hua Tan. Masked autoencoders with multi-window local-global attention are better audio learners. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Q53QLftNkA.

Shuwen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee. Superb: Speech processing universal performance benchmark. In *Interspeech 2021*, pages 1194–1198, 2021. doi: 10.21437/Interspeech.2021-1775.

Zhisheng Zheng, Puyuan Peng, Ziyang Ma, Xie Chen, Eunsol Choi, and David Harwath. Bat: learning to reason about spatial sounds with large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Franz Zotter and Matthias Frank. *XY, MS, and First-Order Ambisonics*, pages 1–22. Springer International Publishing, Cham, 2019. ISBN 978-3-030-17207-7. doi: 10.1007/978-3-030-17207-7_1. URL https://doi.org/10.1007/978-3-030-17207-7_1.

A Ambisonics

A.1 Real Spherical Harmonics

The basis functions $Y_{lm}(\theta, \phi)$ are the angular part of the solution to the wave equation in spherical coordinates, where $\theta \in [0, 2\pi)$ denotes azimuth and $\phi \in [-\pi/2, \pi/2]$ denotes elevation. For real-valued audio signals we use real spherical harmonics defined via the associated Legendre functions $P_l^{|m|}(\sin\phi)$ and a normalization constant N_{lm}:

$$Y_{lm}(\theta, \phi) \;=\; N_{lm}\, P_l^{|m|}(\sin\phi) \cdot \begin{cases} \sin(|m|\,\theta) & m < 0 \\ 1 & m = 0 \\ \cos(m\,\theta) & m > 0 \end{cases} \tag{4}$$

We adopt the three-dimensional (N3D) normalization convention used in Ambisonics:

$$N_{lm}^{\mathrm{N3D}} \;=\; \sqrt{(2 - \delta_{m,0})\,(2l+1)\,\frac{(l-|m|)!}{(l+|m|)!}}\,, \tag{5}$$

where $\delta_{m,0}$ is the Kronecker delta. The associated Legendre functions are taken without the Condon–Shortley phase:

$$P_l^{|m|}(x) \;=\; (1-x^2)^{|m|/2}\,\frac{d^{|m|}}{dx^{|m|}} P_l(x). \tag{6}$$

A.2 Ambisonic Channel Number (ACN) Ordering

We adopt the Ambisonic Channel Number (ACN) component ordering, in which the channel index k for a coefficient of order l and degree m is given by

$$k \;=\; l(l+1) + m, \qquad 0 \le l \le L,\; -l \le m \le l. \tag{7}$$

For first-order Ambisonics ($L = 1$), this yields the four channels (W, Y, Z, X) corresponding to $(l, m) \in \{(0,0), (1,-1), (1,0), (1,1)\}$, with closed-form expressions

$$W = Y_{0,0}(\theta, \phi) = 1, \tag{8}$$

$$Y = Y_{1,-1}(\theta, \phi) = \sqrt{3}\,\cos\phi\,\sin\theta, \tag{9}$$

$$Z = Y_{1,0}(\theta, \phi) = \sqrt{3}\,\sin\phi, \tag{10}$$

$$X = Y_{1,1}(\theta, \phi) = \sqrt{3}\,\cos\phi\,\cos\theta. \tag{11}$$

B Ambisonics Signal Matching (ASM)

B.1 Ambisonics Representation

We define the Ambisonics representation of a signal $\mathbf{s}(k)$ composed of Q plane waves as

$$\mathbf{a_{nm}}(k) = \mathbf{Y}_{\mathbf{\Omega}_Q}^H \mathbf{s}(k), \tag{12}$$

where $\mathbf{Y}_{\mathbf{\Omega}_Q}$ of size $Q \times (N_a + 1)^2$ is the spherical harmonics (SH) matrix. Its columns are $\mathbf{y_{nm}} = [Y_{nm}(\theta_1, \phi_1), \ldots, Y_{nm}(\theta_Q, \phi_Q)]^T$, i.e. the real SH functions of order n and degree m evaluated at the source directions (θ_q, ϕ_q). The vector $\mathbf{a_{nm}}(k)$ collects the Ambisonics coefficients up to order N_a.

B.2 Ambisonics Encoding for Arbitrary Arrays

To bypass the geometric constraints of spherical arrays, Ambisonics encoding can be generalized to arbitrary array geometries via Tikhonov regularization. We assume the standard spatial observation model in which the raw microphone signal vector is the sum of Q plane waves and additive noise:

$$\mathbf{x} = \mathbf{V}\mathbf{s} + \mathbf{n}, \tag{13}$$

with $\mathbf{V} \in \mathbb{C}^{M\times Q}$ the array steering matrix whose (m,q)-th entry is the response of microphone m to a unit-amplitude plane wave arriving from $\mathbf{\Omega}_q$, $\mathbf{s} \in \mathbb{C}^{Q\times 1}$ the plane-wave amplitudes, and $\mathbf{n} \in \mathbb{C}^{M\times 1}$ the microphone noise vector. Encoding is posed as a linear mapping from $\mathbf{x}$ to the estimated Ambisonics coefficients:

$$\hat{a}_{nm} = \mathbf{c}_{nm}^H \mathbf{x}, \quad \forall\, 0 \le n \le N_a,\ -n \le m \le n, \tag{14}$$

where $\mathbf{c}_{nm} \in \mathbb{C}^{M\times 1}$ is the encoding filter. The optimal filter is obtained by minimizing the mean squared error (MSE)

$$\varepsilon_{nm}^{\mathrm{ASM}} = \frac{\mathbb{E}\big[\|\hat{a}_{nm} - a_{nm}\|_2^2\big]}{\mathbb{E}\big[\|a_{nm}\|_2^2\big]}. \tag{15}$$

To obtain a signal-independent solution we assume an isotropic diffuse sound field ($\mathbf{R}_s = \sigma_s^2\mathbf{I}$) and spatially white noise ($\mathbf{R}_n = \sigma_n^2\mathbf{I}$). Substituting (12) and (14) into (15) yields

$$\varepsilon_{nm}^{\mathrm{ASM}} = \frac{\sigma_s^2\|\mathbf{V}^H\mathbf{c}_{nm} - \mathbf{y}_{nm}\|_2^2 + \sigma_n^2\|\mathbf{c}_{nm}\|_2^2}{\sigma_s^2\|\mathbf{y}_{nm}\|_2^2}. \tag{16}$$

Minimizing (16) with respect to $\mathbf{c}_{nm}$ gives the optimal Ambisonics Signal Matching (ASM) filters; the term $\frac{\sigma_n^2}{\sigma_s^2}\mathbf{I}$ serves as the regularizer. Writing the result in the row-vector (Hermitian) form that matches (14):

$$\big[\mathbf{c}_{nm}^{\mathrm{ASM}}\big]^H = \mathbf{y}_{nm}^H\mathbf{V}^H\big(\mathbf{V}\mathbf{V}^H + \tfrac{\sigma_n^2}{\sigma_s^2}\mathbf{I}\big)^{-1}. \tag{17}$$

For efficient downstream application the filters can be stacked to encode all $(N_a+1)^2$ Ambisonics channels simultaneously:

$$\big[\mathbf{C}_{\mathbf{nm}}^{\mathrm{ASM}}\big]^H = \big[\mathbf{c}_{00}^{\mathrm{ASM}}, \ldots, \mathbf{c}_{N_aN_a}^{\mathrm{ASM}}\big]^H = \mathbf{Y}_{\mathbf{\Omega}_Q}^H\mathbf{V}^H\big(\mathbf{V}\mathbf{V}^H + \tfrac{\sigma_n^2}{\sigma_s^2}\mathbf{I}\big)^{-1}. \tag{18}$$

C Binaural Signal Matching (BSM) for Arbitrary Arrays

Starting from the Ambisonics encoding, binaural reproduction is obtained by

$$p^{l,r}(k) = \big[\mathbf{h}_{\mathbf{nm}}^{l,r}(k)\big]^T\mathbf{a}_{\mathbf{nm}}(k), \tag{19}$$

where $\mathbf{h}_{\mathbf{nm}}^{l,r}$ is the vector of SH coefficients of the left- or right-ear HRTF expressed in the real SH basis of Appendix A. Both the HRTF and the Ambisonics encoding are truncated to the same order $N_a = N_h$.

For nearly uniform spherical microphone arrays, plane-wave decomposition (PWD) accurately computes these coefficients, provided the number of microphones satisfies $(N_a+1)^2 \le M$.

Notation: For brevity the wave number k is omitted from all subsequent equations.

Rather than routing through an intermediate Ambisonics representation, Binaural Signal Matching (BSM) extracts the binaural signals directly from the arbitrary microphone array via the linear mapping

$$\hat{p}_{\mathrm{BSM}}^{l,r} = \big[\mathbf{c}_{\mathrm{BSM}}^{l,r}\big]^H\mathbf{x}. \tag{20}$$

The optimal BSM filter $\mathbf{c}_{\mathrm{BSM}}^{l,r} \in \mathbb{C}^{M\times 1}$ is obtained by minimizing the mean squared error (MSE) against the target binaural signal $p^{l,r}$:

$$\varepsilon_{\mathrm{BSM}} = \mathbb{E}\big[\|\hat{p}_{\mathrm{BSM}}^{l,r} - p^{l,r}\|_2^2\big]. \tag{21}$$

Substituting the observation model (13) together with the SH representation $\mathbf{a}_{\mathbf{nm}} = \mathbf{Y}_{\mathbf{\Omega}_Q}^H\mathbf{s}$ into (19) expresses the target as a linear function of the plane-wave amplitudes, $p^{l,r} = \big[\mathbf{h}_{\mathbf{nm}}^{l,r}\big]^T\mathbf{Y}_{\mathbf{\Omega}_Q}^H\mathbf{s}$. Applying the same isotropic diffuse field and white noise assumptions used in ASM, Tikhonov regularization yields the closed-form BSM solution:

$$\big[\mathbf{c}_{\mathrm{BSM}}^{l,r}\big]^H = \big[\mathbf{h}_{\mathbf{nm}}^{l,r}\big]^T\mathbf{Y}_{\mathbf{\Omega}_Q}^H\mathbf{V}^H\big(\mathbf{V}\mathbf{V}^H + \tfrac{\sigma_n^2}{\sigma_s^2}\mathbf{I}\big)^{-1}. \tag{22}$$

D Additional Downstream Analyses

D.1 Array-geometry transfer on dynamic scenes

Table 7 extends the TAU-2019 input-format comparison of Section 5.3 to the dynamic datasets. AUDIOSPHERE transfers to tetrahedral recordings on TAU-2020, TAU-2021, and STARSS23 without retraining the encoder, in both the frozen and fine-tuned regimes (each with the temporal head), and the fine-tuned ASM-FOA model matches or exceeds the supervised SELDNet trained directly on native Ambisonics. A residual gap to native FOA remains. The supervised SELDNet exhibits the same FOA→ASM-FOA degradation when trained from scratch, confirming that the re-encoding reduces the performance also on the supervised baseline. Appendix E quantifies the alignment between reference and ASM-encoded FOA channels and finds consistent azimuth and elevation cue degradation (median azimuth/elevation differences of 7.4° / 7.6° on TAU-2020, 8.0° / 7.6° on TAU-2021 and 9.4° / 8.3° on STARSS23).

Table 7: SELD performance on TAU2020, TAU2021, and STARSS23. *FOA*: the 32-channel Eigenmike reference Ambisonics distributed with each dataset. *ASM-FOA*: tetrahedral microphone signals encoded to first-order Ambisonics through our ASM pipeline. SELDNet is trained from scratch on the ASM-FOA input; AUDIOSPHERE is evaluated frozen and fully fine-tuned, in both cases with the temporal head (TM) of Section 4, i.e., the TM rows of Table 5. Location-dependent metrics as in Table 5 (Appendix I); ER and LE: lower is better; F and LR: higher is better.

Method	TAU2020				TAU2021				STARSS23			
	$ER_{20^\circ}\downarrow$	$F_{20^\circ}\uparrow$	$LE_{CD}\downarrow$	$LR_{CD}\uparrow$	$ER_{20^\circ}\downarrow$	$F_{20^\circ}\uparrow$	$LE_{CD}\downarrow$	$LR_{CD}\uparrow$	$ER_{20^\circ}\downarrow$	$F_{20^\circ}\uparrow$	$LE_{CD}\downarrow$	$LR_{CD}\uparrow$
Baseline (SELDNet)												
Baseline FOA	0.72	37.4	22.8	60.7	0.73	30.7	24.5	40.5	0.57	29.9	22.0	47.7
Baseline ASM-FOA	0.77	32.2	28.0	58.3	0.72	20.5	29.4	39.7	0.62	27.6	26.3	47.5
AUDIOSPHERE (frozen)												
Frozen + TM, FOA	0.41	63.6	15.0	75.0	0.63	37.2	24.3	61.3	0.56	37.4	22.8	58.8
Frozen + TM, ASM-FOA	0.49	55.5	17.7	74.9	0.71	27.0	28.8	60.3	0.58	32.6	24.8	55.4
AUDIOSPHERE (fine-tuned)												
Fine-tune + TM, FOA	**0.40**	**68.4**	**11.1**	**77.8**	**0.51**	**57.7**	**14.6**	**67.4**	**0.49**	**44.7**	**18.3**	**60.5**
Fine-tune + TM, ASM-FOA	0.44	63.7	14.5	77.2	0.58	46.4	20.9	66.7	0.52	40.9	18.9	59.2

D.2 Binaural-input evaluation of AUDIOSPHERE

To test whether the gap between the ASM + AUDIOSPHERE and BSM + Spatial-AST pipelines on RSL2019-dev (Table 6b) can be explained by the test-time input representation alone, we evaluate the frozen AUDIOSPHERE encoder on the same binaural renderings used for Spatial-AST. The BSM binaural renderings are converted to first-order Ambisonics by binaural-to-Ambisonics upmixing (a Tikhonov-regularized least-squares inversion analogous to ASM). This 2-to-4-channel mapping is underdetermined and therefore close to the worst case for ASM-style conversion. Table 8 reports the median DoA error, with and without folding front–back mirrored predictions onto one hemisphere; the corresponding azimuth confusion matrices are shown in Appendix M (Figure 4).

Table 8: Median DoA error (°) on RSL2019-dev when both encoders receive the same binaural renderings. *Folded*: predictions mirrored across the front–back axis before scoring, which removes front–back confusions.

Pipeline	Median DoA ↓	Folded ↓
BSM + Spatial-AST (BAT)	13.1	4.3
AUDIOSPHERE (binaural → FOA)	3.4	2.1
ASM + AUDIOSPHERE (Table 6b)	1.9	–

Operating on the same input spatial audio format as Spatial-AST, the frozen AUDIOSPHERE encoder attains a median DoA error of 3.4° versus 13.1°. Folding reduces AUDIOSPHERE's error from 3.4° to 2.1° and Spatial-AST's from 13.1° to 4.3°: both pipelines exhibit some front–back ambiguity, but AUDIOSPHERE is substantially less affected, and front–back confusion accounts for a substantial but not complete part of the original gap. These results indicate that the test-time input representation alone is insufficient to explain the observed gap, but they do not identify the cause of the remaining difference, which may stem from pre-training spatialization (binaural for Spatial-AST; FOA with

localized and diffuse WHAM! noise for AUDIOSPHERE), objective, supervision, architecture, or optimization. A native binaural AUDIOSPHERE variant, which would isolate encoder quality from input format, is left for future work.

E Analysis of Errors Introduced by ASM

To quantify the spatial fidelity loss introduced by our ASM pipeline, we exploit the fact that recent DCASE challenge releases ship both the raw tetrahedral microphone signals and a reference first-order Ambisonics (FOA) recording for every scene. We carry out this analysis on the FOA development splits of TAU-2020, TAU-2021, and STARSS23. Because STARSS23 contains both synthetic mixtures (folds 1,2) and real recordings of sound scenes (folds 3,4), we report these regimes separately, giving four evaluation subsets in total.

Methodology. For each paired recording we compute the broadband active intensity vector $\mathbf{I}(t,f) = \mathrm{Re}\{W^*(t,f)\,[X,Y,Z]^\top(t,f)\}$ in the short-time Fourier domain (N_{FFT}=1024, hop 256, f_s=24 kHz), and average it within the $[200, 4000]$ Hz band. The corresponding direction of arrival (DoA) is read off as $\theta = \mathrm{atan2}(I_y, I_x)$ (azimuth) and $\phi = \mathrm{atan2}\big(I_z, \sqrt{I_x^2 + I_y^2}\big)$ (elevation), following the convention of Section 3.1. We then compare the per-frame DoA estimates obtained from the reference FOA and from the ASM-FOA, retaining only frames whose joint magnitude (the per-frame minimum of $|W|$ across the two signals) exceeds the 50th percentile of the file. This restricts the comparison to acoustically active content, where intensity-vector DoA is more reliable.

Results. Figure 3 reports the agreement between the two intensity-vector DoA estimates as empirical CDFs of the absolute azimuth and elevation errors, together with the per-dataset error distributions. Median errors remain in the single-digit-degree range across all four subsets, indicating that the spatial cues required by downstream localization tasks are preserved by ASM. Errors are smallest on the TAU datasets, slightly larger on the synthetic STARSS23 fold, and largest on the real STARSS23 fold.

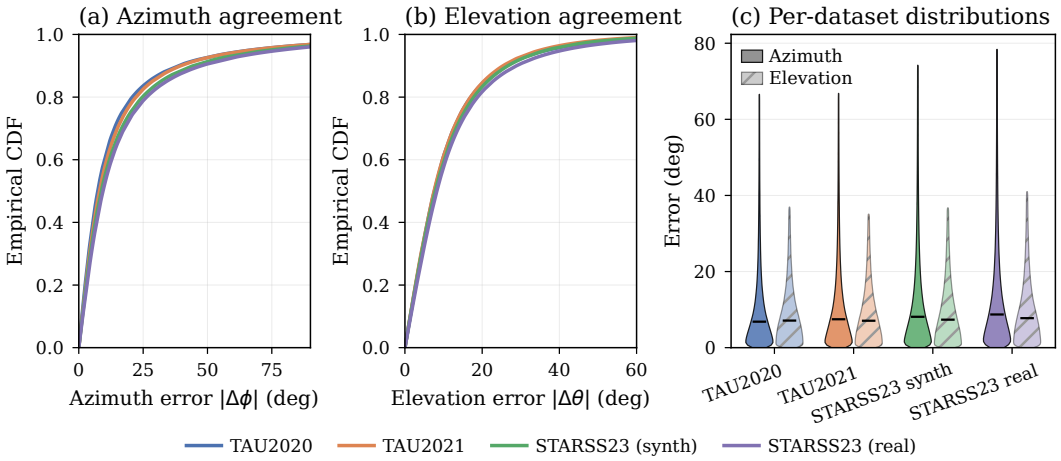

Figure 3: **Spatial-cue degradation introduced by the ASM pipeline.** Per-frame intensity-vector DoA agreement between the reference FOA and our ASM-FOA, computed in the $[200, 4000]$ Hz band on energetic frames only. **(a, b)** Empirical CDFs of the absolute azimuth error (panel axis label $|\Delta\phi|$) and elevation error (panel axis label $|\Delta\theta|$) across the four evaluation subsets. **(c)** Violin plots of the same errors per dataset (azimuth: solid; elevation: hatched), clipped at the 95th percentile.

F Ablations

Table 9 reports the design ablations discussed in Section 3.3: channel-axis masking and reconstruction targets, masking pattern, masking ratio, and reconstruction loss.

Effect of clean data ratio: We investigate the impact of the clean data ratio α, which controls the proportion of real (non-simulated, single channel AudioSet duplicated on all Ambisonic channels)

Table 9: Ablation results with frozen-encoder probes on ESC-50 (classification accuracy, mean ± std over folds) and TAU2020 (location-dependent SELD metrics ER_{20°, F_{20°, LE_{CD}, LR_{CD}; Appendix I). All encoders are pre-trained for 50k steps under an identical methodology except for the ablated factor. Rows marked with † are trained using mask-infilling. In the mask-infilling regime the encoder sees all input tokens, but some content is replaced by learned masked tokens and extra information is added to the input to indicate which parts were masked. This allows for a fair comparison between per-channel masking, which never fully hides a token, and tube masking, which completely masks out selected tokens. Bold marks the best value within each block.

	ESC-50 Acc ↑	**ER_{20°** ↓	**F_{20°** ↑	**LE_{CD}** ↓	**LR_{CD}** ↑
Channel-axis masking / reconstruction target (mask-infilling regime)					
Tube masking†	$\mathbf{72.3 \pm 2.5}$	**0.60**	**41.7**	**20.7**	**63.4**
Per-channel masking†	62.0 ± 4.4	0.66	34.5	22.9	52.3
W+IVs → XYZ†	48.4 ± 3.6	0.74	21.8	45.4	34.0
Mels → IVs†	33.5 ± 2.4	0.98	0.8	147.2	3.8
Masking pattern (drop regime, mask ratio 0.8)					
Frequency	69.8 ± 2.7	0.58	45.1	20.1	62.9
Time	71.2 ± 3.2	0.61	40.4	21.5	62.4
Clustered (block)	73.7 ± 1.3	**0.57**	**46.6**	**19.1**	63.7
Random	$\mathbf{76.1 \pm 2.2}$	0.58	46.4	19.9	**66.7**
Masking ratio (drop regime, random pattern)					
$r = 0.2$	71.9 ± 4.2	0.58	43.0	20.5	63.0
$r = 0.4$	75.0 ± 1.7	0.58	45.5	20.3	**67.9**
$r = 0.6$	$\mathbf{76.4 \pm 1.6}$	**0.53**	**48.4**	19.7	67.1
$r = 0.8$	76.1 ± 2.2	0.58	46.4	19.9	66.7
$r = 0.9$	74.7 ± 1.6	0.56	47.5	**19.5**	65.9
Reconstruction loss (drop regime, random pattern, $r = 0.8$)					
MSE (default)	$\mathbf{76.1 \pm 2.2}$	**0.58**	**46.4**	19.9	**66.7**
Mel MSE + IV cosine	71.2 ± 1.5	0.60	46.3	**19.2**	61.6

Table 10: Ablation over the clean data ratio α. Higher α denotes a larger share of real recordings during pre-training; α=0.0 uses only simulated data on TAU-2019. Here, we evaluated the models trained for 125k steps.

α	**ER** ↓	**F** ↑	**LE°** ↓	**LR** ↑
0.75	**0.22** ± 0.04	**87.3** ± 2.5	14.9 ± 0.5	78.2 ± 0.9
0.5	0.23 ± 0.04	86.8 ± 2.8	**14.4** ± 0.6	**78.7** ± 1.4
0.25	0.23 ± 0.04	86.8 ± 2.5	15.2 ± 0.5	77.9 ± 0.6
0.0	0.24 ± 0.04	85.7 ± 3.1	14.9 ± 1.6	77.0 ± 1.5

recordings relative to simulated data during pre-training. A value of α=0.0 corresponds to a fully simulated training set, while higher values progressively incorporate more clean data. Table 14 reports HEAR scores across all downstream tasks for $\alpha \in \{0.0, 0.25, 0.5, 0.75\}$.

Even a fully simulated pre-training set (α=0.0) yields competitive performance, achieving an average score of 81.1 that already surpasses most AudioSet SSL baselines except Dasheng and ATST-Frame. Progressively incorporating clean data improves the average to 83.2 at α=0.75, with the most pronounced gains in speech tasks: VL improves by 5.2 points and CD by 4.0 points as more real recordings are introduced. Acoustic event tasks follow a similar trend, with DCASE and ESC-50 benefiting from higher clean data ratios. Music tasks, by contrast, remain largely unaffected: Mri-S and Mri-T are saturated above 97 across all configurations. Notably, the majority of the improvement is captured already at α=0.25 (+1.9 points over the fully simulated baseline), indicating diminishing returns as the clean data share grows. The spike in performance after incorporating

clean data signifies that AUDIOSPHERE with fully simulated data experiences domain mismatch, and incorporating clean data randomly serves as a remedy that improves transfer to the monaural HEAR tasks. Furthermore, TAU-2019 results suggest that incorporating clean data does not degrade the performance of AUDIOSPHERE on fully Ambisonics data, and even improves the performance on detection tasks. This finding is also observed in [Devnani et al., 2024].

Spatial cue supervision: To determine whether AUDIOSPHERE leverages directional information for scene analysis, we evaluate on the TAU-2019 dataset under two training objectives: multi-label classification (non-spatial) and ACCDOA (spatially-aware) [Shimada et al., 2021]. As shown in Table 11, the ACCDOA objective yields a substantially lower error rate (0.28) than both the classification-only variant (0.40) and the supervised SELDNet baseline (0.34). This result indicates that the Ambisonics encoder in AUDIOSPHERE captures spatial structure that aids sound event discrimination, and that jointly predicting direction-of-arrival provides a useful inductive bias for high-level acoustic scene understanding.

Table 11: Ablation of spatial cues on TAU-2019 scene analysis. Error Rate (ER, ↓) is reported for AUDIOSPHERE under two training objectives and a supervised baseline.

Model	**ER ↓**
Supervised	
SELDNet	0.34
AUDIOSPHERE	
Multi-label classification	0.40
ACCDOA	**0.28**

F.1 Spatial cues

Table 12: **Ablation of spatial cues on TAU-2019.** We progressively remove or corrupt the spatial information available to AUDIOSPHERE. *Clean* uses the audio recordings from AudioSet without reverberation and noise addition (matching to MW-MAE training), *W* adds reverberation and noise back. *W dup.* replaces the directional X/Y/Z channels with copies of W during inference. Localization metrics collapse without genuine spatial cues, confirming the model relies on directional information rather than spectral shortcuts.

Method	**ER ↓**	**F ↑**	**LE° ↓**	**LR ↑**
AUDIOSPHERE-Clean	0.29	81.9	73.0	64.3
AUDIOSPHERE-W	0.29	82.2	73.2	65.2
AUDIOSPHERE (W dup.)	0.33	79.9	73.5	62.5
AUDIOSPHERE	**0.24**	**85.7**	**14.9**	**77.0**

Table 12 reports the spatial-cue ablation discussed in Section 5.4.

G Downstream Evaluation

Table 13 lists the abbreviations, task descriptions, and the types we have used to benchmark our models.

Table 14 reports the per-task HEAR results summarized in Table 2 of the main text.

We compare the performance and efficiency of AUDIOSPHERE on downstream tasks with state-of-the-art self-supervised audio representation models with a similar number of parameters as AUDIOSPHERE (90 M): MAE-16x16 [Huang et al., 2022], SSAST-patch [Gong et al., 2022], BEATs-iter3 [Chen et al., 2023], MW-MAE-B-200-4x16 [Yadav et al., 2024], SSAM [Yadav and Tan, 2024]; self-supervised speech representation models Wav2Vec 2.0 Base [Baevski et al., 2020], HuBERT Base [Hsu et al., 2021], WavLM Base [Chen et al., 2021]. Furthermore, we included BAT [Zheng et al., 2024] because it is the only model trained on spatial sound scenes. Table 15 summarizes, for each compared model, its objective and supervision, pre-training corpus and scale, native input format, evaluation-time format conversion, downstream protocol, and whether the reported numbers are reproduced by us or cited.

Table 13: Overview of the HEAR tasks.

Abbreviation	Task Name	Description	Type
DCASE	DCASE-2016 Task 2 [Mesaros et al., 2018]	Event detection of overlapping office sounds in synthetic mixtures	Scene Analysis
FSD50K	FSD50k [Fonseca et al., 2022]	Multilabel, large scale audio tagging	Scene Analysis
LC	LibriCount [Stöter et al., 2018]	Speaker Count Identification, Simulated Cocktail Party	Scene Analysis
ESC-50	ESC-50 [Piczak]	Environmental Sound Classification	Environmental Sound Classification
CD	Crema-D [Cao et al., 2014]	Emotion Recognition	Speech Analysis
VL	VoxLingua107 Top10 [Valk and Alumäe, 2021]	Spoken language identification	Speech Analysis
SC-5	Speech Command 5h [Warden, 2018]	Keyword Spotting, reduced training subset	Speech Analysis
NS	NSynth Pitch 5h [Engel et al., 2017]	Pitch Classification, reduced training subset	Pitch Classification
BO	Beijing Opera [Tian et al., 2014]	Classifying percussion instruments	Percussion
Mri-S	Mridangam Stroke [Anantapadmanabhan et al., 2013]	Stroke classification in pitched percussion instruments	Percussion
Mri-T	Mridangam Tonic [Anantapadmanabhan et al., 2013]	Tonic classification in pitched percussion instruments	Percussion

Table 14: Performance on HEAR. Reported values reflect the average performance $\pm$ standard deviation, calculated using n-fold cross-validation as specified by the HEAR pipeline. Bold numbers indicate the best-performing self-supervised model on the specific task. Grayed-out rows indicate supervised models. Higher α denotes a larger share of real recordings during pre-training; α=0.0 uses only simulated data. Tasks are specified in Appendix G. **LS**: LibriSpeech 960 h, **AS**: AudioSet. The mean column is computed from the single-seed per-task values shown.

		Acoustic Events and Scene Analysis				Speech			Music				
Model	**Corpus**	**DCASE**	**FSD50K**	**LC**	**ESC-50**	**CD**	**VL**	**SC-5**	**NS**	**BO**	**Mri-S**	**Mri-T**	**mean**
Wav2Vec 2.0	LS	23.5	29.4	69.9 ± 2.1	46.4 ± 1.8	57.3 ± 1.1	34.9 ± 2.4	85.3	17.4	81.4 ± 4.8	90.7 ± 0.8	77.0 ± 0.9	55.7
HuBERT	LS	78.3	32.8	63.3 ± 1.2	58.6 ± 2.8	71.2 ± 1.2	65.2 ± 2.9	94.0	19.8	93.2 ± 5.9	94.6 ± 0.4	85.0 ± 2.5	68.7
WavLM	LS	27.0	25.7	61.3 ± 2.3	49.5 ± 3.8	64.3 ± 1.3	60.1 ± 3.2	93.8	18.2	84.3 ± 6.3	88.8 ± 1.0	76.8 ± 0.5	59.1
MAE	AS	–	33.4	62.3 ± 1.1	72.9 ± 2.1	60.8 ± 1.8	21.3 ± 5.8	66.6	63.6	94.5 ± 5.6	94.8 ± 0.6	85.1 ± 10.4	65.5
SSAST	AS + LS	–	21.4	57.8 ± 3.3	58.3 ± 2.6	48.0 ± 2.1	15.4 ± 2.6	22.0	64.2	95.8 ± 4.3	90.2 ± 5.9	89.1 ± 8.0	56.2
BEATs	AS	–	54.1	77.8 ± 1.2	85.8 ± 2.9	66.9 ± 2.5	39.7 ± 4.3	86.9	68.6	94.1 ± 3.5	95.5 ± 0.4	96.6 ± 0.5	76.6
MWMAE	AS	94.2	51.8	80.3 ± 1.9	82.2 ± 3.2	74.4 ± 1.5	45.5 ± 1.7	91.6	69.4	95.8 ± 4.3	97.5 ± 0.4	97.6 ± 0.6	80.0
SSAM	AS	87.3	53.5	75.5 ± 1.4	82.9 ± 3.6	70.2 ± 0.4	56.4 ± 5.2	89.3	72.6	93.2 ± 3.5	**97.8** ± 0.5	96.9 ± 0.5	79.6
Dasheng	Mix	94.3	51.2	75.9 ± 2.3	82.6 ± 2.6	**79.6** ± 1.2	**73.1** ± 3.9	**96.6**	71.0	95.3 ± 4.6	96.5 ± 0.6	94.5 ± 0.7	82.8
ATST-Clip	AS	–	**58.5**	79.3 ± 1.8	**92.9** ± 1.8	75.9 ± 0.6	42.1 ± 3.2	93.2	66.8	95.3 ± 3.8	97.5 ± 0.4	96.2 ± 0.7	79.8
ATST-Frame	AS	94.9	55.9	75.1 ± 2.1	87.0 ± 1.3	77.5 ± 1.3	64.6 ± 5.8	91.6	66.4	94.9 ± 3.6	97.3 ± 0.5	96.0 ± 0.5	81.9
PASST	AS	–	64.1	60.7 ± 3.7	**94.8** ± 0.3	61.8 ± 1.1	25.9 ± 2.6	68.7	24.2	96.6 ± 3.2	96.4 ± 0.7	87.8 ± 1.2	68.1
Spatial-AST	AS	–	54.7	72.6 ± 1.5	90.3 ± 1.7	62.2 ± 1.3	29.1 ± 1.9	80.6	69.8	96.2 ± 5.3	96.2 ± 0.4	94.6 ± 0.6	74.6
AudioSphere (α=0.0)	AS	94.3	53.0	79.4 ± 1.5	85.9 ± 1.5	71.9 ± 1.9	53.7 ± 1.2	89.6	73.8	94.9 ± 4.9	97.6 ± 0.5	**98.5** ± 0.4	81.1
AudioSphere (α=0.25)	AS	95.1	54.7	80.8 ± 1.7	87.7 ± 1.0	75.8 ± 1.1	57.3 ± 3.6	94.5	75.8	**96.6** ± 3.8	97.5 ± 0.5	98.4 ± 0.3	83.0
AudioSphere (α=0.5)	AS	94.4	55.9	**81.7** ± 1.6	87.5 ± 2.2	75.1 ± 1.1	57.5 ± 4.9	93.0	75.0	94.9 ± 4.9	97.6 ± 0.3	98.1 ± 0.5	82.8
AudioSphere (α=0.75)	AS	94.8	55.8	80.8 ± 1.6	88.3 ± 2.1	75.9 ± 1.2	58.9 ± 3.9	93.2	**77.8**	94.1 ± 4.8	97.6 ± 0.3	98.4 ± 0.3	**83.2**
AudioSphere-Clean	AS	**95.3**	56.8	81.3 ± 1.8	87.5 ± 2.3	75.1 ± 0.6	57.3 ± 3.4	93.5	75.8	95.8 ± 3.7	97.4 ± 0.3	98.0 ± 0.2	83.1

G.1 Extracting frame and clip level embeddings for downstream tasks

To obtain frame-level embeddings (80 ms resolution), we encode all patches $P_1, \ldots, P_n$ using the frozen AUDIOSPHERE encoder and follow the exact patch aggregation process from [Niizumi et al., 2022]. Patches are stacked and concatenated along the frequency axis, yielding one embedding per 80 ms frame that is suitable for dense prediction tasks. For clip-level embeddings, we split the input audio into non-overlapping 2-second chunks, concatenate the embedded patches of each chunk over time, and then average along the time and frequency axis. This produces a fixed-dimensional clip embedding that is independent of the original audio duration, ideal for scene classification tasks.

Table 15: Baseline summary. Durations are as reported by the respective papers; the AudioSet copy used for AUDIOSPHERE contains ∼1.74 M clips (∼4.8 k nominal hours). HEAR probes use the HEAR shallow MLP protocol [Turian et al., 2022]. SSL = self-supervised.

Model	**Objective / supervision**	**Pre-training corpus (scale)**	**Native input**	**Eval-time conversion**	**Downstream protocol**	**Results**
Wav2Vec 2.0, HuBERT, WavLM (Base)	SSL, masked prediction	LibriSpeech (960 h)	mono waveform	none	HEAR MLP probe	reproduced (HEAR API)
MAE, SSAST, BEATs, MW-MAE, SSAM, ATST	SSL, masked autoencoding / latent prediction	AudioSet (∼4.8–5.8 k h; SSAST also LibriSpeech)	mono spectrogram	none	HEAR MLP probe	reproduced (HEAR API)
Dasheng	SSL, masked autoencoding	Mixed (∼272 k h)	mono spectrogram	none	HEAR MLP probe	reproduced (HEAR API)
PaSST	supervised (AudioSet labels)	AudioSet	mono spectrogram	none	HEAR MLP probe	reproduced (HEAR API)
Spatial-AST (BAT)	supervised (event, DoA, distance labels)	AudioSet + simulated binaural RIRs	binaural	BSM (TAU-2019, RSL2019-dev)	HEAR MLP probe (HEAR, REALSELD, RSL2019-dev)	reproduced
ELSA	contrastive (audio–text); supervised spatial-branch pre-training	Spatially augmented Clotho + AudioCaps + Freesound (∼4.7 k h); Spatial LibriSpeech	FOA	– (model not public)	–	cited (TUT-2018 ov. 1 LE)†
Spatial LibriSpeech	supervised (DoA)	Spatial LibriSpeech (synthetic)	FOA	–	–	cited (TUT-2018 ov. 1 LE)†
w2v-SELD	SSL, wav2vec2-style	L3DAS21 + TUT-2018 + TAU-2019/20/21 (∼200 h)	FOA waveform	none	frozen probe as reported	cited (TAU-2019)
SELDNet, DOAnet, MSEDnet, SEDnet	supervised, per dataset	TUT-2018 Real training split	FOA	none	–	cited
SELDNet (TAU-2019/2020/2021, STARSS23)	supervised, per dataset	each dataset's training split	FOA	ASM for ASM-FOA rows	trained from scratch	cited§; ASM-FOA row retrained by us
MUSIC	signal processing, no training	–	raw planar array	none	–	re-implemented by us
AUDIOSPHERE (ours)	SSL, masked reconstruction of mels + IVs	AudioSet (∼4.8 k h) + simulated Ambisonics RIRs + WHAM!	FOA	ASM (tetrahedral, planar)	HEAR MLP probe; temporal head; fine-tuning	this work

§SELDNet numbers for TAU-2019 are from Adavanne et al. [2019b], for TAU-2020 and STARSS23 from the official DCASE baseline repositories, and for TAU-2021 from Politis et al. [2021]. †REAL subset of TUT-2018.

H RealSELD

We extend the HEAR Benchmark with ACCDOA [Shimada et al., 2021] framework to solve newly introduced real world SELD tasks. This framework jointly models sound event detection and localization across target sound classes. A class is considered active at frame t when the predicted Cartesian coordinate magnitude $\|\mathbf{c}_t\| > 0.5$, where $\mathbf{c}_t \in \mathbb{R}^3$ represents the unit direction vector. We do not perform any post-processing steps unlike the HEAR protocol. All probes are the HEAR shallow MLP classifier [Turian et al., 2022] trained on frozen embeddings; for SELD its output layer is a $3 \times C$ ACCDOA vector (C classes), and the temporal head of Section 4 replaces the MLP by a 256-d projection, frequency pooling, a GRU and multi-head self-attention in front of the same ACCDOA output.

To adapt the datasets to the HEAR format, we processed the ground truth timestamp labels into fixed-duration segments. For static sources, we extracted start and end times alongside fixed azimuth and elevation coordinates. For moving sources, we extracted active segments and their corresponding time-varying directions. In frames where no sources are active, the ACCDOA target is defined as a zero vector. Crucially, the dynamic datasets in RealSELD utilize 100ms segments (10 Hz resolution). To align the output resolution of the audio embedding model with these labels, we applied average

pooling to the model representations. This ensures the AUDIOSPHERE model output is temporally aligned with the 100ms ground truth segments.

After the mapping, we extract the time-stamp embeddings and their corresponding ground truth labels using the HEAR evaluation kit, and train a shallow MLP probe on the extracted embeddings following the HEAR evaluation protocol. The datasets are characterized in Table 1 in the main text.

Variance across downstream seeds We quantify the variance of the downstream protocol by retraining the shallow MLP probe with three seeds on the frozen backbone (Table 16). The same procedure on HEAR gives an aggregate of 83.2 ± 0.1 for AUDIOSPHERE (α=0.75; seeds 83.2, 83.3, 83.1) and 83.0 ± 0.2 for AUDIOSPHERE-Clean (83.1, 82.8, 83.1).

Table 16: REALSELD results of the frozen AUDIOSPHERE backbone with a shallow MLP probe, mean $\pm$ SD over three downstream seeds. Static datasets use the location-independent ER/F and the static LE/LR; dynamic datasets use ER_{20°, F_{20°, LE_{CD}, LR_{CD}(Appendix I).

Dataset	**ER** $\downarrow$	**F** $\uparrow$	**LE**$^\circ$ $\downarrow$	**LR** $\uparrow$
Static scenes (ER, F, LE, LR)				
TUT-2018 Real, ov. 1	0.292 ± 0.007	82.1 ± 0.8	13.41 ± 1.71	83.2 ± 1.0
TUT-2018 Real, ov. 2	0.382 ± 0.002	74.7 ± 0.3	20.80 ± 1.93	48.0 ± 1.7
TUT-2018 Real, ov. 3	0.423 ± 0.006	71.1 ± 0.2	25.31 ± 2.00	26.1 ± 2.0
TAU-2019	0.277 ± 0.054	84.6 ± 2.1	14.05 ± 1.26	78.5 ± 0.6
Dynamic scenes (ER_{20°, F_{20°, LE_{CD}, LR_{CD})				
TAU-2020	0.571 ± 0.019	46.6 ± 3.0	23.54 ± 1.86	63.2 ± 0.6
TAU-2021	0.747 ± 0.007	21.0 ± 0.8	36.89 ± 0.98	49.1 ± 0.4

I Evaluation Metrics

We evaluate the sound localization performance on the newly generated sound localization tasks, and RSL2019-dev by calculating the Direction of Arrival (DoA) error θ between the $[x, y, z]$ coordinates of the target sound source on the unit sphere using the arc cosine of the dot product of the unit vectors: $\theta = \arccos(v \cdot \hat{v})$.

SELD metrics on RealSELD Dynamic Motion: We used four joint localization and detection metrics [Politis et al., 2020b], which are widely-used in audio-only SELD tasks [Politis et al., 2022]. Two metrics, referred to as location-aware detection, are the error rate (ER_{20°) and F-score (F_{20°) for one-second non-overlapping segments. We consider a prediction to be a true positive (TP) if the prediction and the reference class are the same and the angle difference is less than 20°. F_{20° is calculated from location-aware precision and recall, whereas ER_{20° is the sum of insertion, deletion, and substitution errors, divided by the total number of references. The other two metrics, referred to as class-aware localization, are localization error (LE_{CD}) in degrees and localization recall (LR_{CD}) in one-second non-overlapping segments, where the subscript denotes classification-dependent. Unlike location-aware detection, we do not use a threshold; instead, we estimate the difference between the correct prediction and the reference. LE_{CD} expresses the average angular difference between the same class's predictions and references. LR_{CD} reports the true positive rate as the number of localization estimates detected in a class out of the total number of class instances. We used the macro mode of computation, which does not apply to ER_{20° because it includes substitution errors between two classes. We first computed the metrics for each class and then averaged them for the other three metrics to obtain the final system performance.

SELD metrics on RealSELD Static Motion: For testing the model on the TUT-2018 REAL and TAU-2019, we used the non-localization-dependent counterparts of the aforementioned metrics. ER and F-score calculated in segments of one second with no overlap, as proposed. The segment-wise results are obtained from the classifier's frame-level predictions, treating a sound event as active across the entire segment if it is active in any frame. Similarly, we obtain labels for one-second segments of the reference from its framewise annotation and calculate segment-wise ER and F-scores. Additionally, in order to account for time frames where the number of estimated and reference DOAs are unequal, we report the localization recall, calculated as TP/(TP + FN) in percentage, where true

positives TP is the total number of time frames in which the number of DOAs predicted is equal to reference, and false negatives FN is the total number of frames where the predicted and reference DOA are unequal.

J Class-Level Analysis for STARSS23

To better understand the behavior of AUDIOSPHERE with a shallow MLP probe during out-of-domain evaluation, we provide a granular breakdown of its performance across the 13 target classes of the STARSS23 dataset. Table 17 details the location-dependent F-score (F_{20°) and the localization error (LE_{CD}) for the direct transfer setting.

Table 17: Class-wise metrics for STARSS23 Direct Transfer. We report location-dependent F-score (F_{20°, higher is better) and Localization Error (LE_{CD}, lower is better). The DCASE convention assigns maximum error (180°) to undetected classes.

Class	$\boldsymbol{F_{20^\circ}} \uparrow$	$\boldsymbol{LE_{CD}} \downarrow$
Fem. Speech	0.52	25.27
Male Speech	0.63	15.17
Clap	0.20	18.19
Phone	0.01	36.21
Laugh	0.36	17.63
Domestic sounds	0.30	20.54
Footsteps	0.01	37.45
Door	0.00	180.00
Music	0.24	37.12
Music Instr.	0.40	20.91
Faucet	0.28	13.19
Bell	0.00	180.00
Knock	0.13	10.31

Strong Performance on Speech and Harmonic Sounds AUDIOSPHERE demonstrates exceptional direct-transfer capabilities on human vocalizations. "Male Speech" and "Female Speech" are among the highest performing categories, achieving F_{20° scores of 0.63 and 0.52, respectively, alongside strong localization precision. Similarly, continuous and harmonic sounds such as "Laugh" (17.63°) and "Music Instr." (20.91°) yield highly robust localization. This suggests that our spatial-acoustic pre-training excels at learning stable representations of sounds with sustained spectral energy and frequent occurrence across diverse acoustic environments.

Failure Modes and Limitations Conversely, the model struggles significantly with specific environmental transients and quiet events, notably "Door," "Bell," and "Footsteps." For "Door" and "Bell," the model fails to successfully detect and localize the events within the 20° tolerance threshold, resulting in F_{20° scores of 0.00 and a maximum localization error penalty of 180.00°. Footsteps (F_{20° of 0.01) similarly suffer, likely because they are low-energy events heavily susceptible to masking by reverberation and ambient noise in real-world dynamic scenes. These failure cases highlight an ongoing challenge in zero-shot SELD transfer: bridging the gap for rare, highly reverberant, or low-energy impulsive sounds.

Effect of temporal integration We hypothesized that limited temporal integration is a main contributor to the failures on brief, intermittent, low-energy or moving events. Fitting the SELDNet-style temporal head of Section 4 to the frozen backbone (overall STARSS23: F_{20° 37.4, LE_{CD} 22.8°, LR_{CD} 58.8) improves these classes (Table 18) and thereby reduces class-wise variability. This experiment demonstrates that the frozen frame-level embeddings carry spatial-acoustic information that a standard sequence model can integrate over time.

Table 18: STARSS23 failure classes with the shallow MLP probe versus the temporal head on the same frozen backbone.

Class	$F_{20^\circ} \uparrow$ (MLP $\rightarrow$ temporal)	$LE_{CD} \downarrow$ (MLP $\rightarrow$ temporal)
Door	0.00 $\rightarrow$ 0.30	180.0 $\rightarrow$ 11.4
Bell	0.00 $\rightarrow$ 0.17	180.0 $\rightarrow$ 31.7
Footsteps	0.01 $\rightarrow$ 0.12	37.5 $\rightarrow$ 27.1
Phone	0.01 $\rightarrow$ 0.16	36.2 $\rightarrow$ 26.8

K SoundSpaces 2.0 Specifications

We generate our ARIRs using the simulator provided by SoundSpaces 2.0 [Chen et al., 2022]. Our hyperparameters for the simulator is depicted in Table 19.

Table 19: Acoustic configuration parameters utilized in SoundSpaces 2.0 to generate our ARIRs.

Parameter	Value	Parameter	Value
directSHOrder	3	indirectSHOrder	3
sampleRate	32000	frequencyBands	8
maxDiffractionOrder	10	transmission	True
indirect	True	indirectRayCount	15000
indirectRayDepth	400	sourceRayCount	200
sourceRayDepth	20	threadCount	16
agentHeight	1.5m		

L Detailed Pre-Training HyperParameters

Table 20 summarizes all hyperparameters. We train AUDIOSPHERE for 500K steps on a single H100 96 GB GPU with 16 CPU cores, minimizing the mean squared error (MSE) between predicted and ground-truth patches at masked positions. We use AdamW [Loshchilov and Hutter, 2017] with weight decay 0.01, gradient clipping at max norm 5.0, and a cosine learning rate schedule with 10K-step warm-up from 2×10^{-4} decaying to 0. Training converges in approximately 90 GPU hours.

Table 20: Pre-training hyperparameters for AUDIOSPHERE.

Parameter	Value
Encoder	ViT-Base (12 layers, 768 dim, 12 heads)
Decoder	8 layers, local-global attention
Patch size	$7 \times 8 \times 16$ (channels $\times$ time $\times$ freq)
Embedding dim	768
Mask ratio	80% (tube masking)
Segment length	2 s (200 frames)
Segments per clip	16
Nominal / effective batch size	96 / 1,536
Optimizer	AdamW (β_1=0.9, β_2=0.95)
Learning rate	2×10^{-4} (cosine decay to 0)
Warm-up steps	10,000
Weight decay	0.01
Gradient clipping	max norm 5.0
Training steps	500K

M Confusion Matrices for BAT and AUDIOSPHERE on RSL2019-dev

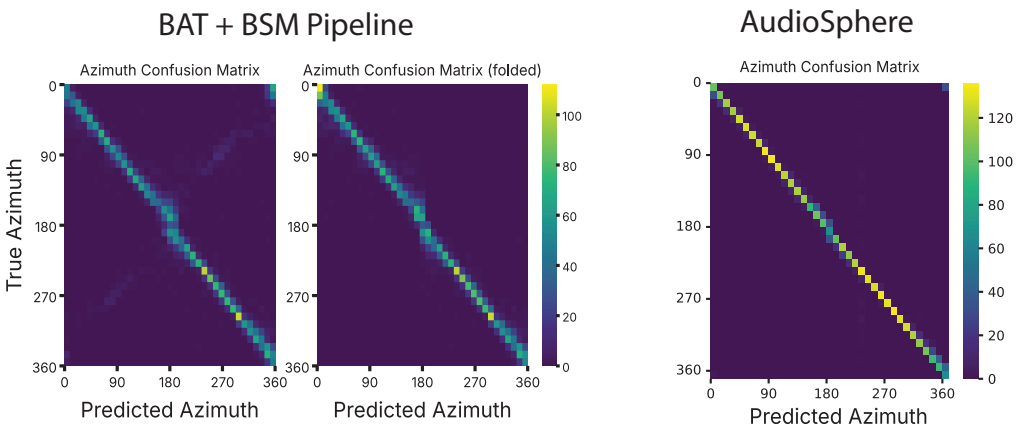

Figure 4: **Confusion Matrices for BAT + BSM Pipeline and AUDIOSPHERE.** Here, we depict the azimuth confusion matrices for BAT and AUDIOSPHERE on RSL2019-dev dataset. BAT + BSM's errors concentrate on front–back mirrored directions; see Appendix D.2 for the folded analysis.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our main claims are: (i) AUDIOSPHERE is a self-supervised spatial-acoustic representation model whose frozen embeddings transfer, with a shallow MLP probe, to clip-level classification of environmental sound, speech and music and to frame-level sound event detection (HEAR, Section 5.1), and to real-world sound event localization and detection (RealSELD, Section 5.2). (ii) RealSELD, an evaluation kit that adapts five real-world SELD datasets (seven tasks) to a frozen-backbone protocol (Section 4, Appendix H). (iii) Microphone flexibility via ASM, supported by the ASM literature and by our experiments on tetrahedral arrays (TAU-2019 in Section 5.3; TAU-2020, TAU-2021 and STARSS23 in Appendix D) and on a planar array (RSL2019-dev). The comparison with Spatial-AST on RSL2019-dev is stated at the level of complete evaluated pipelines (Section 5.3, Appendix D), and the scope of the spatial evaluation (detection and direction-of-arrival estimation only) is stated in Section 6.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In Section 6 we discuss the limitations of our work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A].

   Justification: Our paper does not derive new theoretical results.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We describe the data-generation pipeline, model and training objective in Section 3 and the evaluation protocols in Section 4, with full details in the appendices: ASM encoding (B), the HEAR tasks, embedding extraction and the protocol of every baseline (G), the construction of RealSELD and the downstream heads (H), evaluation metrics (I), simulator settings (K) and pre-training hyperparameters (L).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code, pre-trained checkpoints, the RealSELD evaluation kit (scripts and dataset-split files) and the preprocessing pipeline are released on GitHub and HuggingFace (footnote 1).

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Pre-training data generation is described in Section 3 and Appendix K, and the optimizer and all training hyperparameters in Appendix L. The downstream protocols (shallow MLP probing following HEAR, the temporal head, fine-tuning and its learning-rate sweep) are given in Section 4 and Appendix H, and the RSL2019-dev splits in Section 4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Error bars capture two sources of variability: (i) n-fold cross-validation as specified by the HEAR pipeline, mean $\pm$ standard deviation (per-task HEAR results, Appendix G); and (ii) variability over three downstream seeds of the probe trained on the frozen backbone, mean $\pm$ one standard deviation (HEAR aggregate in Section 5.1; RealSELD static and dynamic results in Section 5.2 and Appendix H).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix L reports the hardware (a single H100 96 GB GPU with 16 CPU cores) and the pre-training cost ($\sim$90 GPU hours for 500k steps). Ablation models were pre-trained for 50k steps (Appendix F) or 125k steps (Section 5.4), and downstream probes were repeated over three seeds (Appendix H).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

Justification: Our paper conforms with the ethics.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: Our work addresses spatial audio representation learning, and does not possess negative societal impacts.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: Our work addresses spatial audio representation learning, and does not possess negative societal impacts.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cited each dataset and the model code we've used. Furthermore, our released datasets and benchmarks include all necessary licenses.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The AUDIOSPHERE checkpoints and pre-training data are released on HuggingFace and the RealSELD evaluation kit on GitHub, each documented alongside the asset.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: We did not perform crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: We did not use LLMs to develop methodological parts of this research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.