STRUCTURING SPATIAL AUDIO REPRESENTATIONS VIA INVARIANCE-EQUIVARIANCE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

A spatial audio representation must jointly capture scene semantics to recognize *what* is sounding and scene geometry to understand *where* sound events occur. Self-supervised objectives often favor semantic invariance, treating direction as a nuisance, making spatial geometry difficult to preserve. We propose a framework for structuring spatial audio representations through invariance–equivariance learning, using geometric transformations as a training signal. In first-order ambisonics (FOA), rotating a sound field is a matrix operation, providing rotated views and their relative rotations without labels. We train encoders to remain stable under spectral nuisances while predicting how their representations change under rotation. Within this framework, we compare representations explicitly partitioned into invariant and equivariant subspaces with an unsplit predictive latent. The latter, built on seq-JEPA, is to our knowledge the first joint-embedding predictive architecture (JEPA) for spatial audio that learns invariance and equivariance jointly. We show that this latent representation organizes direction as an ordered spherical map while preserving linearly separable sound classes without an architectural split. When frozen, it achieves a lower SELD score on real recordings than supervised baselines and improves spatial audio captioning and retrieval, with the largest gains in spatial accuracy. These results demonstrate a path toward robust, general-purpose spatial audio representations that jointly encode what is present and where it occurs in space.

1 INTRODUCTION

Hearing is spatial: it tells us not only *what* is sounding but also *where*, conveying the location of sources around a listener, including those outside the field of view (Blauert, 1996). Machines that perceive and act in the physical world need the same ability, from robots that navigate toward a sound (Chen et al., 2022) to assistive and AR/VR devices that place sounds around users (Engel et al., 2023). Supporting these technologies requires spatial audio encoders that recognize and localize sounds robustly across the wide variety of real-world events and acoustic environments.

Building such encoders is an active stream of work. Sound direction of arrival (DoA) is achieved by processing subtle phase and magnitude differences between the channels of a multichannel recording, cues that reverberation and noise distort. Classical array processing estimates direction from these cues (Krim & Viberg, 1996), but it does not recognize what is sounding and degrades in real, noisy environments. Sound event localization and detection (SELD) addresses both at once, training deep networks to detect sound events and estimate their DoA (Adavanne et al., 2019; Politis et al., 2020). Most SELD systems, however, are supervised (Cao et al., 2021; Hu et al., 2024), and spatial annotations are costly, so labelled data remain scarce and cover a limited set of environments and sound classes (Shimada et al., 2023; Krause & Politis, 2024), limiting generalization. Self-supervised learning (SSL) removes the need for labels and has produced strong general-purpose audio encoders (Huang et al., 2022; Chen et al., 2023; Niizumi et al., 2021; Fei et al., 2023), yet most are trained on monaural audio, capturing *what* a sound is but discarding *where* it comes from.

The few SSL encoders that accept multichannel input follow two dominant paradigms. GRAM (Yuksel et al., 2025b), a multichannel masked autoencoder, reconstructs spatialized scenes, outperforming monaural foundation models on spatial benchmarks and localizing sources in simulated and real recordings. Contrastive approaches align paired views or modalities (Wang et al.,

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

2022). Neither objective organizes the representation around scene geometry: reconstruction retains whatever is needed to rebuild the input, leaving spatial cues entangled with other information, while contrastive learning discards whatever differs between paired views. The SARL probing study (Chen et al., 2026) finds that sound classes are consistently decodable across supervised and self-supervised encoders, whereas localization is much less consistently encoded. Thus, neither multichannel input nor reconstruction or contrastive objectives alone guarantees a representation that is structured around spatial geometry.

What is missing is an objective that specifies how a representation should change when the spatial scene changes. Rotation provides such a signal. In standard spatial audio formats, rotating the sound field is an exact operation on the audio channels (Zotter & Frank, 2019), so any recording can be rotated by a known amount, without labels. We hypothesize that *learning rotation equivariance as a pretext induces a representation of scene geometry that encodes source localization*: to predict how its latent changes under arbitrary rotations, a network is pushed to represent where the sources are. Rotation equivariance alone, however, does not ask the representation to capture what is sounding. Invariance is a well-established principle for learning semantics (Misra & van der Maaten, 2020), but invariance to rotation would discard direction (Garrido et al., 2023). We therefore combine the two through invariance–equivariance learning: the representation is invariant to spectral transformations such as noise and gain, and equivariant to rotations. No prior work has examined this hypothesis empirically. We investigate it through three contributions:

- An invariance–equivariance framework for spatial audio that uses sound-field rotations as a label-free signal for learning source localization, introducing spatial audio as a new domain for studying invariance–equivariance learning.
- A comparison of three learning methods: two that split the latent into invariant and equivariant subspaces, and a predictive one that aggregates rotated views into a single latent.
- A systematic study of how these encoders represent semantic and spatial structure, measured by linear probing, and how they use it for joint detection-localization, and spatial audio-language understanding.

2 Related Work

Spatial audio representations have been learned mainly for sound event localization and detection (SELD), the joint task of recognizing active sound events and estimating their directions of arrival (Politis et al., 2020). Supervised SELD models have steadily improved, from convolutional-recurrent networks (Adavanne et al., 2019) to track-wise formulations (Cao et al., 2021) and pre-training on large synthetic corpora (Hu et al., 2024), and their encoders now also supply the spatial features of audio–language models (Zheng et al., 2024; Seki et al., 2025). Some SELD systems also rotate ambisonic recordings, together with their direction labels, as data augmentation (Mazzon et al., 2019; Wang et al., 2023). Self-supervised encoders avoid labels by learning from audio-visual spatial alignment (Morgado et al., 2020; Wang et al., 2022), masked reconstruction (Yuksel et al., 2025b), masked prediction (Dimitriadis et al., 2023) or latent prediction (Yuksel et al., 2025a), although the latter encodes space less reliably than reconstruction (Chen et al., 2026). Across these works, rotation serves at most as an augmentation of the input and no objective specifies how the representation itself should change when the scene geometry rotates.

Specifying how the representation should change when the scene geometry rotates is the role of equivariance, which can first be built into the architecture. Group-equivariant and spherical CNNs (Cohen & Welling, 2016; Cohen et al., 2018; Esteves et al., 2018) and networks for volumetric data, point clouds and molecules (Weiler et al., 2018; Thomas et al., 2018; Fuchs et al., 2020) guarantee that features transform with rotations of the input, and the same idea has recently reached ambisonic SELD (Sato et al., 2021; Yuksel et al., 2026). Such models fix the group representation in the architecture, and in audio they remain supervised. The alternative is to learn equivariance from data, leaving the backbone unconstrained. Following transforming autoencoders (Hinton et al., 2011), self-supervised methods first predicted the transformation applied to an input, such as an image rotation (Gidaris et al., 2018; Lee et al., 2021; Dangovski et al., 2022), and later its effect in latent space via a conditioned predictor (Park et al., 2022). Paired with an invariance objective to preserve semantics, these methods differ in where each property is represented: SIE splits the latent into invariant and equivariant subspaces (Garrido et al., 2023), EquiMod derives both from one

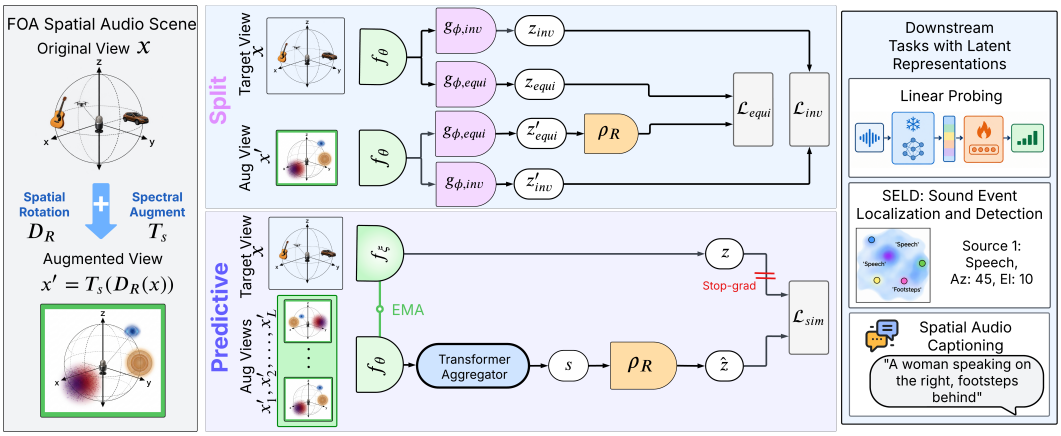


Figure 1: SIE and SimCLR-EquiMod (top): the representation is *split* into an invariant part, trained to match across views, and an equivariant part, trained to match after the latent action ρ_R. seq-JEPA (bottom): L rotated views are embedded and pooled by a Transformer into a single latent s, from which a predictor conditioned on the rotation regresses an EMA target.

backbone output via separate projection heads (Devillers & Lefort, 2023), and seq-JEPA uses no split, aggregating a sequence of views and their actions to predict a held-out view (Ghaemi et al., 2025).

These methods have been studied on image augmentations and rendered objects (Garrido et al., 2023), and in audio only for one-dimensional transformations such as pitch and tempo (Gfeller et al., 2020; Riou et al., 2023; Guinot et al., 2024). There, it is hard to tell whether learned structure comes from the objective or from the transformations and data, a question that also runs through predictive architectures, where latent prediction alone can induce structure (Assran et al., 2023; Farid et al., 2025) yet recent work adds it by design (Jha et al., 2026; Maes et al., 2026). Spatial audio is well suited to this question, because a rotation of the sound field is exact for any recording, real or simulated. We therefore study invariance–equivariance learning under 3D rotations of real acoustic scenes, comparing split and predictive designs under identical conditions.

3 Approach

3.1 Invariance–Equivariance Framework

Let x be a view of an acoustic scene and f_θ an encoder that maps it to a latent $z = f_\theta(x)$. A transformed view is $x' = T_s(x)$, where T_s is a spectral transformation (noise, gain and time–frequency masking), altering the signal but not the scene. We train the representation to be *invariant* to them:

$$f_\theta\big(T_s(x)\big) = f_\theta(x). \tag{1}$$

The rotation of the sound field is a special transformation, $x' = D_R(x)$ with $R \in \mathrm{SO}(3)$, which changes the geometry of the scene. We train the representation to be *equivariant* to it, so that the latent moves under a latent action ρ_R conditioned on the rotation:

$$f_\theta\big(D_R(x)\big) = \rho_R\big(f_\theta(x)\big). \tag{2}$$

The action ρ_R is learned and need not resemble D_R. Invariance is the special case $\rho_R = \mathbf{I}$, which maps every rotated version of a scene to the same latent and leaves the rotation unreadable. The objectives of Section 3.4 impose both properties on a single encoder, differing in how the latent is organized and how collapse is avoided.

3.2 Transformations for Spatial Audio

Ambisonics represents a three-dimensional sound field independently of the recording array by projecting it onto spherical harmonics up to an order ℓ (Zotter & Frank, 2019). First-order ambisonics

(FOA), the $\ell = 1$ encoding, has four channels: an omnidirectional channel W and three figure-of-eight dipoles X, Y, Z aligned with the Cartesian axes. It is an industry-standard format, exchanged in the AmbiX convention with channels in ACN order (W, Y, Z, X) (Nachbar et al., 2011). Because the dipoles encode each source direction in Cartesian coordinates, rotating the sound field rotates them with the scene and leaves W unchanged. For an FOA recording $x = [x_1, \dots, x_T] \in \mathbb{R}^{4\times T}$, where x_t holds the four channels at time t, a rotation $R \in \mathrm{SO}(3)$ therefore acts on each sample through the Wigner-D matrix

$$x'_t = D_R\, x_t, \qquad D_R = \begin{pmatrix} 1 & \mathbf{0}^\top \\ \mathbf{0} & MRM^\top \end{pmatrix}, \tag{3}$$

where M permutes Cartesian (X, Y, Z) into ACN order and the zero blocks mean that no energy moves between orders. Rotating a recording is thus an exact, invertible 4×4 multiply per sample, applied to the raw waveform of any real or simulated recording without labels. Eq. 3 supplies in closed form the input action D_R of Eq. 2.

Views. Every view is therefore defined by: a rotation R, applied as D_R to the raw FOA waveform before feature extraction; a spectral transformation T_s (Gaussian noise, gain scaling, time–frequency masking). Because D_R is analytic, views are synthesized on the fly at negligible cost and their relative rotations are known, so this supervision is label free.

3.3 Shared Encoder

All three methods use the same encoder f_θ, so that differences in what they learn can be attributed to the objective rather than to the backbone. We adopt the Conformer of PSELDnets (Hu et al., 2024), a CNN8 front-end followed by Conformer blocks (Gulati et al., 2020). Its input is a 5 s FOA clip at 24 kHz, represented as seven stacked time–frequency channels: the log-mel spectrograms of the four FOA channels and three acoustic intensity vectors, formed from the omnidirectional channel and each dipole, which point toward the dominant source direction in each bin (Appendix A.2). Its sound event detection (SED) and direction-of-arrival (DoA) streams are pooled into a per-clip embedding and linearly projected to a 512-dimensional latent.

3.4 Invariance–Equivariance Objectives

We implement the framework with two families of architectures that differ in how they obtain invariant and equivariant representations and in how they avoid collapse (Appendix A.1). In both, the rotation R relating two views enters as a token $q \in \mathbb{R}^4$, the unit quaternion of R; roll is fixed, so R varies only in azimuth and elevation.

3.4.1 Split Invariance–Equivariance Objectives

SIE (Garrido et al., 2023) and SimCLR-EquiMod (Chen et al., 2020; Devillers & Lefort, 2023) use two projection heads, $g_{\phi,\mathrm{inv}}$ and $g_{\phi,\mathrm{equi}}$, to map each view to an invariant latent z_{inv} and an equivariant latent z_{equi} (Figure 1, top). For a target view x and an augmented view x', an invariance loss $\mathcal{L}_{\mathrm{inv}} = d(z_{\mathrm{inv}}, z'_{\mathrm{inv}})$ makes the invariant latents discard rotation and spectral augmentations. An equivariance loss $\mathcal{L}_{\mathrm{equi}} = d\big(z_{\mathrm{equi}}, \rho_R(z'_{\mathrm{equi}})\big)$ asks the predictor ρ_R, a hypernetwork that generates a latent linear map from q, to undo the rotation, so the equivariant latent retains orientation. SIE splits the 512-dimensional encoder output into two halves, one per head, uses the squared ℓ_2 distance for d, and prevents collapse with a VICReg variance–covariance regularizer (Bardes et al., 2022). SimCLR-EquiMod feeds the 512-dimensional output to both heads, so the separation comes from the losses alone, and uses an InfoNCE loss for d, whose in-batch negatives prevent collapse.

3.4.2 Predictive Invariance–Equivariance Objectives

seq-JEPA (Ghaemi et al., 2025) obtains invariance differently: it aggregates a set of views into a common representation (Figure 1, bottom). The encoder f_θ embeds L augmented views $x'_1, \dots, x'_L$ of a clip, each embedding is concatenated with the token q_i of the relative rotation to the next view, the last token leading to a held-out target view x, and a Transformer aggregator combines the set into a single latent s. Because s already aggregates spectral augmentations and rotations, the predictor ρ_R is a linear map that predicts a 512-dimensional target latent, $\hat{z} = \rho_R(s)$. The target z comes from

f_ξ, an exponential moving average (EMA) of f_θ, and $\mathcal{L}_{\text{sim}}$ is the negative cosine similarity between $\hat{z}$ and z, with a stop-gradient on z; this asymmetry alone prevents collapse. Neither invariance nor equivariance is enforced by a loss; both emerge from the architecture (Ghaemi et al., 2025). For downstream evaluation, we use the view embedding $\hat{z}$.

4 Experimental Setup

We evaluate the representations on three tasks: **Linear probing** measures what semantic and spatial information each representation makes available; **Sound Event Localization and Detection (SELD)** tests the detection and localization of concurrent sources in realistic spatial scenes; and **Spatial captioning and retrieval** tests audio language understanding from the representations.

4.1 SSL Pretraining

Data. Pretraining uses on-the-fly synthetic mixtures, dry source clips drawn from a 34-class sound-event vocabulary (a curated subset of the FSD50K ontology, Fonseca et al., 2022) are convolved with room impulse responses (RIRs) generated with AudibleLight (Cheston et al., 2025) on Gibson meshes (Xia et al., 2018), sampling source azimuth in $\pm 180°$, elevation in $\pm 60°$, and distance in $[0.5, 2.5]\,\mathrm{m}$. To expose the encoder to room acoustics and event diversity, we add unlabelled real FOA recordings from Sphere360 (Liu et al., 2025), retaining roughly 100 hours after curation. Full generation and curation details are in Appendix A.2.

Curriculum. Polyphony and real data are introduced in stages rather than from the start: single-source synthetic mixtures until epoch 150, up to two simultaneous sources from epoch 150 and three from epoch 200, with Sphere360 mixed in from epoch 250 at a fixed 25% of each batch. This lets the encoder learn invariant and equivariant structure from clean single-source scenes before facing polyphony and real recordings.

Optimization and model selection. All methods share one recipe: AdamW with a 10-epoch linear warmup and cosine decay. Pretraining is fully self-supervised, but pretext objectives can collapse or diverge, so we select models with an auxiliary supervised score, following the online evaluator of SIE (Garrido et al., 2023). A linear probe is trained online on the frozen representation, with no gradient reaching f_θ, and scored on a held-out synthetic validation set whose class and DoA labels come free from the simulator. Its composite score (0.5 classification $+$ 0.25 rotation $+$ 0.25 DoA) drives a per-method Bayesian search over loss weights, early stopping, and the choice of backbone. Additionally, during training we use representation statistics to detect collapse (Appendix A.3).

4.2 Evaluation Tasks

Linear probing test whether semantic and spatial information is linearly decodable. Following SARL (Chen et al., 2026), we freeze the encoder and fit small heads for three tasks on held-out single-source mixtures whose RIRs and sound assets are disjoint from pretraining: sound-event classification (34 classes), relative rotation estimation between two augmented views, and DoA estimation. Rotation and DoA are cast as classification over discretized $10°$ azimuth/elevation bins; the classification and DoA probes are single linear layers, while the rotation probe is a two-layer MLP on the concatenation of both views' embeddings. Backbone mean pooling collapses each clip to a single vector before probing; for split methods we probe the invariant part (Inv.), the equivariant part (Equi.), their concatenation (All), while other methods have All only. We report Top-1 accuracy and mean angular error (MAE, degrees), decoded by arg-max from Gaussian-smoothed ($\sigma = 2°$) bin predictions with azimuth error under circular wraparound. Reference models include: the same Conformer backbone trained end-to-end (E2E) on each task as top-line; two supervised encoders (Sup): SELDnet (Adavanne et al., 2019), EINV2 (Cao et al., 2021); and two self-supervised encoders (SSL): GRAM-T (Yuksel et al., 2025b) and AVSA (Wang et al., 2022).

Sound Event Localization and Detection (SELD) tests whether representations support joint event detection and localization in dynamic acoustic scenes. We pair each encoder with a multi-ACCDOA decoder trained with ADPIT (Shimada et al., 2022), under two protocols: frozen encoder (FE) and fine-tuned (FT). Because the decoder needs frame-level rather than clip-level features, pooling is skipped: the time axis is unwrapped into the batch dimension, every frame passes through the same

Table 1: Linear probing of learned representations on the held-out single-source synthetic set. Top-1 classification accuracy (%), Rotation MAE (°), Azimuth MAE (°), and Elevation MAE (°). Methods without Inv./Equi split are shown as **All**. Within each task, the best result is in **bold** and the second best is underlined; the end-to-end Conformer is a top line and is excluded from this ranking.

	Method	CLS (Top-1) ↑			Rot (MAE) ↓			DoA Azi (MAE) ↓			DoA Ele (MAE) ↓		
		All	**Inv.**	**Equi.**	**All**	**Inv.**	**Equi.**	**All**	**Inv.**	**Equi.**	**All**	**Inv.**	**Equi.**
Ours	SIE	59.1	58.3	30.0	6.4	60.7	**5.7**	38.3	88.8	38.3	24.1	48.6	24.9
	SimCLR-EquiMod	67.0	57.3	50.2	7.4	50.4	12.6	<u>23.7</u>	44.7	23.88	**12.0**	28.0	21.0
	seq-JEPA	67.7			<u>6.6</u>			**23.6**			**12.0**		
SSL	AVSA	49.2			57.29			89.83			47.7		
	GRAM-T	**82.4**			8.92			30.2			<u>15.7</u>		
Sup	SELDnet	25.5			56.67			88.15			45.23		
	EINV2	<u>79.30</u>			19.75			40.89			28.66		
E2E	Conformer	68.8			11.9			17.3			8.6		

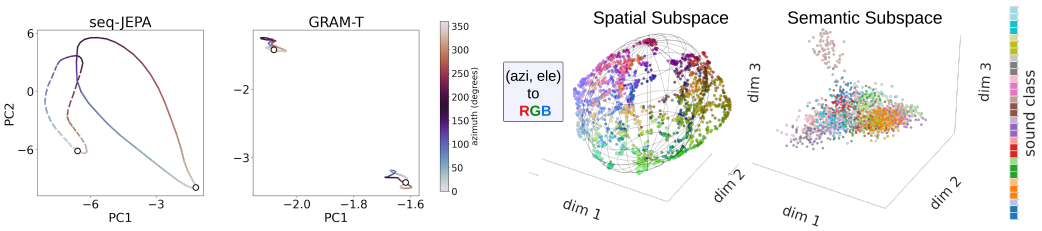
frozen projector stack, and the result is re-wrapped to $[B, T, \dim]$. Training and validation use disjoint room partitions of the synthetic DCASE-Synth dataset (Krause & Politis, 2024). Evaluation spans three sets on diverse environments: the full STARSS23 dataset (Shimada et al., 2023), real recordings and rooms (168 clips, 7.4 hours); ScaperSynth dataset genearted using SpatialScaper (Roman et al., 2024) with 11 unseen room environments using STARSS23 vocabulary (110 clips, 1.8 hours); and the TAU-NIGENS dev-test split (Politis et al., 2021) (100 clips, 1.4 hours), a different dataset and recording setup scored only on the 6 classes its taxonomy shares with STARSS23. We report the standard joint metrics at a 20° threshold (Mesaros et al., 2019; Politis et al., 2020): location-sensitive error rate (ER) and F-score, class-sensitive localization error (LE, degrees) and localization recall (LR), and the aggregate SELD score $= \frac{1}{4}(\mathrm{ER} + (1 - F) + \mathrm{LE}/180 + (1 - \mathrm{LR}))$. ER and F-score enter only via the SELD score. We compare to a supervised Conformer, SELDnet (Adavanne et al., 2019) and EINV2 (Cao et al., 2021) trained end-to-end on DCASE-Synth.

Spatial captioning and retrieval asks whether the representation carries enough spatial detail for language to be grounded in it. Following Spatial-CLAP (Seki et al., 2025), we convolve the monaural assets of AudioCaps2.0 (Kim et al., 2019) (91,256 training, 2,475 validation, 975 test) with our RIRs, discretize source position into four 90° quadrants ('front', 'left', 'right', 'behind'), and build captions with the template '[source caption] + [location]'. The audio encoder is kept fully frozen for both retrieval and captioning. For captioning we train only a lightweight GPT-2 decoder, so that a compact head cannot compensate for missing spatial information with text priors. Two metric families behave differently here and we separate them in advance: lexical and semantic agreement with the reference caption (BLEU, ROUGE-L, METEOR, CIDEr, SPICE, SPIDEr, BERTScore, SBERT) rewards fluent captions whether or not their spatial claim is correct, whereas direction-weighted SBERT (DW-SBERT) and spatial accuracy score that claim directly. Retrieval ranks by audio–text cosine similarity and is scored by recall at ranks 1 and 5, both text-to-audio and audio-to-text. Directional accuracy instead checks the fraction of clips whose audio embedding is closest to the text of the correct quadrant. We compare against Spatial-CLAP (S-CLAP). We keep its LAION-CLAP text encoder (Wu et al., 2023) and decoder fixed and replace only its EINV2 spatial encoder, so any difference is attributable to the representation. For captioning, we use an EINV2 checkpoint pretrained on AudioSet (matching the AudioCaps taxonomy) from PSELDnets (Hu et al., 2024).

5 Results

5.1 Linear Probing

Table 1 shows that invariance–equivariance learning makes both semantic and spatial information linearly recoverable from a single frozen representation, and all three methods achieve it. Every supervised or SSL method fails on the spatial tasks except GRAM-T, which captures semantics at their expense. Our three proposed methods differ only in how they arrange invariance and equivariance. The split methods specialize as designed, but only partially. SIE separates cleanly: its invariant half



(a) Latent space equivariance quality (b) seq-JEPA: spatial and semantic analysis

Figure 2: **(a)** PCA projection of the latents of two held-out clips, each swept through a full 360° azimuth rotation and colored by azimuth; circles mark the starting point (0°). Each model has its own projection, so compare the size of each trajectory with the distance between the two clips rather than axis values across panels. seq-JEPA traces a closed loop per clip; GRAM-T keeps each clip in a small, jagged cluster far from the other. **(b)** seq-JEPA's latent decomposed linearly. Left: UMAP of the DoA-predictive subspace, colored by source direction (azi/ele mapped to RGB). Right: LDA projection of the orthogonal residual, colored by sound class.

outperforms its equivariant half on classification (58.3% vs. 30.0%), while its equivariant half outperforms the invariant half on every spatial task. SimCLR-EquiMod leaks between splits, shown by its equivariant head still reaching 50.2% classification accuracy. For both SimCLR-EquiMod and SIE, All outperforms either sub-latent on the spatial tasks, indicating incomplete disentanglement of spatial information. SIE's equivariant latent does attain the lowest rotation error, but this is the pretext task on which that subspace is trained directly and does not transfer to DoA, where SIE underperforms the other methods. This suggests that the apparent disentanglement is task-specific rather than a general separation of spatial information. The unsplit latent bears this out, with seq-JEPA leading the three methods in classification and azimuth, tying SimCLR-EquiMod in elevation, and coming within 0.2° of SIE in rotation. Explicit partitioning thus seems unnecessary for both.

Among self-supervised encoders, AVSA, trained by audio-visual spatial alignment, underperforms across the board; its azimuth error (89.8°) is close to what a uniform guess over the full circle would give. GRAM-T is the more informative comparison, highlighting the tradeoff between semantic and spatial encoding. Its masked-reconstruction pretraining gives it the best classification accuracy (82.4%), but it is worse than seq-JEPA on spatial tasks. Reconstruction rewards keeping whatever is needed to rebuild the input, but nothing in the objective asks the latent to change coherently with changes of the scene geometry. Learning rotation equivariance therefore favors spatial precision.

On the other hand, supervised SELD encoders show that task supervision does not by itself yield a spatially structured representation. EINV2 has the second-best classification accuracy (79.3%) but transfers poorly to rotation and DoA, and is worse than all three of our encoders on every spatial task despite being trained with DoA labels. SELDnet is weak on every task, due to its small architecture, which is prone to overfitting to training conditions rather than learning transferable spatial features.

The task-specific Conformer top line leads on classification and absolute DoA, as expected: it is optimized end-to-end for each target, while the SSL representations are frozen and read out by a linear probe. It sets the reference for absolute DoA (17.3° azimuth, 8.6° elevation). Even so, on classification, seq-JEPA (67.7) and SimCLR-EquiMod (67.0) come within 1.8 points of it and all three invariance–equivariance encoders estimate relative rotation more accurately.

Table 1 shows what can be decoded out of each latent but not how that information is organized. We complement it with a qualitative look at seq-JEPA, which learns from rotation, and GRAM-T, which does not (Figure 2). First, we sweep two held-out clips through 360° of azimuth and project each model's latents with PCA (Figure 2–(a)). seq-JEPA traces a smooth loop for each clip, around azimuth, that closes when the rotation returns to its start, as an equivariant representation should. GRAM-T shows the opposite: each clip stays in a small, jagged cluster, far smaller than the distance between clips, so its latent separates clips by semantics and barely registers the rotation, consistent with linear probing. We further decompose seq-JEPA's latent by linear regression into a DoA-predictive subspace and its orthogonal residual (Figure 2–(b)). Under UMAP the spatial subspace varies smoothly with source direction, so DoA is encoded as an ordered map. The residual, with the

Table 2: Real-data SELD on STARSS23. Our encoders (blue) are frozen (FE) or fine-tuned (FT); baselines are trained end-to-end (E2E). Best in **bold**, second underlined; 3 seeds.

Method	Backbone	Protocol	SELD↓	LE (°)↓	LR (%)↑
SIE	Conformer	FE	0.611 ± 0.004	30.561 ± 1.191	0.363 ± 0.007
SIE	Conformer	FT	0.564 ± 0.010	26.352 ± 1.245	0.454 ± 0.016
seq-JEPA	Conformer	FE	0.570 ± 0.003	24.182 ± 0.587	0.392 ± 0.008
seq-JEPA	Conformer	FT	**0.540 ± 0.015**	**23.573 ± 0.818**	0.443 ± 0.027
SimCLR-EquiMod	Conformer	FE	0.626 ± 0.007	29.500 ± 0.550	0.332 ± 0.005
SimCLR-EquiMod	Conformer	FT	0.601 ± 0.018	30.206 ± 2.449	0.411 ± 0.048
Supervised	SELDnet	E2E	0.695 ± 0.021	36.015 ± 0.476	0.249 ± 0.059
Supervised	EINV2	E2E	0.578 ± 0.005	25.593 ± 1.227	**0.464 ± 0.017**
Supervised	Conformer	E2E	0.618 ± 0.015	30.708 ± 1.566	0.443 ± 0.038

Table 3: Cross-dataset SELD evaluation on ScaperSynth and TAU-NIGENS with the models of Table 2, without retraining. Best in **bold**, second underlined; 3 seeds.

			ScaperSynth		TAU-NIGENS	
Method	Backbone	Protocol	LE (°)↓	LR (%)↑	LE (°)↓	LR (%)↑
SIE	Conformer	FE	22.549 ± 0.549	0.528 ± 0.008	27.957 ± 1.022	0.358 ± 0.017
SIE	Conformer	FT	20.399 ± 0.870	**0.599 ± 0.016**	28.019 ± 0.540	0.455 ± 0.042
seq-JEPA	Conformer	FE	22.704 ± 0.486	0.494 ± 0.025	26.598 ± 0.866	0.331 ± 0.014
seq-JEPA	Conformer	FT	**19.797 ± 0.568**	0.585 ± 0.036	**25.337 ± 1.429**	**0.467 ± 0.052**
SimCLR-EquiMod	Conformer	FE	22.721 ± 0.437	0.506 ± 0.016	28.032 ± 0.947	0.315 ± 0.018
SimCLR-EquiMod	Conformer	FT	21.294 ± 0.481	0.525 ± 0.026	28.068 ± 2.029	0.387 ± 0.036
Supervised	SELDnet	E2E	33.207 ± 4.005	0.388 ± 0.061	35.981 ± 1.665	0.262 ± 0.062
Supervised	EINV2	E2E	20.567 ± 1.190	0.543 ± 0.014	26.436 ± 1.271	0.434 ± 0.027
Supervised	Conformer	E2E	20.663 ± 0.368	0.549 ± 0.025	30.832 ± 1.725	0.415 ± 0.030

linear spatial component removed, still groups sounds by class under LDA, although several classes overlap. Appendix A.5 shows this decomposition at a larger size, next to GRAM-T's for comparison. The separation a split architecture provides can therefore be recovered from seq-JEPA's single latent after training; this is the representation the decoders in Sections 5.2 and 5.3 build on.

5.2 Sound Event Localization and Detection

SELD adds what linear probing lacks: overlapping, moving sources, real reverberation and unseen rooms/environments (STARSS23 in Table 2; ScaperSynth and TAU-NIGENS in Table 3).

Self-supervised pretraining outperforms task supervision. Fine-tuned, all three invariance–equivariance encoders achieve a better SELD score than the supervised SELDnet and Conformer, and seq-JEPA surpasses every supervised system even when frozen (0.570 against 0.578 for EINV2). The supervised models learn only from DCASE-Synth, whereas our encoders were pretrained without labels across hundreds of simulated rooms and real recordings. Fine-tuning improves every metric except SimCLR-EquiMod's LE (29.5° to 30.2°). The invariance–equivariance encoders are strongest on localization error, with seq-JEPA best under both protocols, but they do not lead on localization recall (LR). LR counts the ground-truth events recovered with the correct class, so it depends on recognizing what is sounding rather than on where it is. This is where fine-tuning matters: it barely changes seq-JEPA's LE (24.2° to 23.6°) but raises its LR from 0.392 to 0.443, and SIE and SimCLR-EquiMod follow the same pattern. EINV2, the only competitive supervised system, leads LR (0.464), matching its strong probe semantics (79.3%) and track-wise output.

Without retraining, ScaperSynth keeps the STARSS23 classes but places events in 11 unseen rooms, while TAU-NIGENS uses a different recording setup and shares only 6 classes. Fine-tuned seq-JEPA has the best LE on both (19.8° and 25.3°). Best LR goes to fine-tuned SIE on ScaperSynth and seq-JEPA on TAU-NIGENS; EINV2 is about a degree behind on TAU-NIGENS in LE.

SELD also reorders the invariance–equivariance methods. SimCLR-EquiMod probes better than SIE on classification and DoA (Table 1), yet SIE is better on almost every SELD metric. Linear probes use single, clean sources, whereas SELD adds overlapping events in real rooms and more classes. EquiMod's contrastive objective, which repels clips that share a class, may therefore not transfer as well to this setting. seq-JEPA generalizes best across both settings. Three design differences may explain this: it needs no partition of the latent, which SIE and EquiMod achieve partially based on linear probing; its objective is predictive, with no negatives; and it infers invariance from many rotated views, not from each pair. We therefore use seq-JEPA for the language tasks of Section 5.3.

5.3 Spatial captioning and retrieval

Finally, we answer whether the representation is expressive enough for another modality to be grounded in it. Table 4 depicts the effect of replacing the Spatial-CLAP encoder (S-CLAP) with seq-JEPA. The encoder is kept fully frozen for both retrieval and captioning. Our encoder improves retrieval in both settings, raising R@1 by 2 points (text-to-audio) and 3.9 points (audio-to-text) and R@5 by a remarkable 7 points in each. The improvement in directional accuracy is the most significant, with 10.9 points improvement. These results show that our frozen representation transfers spatial information to language more effectively than one trained directly for spatial detection in a supervised scheme. Appendix A.6 reports what happens when the encoder is instead fine-tuned with the captioning heads.

Table 4: Spatial audio–text retrieval (R@k: recall at rank k) and directional accuracy (Dir. Acc.) for S-CLAP and our frozen seq-JEPA encoder.

	Text-to-Audio		Audio-to-Text		Dir.
Method	**R@1 ↑**	**R@5 ↑**	**R@1 ↑**	**R@5 ↑**	**Acc. ↑**
S-CLAP	15.8	38.8	14.7	38.3	67.9
Ours	**17.8**	**45.9**	**18.6**	**45.9**	**78.8**

Captioning tests how the representation transfers to text (Table 5; all metrics in Appendix Table 9). Our encoder improves on every metric, with the gains concentrating on the metrics that score the spatial understanding: spatial accuracy rises from 0.459 to 0.657 and DW-SBERT from 0.188 to 0.228. Content metrics improve more modestly: SPIDEr, the standard captioning metric combining n-gram and scene-graph agreement, rises from 0.463 to 0.492, while SBERT, which compares sentence meaning, is essentially unchanged. Although linear probing shows lower source classification for seq-JEPA than for EINV2, the encoder used by S-CLAP (67.7% against 79.3%; Table 1), seq-JEPA retains enough semantics to outperform EINV2 in both retrieval and captioning, including the metrics that primarily measure semantics. The improvement is largest for spatial information, where the limitations of EINV2 are most apparent. Because the compact decoder cannot compensate for missing directions with text priors, this spatial gain remains directly visible.

Table 5: Spatial captioning. SPIDEr, SBERT: content; DW-SBERT, Spatial Acc.: direction.

Metric	**S-CLAP**	**Ours**
SPIDEr	.4630	**.4918**
SBERT	.5666	.5677
DW-SBERT	.1880	**.2281**
Spatial Acc.	.4590	**.6570**

6 Conclusion

We have introduced a framework for structuring spatial audio representations through invariance–equivariance learning, using FOA rotations as a label-free training signal. Learning equivariance to rotation organizes source direction as a structured axis of the latent space, while invariance preserves what is sounding. On single-source probes, this yields a clear spatial advantage over all baselines, with lower classification accuracy than reconstruction-based or supervised encoders. In realistic tasks, however, the representation generalizes better: seq-JEPA achieves the best SELD score on real recordings, surpassing supervised systems even when frozen, and outperforms the supervised Spatial-CLAP encoder in retrieval and captioning, including on semantic metrics. Because the objective requires no labels, pretraining can scale well beyond the simulated and real data used here. The same principle extends to other transformations whose effect on a scene can be simulated, such as source translations and changes in distance, and to the composition of sources within a scene. More broadly, by encoding what is sounding and where it occurs in a single representation, invariance–equivariance learning offers a step toward general-purpose spatial audio foundation models that can be shared across modalities.

AI USE STATEMENT

In this work, we used generative AI tools for sentence paraphrasing, and drafting code after which they carefully reviewed and revised the content before using it within the framework and manuscript. We have not used generative AI tools for generation of research ideas, experimental design development or methologies. All technical and scientific contributions were conceptualized and developed by the authors listed in this article. Generated code development was fully reviewed by the authors. The authors take full responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All the claims presented in this work are backed up by code and datasets. The source code, model checkpoints and datasets will be made publicly available upon article acceptance. Detailed description of the datasets used is presented in the Appendix, and the released source code will contain documentation and scripts to reproduce the relevant experiments presentend in this work.

ACKNOWLEDGMENTS

-

REFERENCES

Sharath Adavanne, Archontis Politis, Joonas Nikunen, and Tuomas Virtanen. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 13(1):34–48, 2019.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Jens Blauert. *Spatial hearing: the psychophysics of human sound localization*. The MIT press, 1996.

Yin Cao, Turab Iqbal, Qiuqiang Kong, Fengyan An, Wenwu Wang, and Mark D. Plumbley. An improved event-independent network for polyphonic sound event localization and detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

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 *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2022.

Chuyang Chen, Sivan Ding, Adrian S Roman, and Juan P Bello. Probing spatial structure in pretrained audio representations. *arXiv preprint arXiv:2606.05544*, 2026.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. BEATs: Audio pre-training with acoustic tokenizers. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

Huw Cheston, Adrian Stepien, Juan Azcarreta, Adrian S. Roman, Chuyang Chen, Cagdas Bilen, and Iran R. Roman. AudibleLight: A controllable, end-to-end API for soundscape synthesis across ray-traced & real-world measured acoustics. In *DMRN+20: Digital Music Research Network One-Day Workshop*, 2025.

Taco S. Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2016.

Taco S. Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical CNNs. In *International Conference on Learning Representations (ICLR)*, 2018.

Rumen Dangovski, Li Jing, Charlotte Loh, Seungwook Han, Akash Srivastava, Brian Cheung, Pulkit Agrawal, and Marin Soljačić. Equivariant contrastive learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Alexandre Devillers and Mathieu Lefort. EquiMod: An equivariance module to improve self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2023.

Antoni Dimitriadis, Siqi Pan, Vidhyasaharan Sethu, and Beena Ahmed. Spatial HuBERT: Self-supervised spatial speech representation learning for a single talker from multi-channel audio. *arXiv preprint arXiv:2310.10922*, 2023.

Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. Project Aria: A new tool for egocentric multi-modal AI research. *arXiv preprint arXiv:2308.13561*, 2023.

Carlos Esteves, Christine Allen-Blanchette, Ameesh Makadia, and Kostas Daniilidis. Learning SO(3) equivariant representations with spherical CNNs. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

Karim Farid, Rajat Sahay, Yumna Ali Alnaggar, Simon Schrodi, Volker Fischer, Cordelia Schmid, and Thomas Brox. What drives compositional generalization? the importance of continuous training objectives in visual generative models. *arXiv preprint arXiv:2510.03075*, 2025.

Zhengcong Fei, Mingyuan Fan, and Junshi Huang. A-JEPA: Joint-embedding predictive architecture can listen. *arXiv preprint arXiv:2311.15830*, 2023.

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.

Fabian B. Fuchs, Daniel E. Worrall, Volker Fischer, and Max Welling. SE(3)-transformers: 3D roto-translation equivariant attention networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Quentin Garrido, Laurent Najman, and Yann LeCun. Self-supervised learning of split invariant equivariant representations. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Beat Gfeller, Christian Frank, Dominik Roblek, Matt Sharifi, Marco Tagliasacchi, and Mihajlo Velimirović. SPICE: Self-supervised pitch estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1118–1128, 2020.

Hafez Ghaemi, Eilif B. Muller, and Shahab Bakhtiari. seq-JEPA: Autoregressive predictive learning of invariant-equivariant world models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations (ICLR)*, 2018.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Julien Guinot, Elio Quinton, and György Fazekas. Leave-one-equivariant: Alleviating invariance-related information loss in contrastive music representations. *arXiv preprint arXiv:2412.18955*, 2024.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented transformer for speech recognition. In *Proceedings of Interspeech*, 2020.

Geoffrey E. Hinton, Alex Krizhevsky, and Sida D. Wang. Transforming auto-encoders. In *International Conference on Artificial Neural Networks (ICANN)*, pp. 44–51, 2011.

Jinbo Hu, Yin Cao, Ming Wu, Fang Kang, Feiran Yang, Wenwu Wang, Mark D. Plumbley, and Jun Yang. PSELDNets: Pre-trained neural networks on a large-scale synthetic dataset for sound event localization and detection. *arXiv preprint arXiv:2411.06399*, 2024.

Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Divyansh Jha, Yuanfang Xie, Varan Mehra, and Brennen Yu. BRo-JEPA: Learning modular arithmetic in latent space. *arXiv preprint arXiv:2606.01372*, 2026.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pp. 119–132, 2019.

Daniel Aleksander Krause and Archontis Politis. [DCASE2024 Task 3] Synthetic SELD mixtures for baseline training, 2024. URL `https://zenodo.org/records/10932241`.

Hamid Krim and Mats Viberg. Two decades of array signal processing research: the parametric approach. *IEEE signal processing magazine*, 13(4):67–94, 1996.

Hankook Lee, Kibok Lee, Kimin Lee, Honglak Lee, and Jinwoo Shin. Improving transferability of representations via augmentation-aware self-supervision. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Huadai Liu, Tianyi Luo, Qikai Jiang, Kaicheng Luo, Peiwen Sun, Jialei Wan, Rongjie Huang, Qian Chen, Wen Wang, Xiangtai Li, Shiliang Zhang, Zhijie Yan, Zhou Zhao, and Wei Xue. Omni-Audio: Generating spatial audio from 360-degree video, 2025. URL `https://arxiv.org/abs/2504.14906`.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. Leworld-model: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint arXiv:2603.19312*, 2026.

Luca Mazzon, Yuma Koizumi, Masahiro Yasuda, and Noboru Harada. First order ambisonics domain spatial augmentation for DNN-based direction of arrival estimation. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2019.

Annamaria Mesaros, Sharath Adavanne, Archontis Politis, Toni Heittola, and Tuomas Virtanen. Joint measurement of localization and detection of sound events. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019.

Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6707–6717, 2020.

Pedro Morgado, Yi Li, and Nuno Vasconcelos. Learning representations from audio-visual spatial alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Christian Nachbar, Franz Zotter, Etienne Deleflie, and Alois Sontacchi. AmbiX – a suggested ambisonics format. In *Proceedings of the Ambisonics Symposium*, 2011.

Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. BYOL for audio: Self-supervised learning for general-purpose audio representation. In *International Joint Conference on Neural Networks (IJCNN)*, 2021.

Jung Yeon Park, Ondrej Biza, Linfeng Zhao, Jan Willem van de Meent, and Robin Walters. Learning symmetric embeddings for equivariant world models. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2022.

Archontis Politis, Sharath Adavanne, 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, 2020.

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 Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2021.

Alain Riou, Stefan Lattner, Gaëtan Hadjeres, and Geoffroy Peeters. PESTO: Pitch estimation with self-supervised transposition-equivariant objective. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 535–544, 2023.

Iran R. Roman, Christopher Ick, Sivan Ding, Adrian S. Roman, Brian McFee, and Juan P. Bello. Spatial scaper: A library to simulate and augment soundscapes for sound event localization and detection in realistic rooms. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

Ryotaro Sato, Kenta Niwa, and Kazunori Kobayashi. Ambisonic signal processing DNNs guaranteeing rotation, scale and time translation equivariance. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021. doi: 10.1109/TASLP.2021.3069193.

Kentaro Seki, Yuki Okamoto, Kouei Yamaoka, Yuki Saito, Shinnosuke Takamichi, and Hiroshi Saruwatari. Spatial-CLAP: Learning spatially-aware audio-text embeddings for multi-source conditions, 2025. URL `https://arxiv.org/abs/2509.14785`.

Kazuki Shimada, Yuichiro Koyama, Shusuke Takahashi, Naoya Takahashi, Emiru Tsunoo, and Yuki Mitsufuji. Multi-ACCDOA: Localizing and detecting overlapping sounds from the same class with auxiliary duplicating permutation invariant training. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

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 *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023.

Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation- and translation-equivariant neural networks for 3D point clouds. *arXiv preprint arXiv:1802.08219*, 2018.

Qing Wang, Jun Du, Hua-Xin Wu, Jia Pan, Feng Ma, and Chin-Hui Lee. A four-stage data augmentation approach to ResNet-Conformer based acoustic modeling for sound event localization and detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1251–1264, 2023.

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.

Maurice Weiler, Mario Geiger, Max Welling, Wouter Boomsma, and Taco Cohen. 3D steerable CNNs: Learning rotationally equivariant features in volumetric data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

Fei Xia, Amir R Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9068–9079. IEEE, 2018.

Goksenin Yuksel, Pierre Guetschel, Michael Tangermann, Marcel van Gerven, and Kiki van der Heijden. WavJEPA: Semantic learning unlocks robust audio foundation models for raw waveforms. *arXiv preprint arXiv:2509.23238*, 2025a.

Goksenin Yuksel, Marcel van Gerven, and Kiki van der Heijden. Gram: Spatial general-purpose audio representation models for real-world applications. *arXiv preprint arXiv:2506.00934*, 2025b.

Goksenin Yuksel, Marcel van Gerven, and Kiki van der Heijden. EquiSELD: Efficient training of equivariant sound event localization and detection networks. *arXiv preprint arXiv:2609.23156*, 2026.

Zhisheng Zheng, Puyuan Peng, Ziyang Ma, Xie Chen, Eunsol Choi, and David Harwath. BAT: Learning to reason about spatial sounds with large language models. *arXiv preprint arXiv:2402.01591*, 2024.

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

A Appendix

A.1 Method Comparison

Table 6: The three methods along the axes that determine what their representations encode (Section 3.4). The rotation R reaches the predictor in all three; what differs is whether the latent is partitioned.

Method	Views	Latent	How R enters	Objective (collapse prev.)
SIE	2	Fixed partition	ρ_R on equi. head	ℓ_2 regression (VICReg)
SimCLR-EquiMod	2	Two full heads	ρ_R on equi. head	InfoNCE (negatives)
seq-JEPA	$L{+}1$	Single, unsplit	Per-view token + predictor	cosine (EMA asym.)

Changes from the original implementations. All three methods were developed for images, and we adapt each to spatial audio in the same way, so that the comparison of Section 3.4 isolates the objective. The group action is the analytic Wigner-D rotation of Eq. 3, applied to the raw FOA waveform before feature extraction, in place of image augmentations or re-rendered views. The relative rotation reaches the networks as a unit quaternion, as in SIE, but with roll fixed, so that only azimuth and elevation vary. The spectral transformation T_s (Gaussian noise, gain scaling, time–frequency masking) replaces photometric augmentations. The shared backbone is the audio Conformer of Section 3.3 rather than a ResNet, and its pooled output is projected to 512 dimensions before any split. Training adds the polyphony and real-data curriculum of Section 4.1, with loss weights tuned per method. For seq-JEPA we probe and transfer the view embedding $\hat{z}$, a single unsplit vector, rather than the aggregate s; it is the representation reported throughout Section 5.

A.2 Pretraining Details

Synthetic data generation. 12,000/3,600/3,600 train/validation/test room impulse responses (RIRs) are generated with AudibleLight (Cheston et al., 2025) on Gibson meshes (Xia et al., 2018), which yields geometrically faithful simulation at scale. Source positions are sampled uniformly in azimuth ($\pm 180°$), elevation ($\pm 60°$) and distance ($[0.5, 2.5]$ m); dry sources are drawn from a 34-class subset of the FSD50K ontology (Fonseca et al., 2022).

Sound event curation. For monophonic sound events, we curate single-source stems from FSD50K (Fonseca et al., 2022). We follow the data curation strategy of PSELDnets (Hu et al., 2024) and filter out classes with fewer than 30 clips. As a result, we select 190 classes, yielding 37,999 single sound source files, which we split into training and validation sets.

Input features. Clips are 5 s at 24 kHz. Each is converted to 7-channel FOA log-mel (4 channels) and intensity-vector (3 channels) features with a 300-sample hop and 128 mel bins, giving inputs of shape $(B, 7, 401, 128)$.

Real data curation. Sphere360 (Liu et al., 2025) collects 360° YouTube videos with paired first-order ambisonic audio. We curate the FOA tracks and retain approximately 100 hours. The curation consisted of removing clips whose channels were muted by simply discaring those with no signal on all 4 FOA channels. These recordings carry no class or DoA labels; their role is purely to expose the encoder to naturally recorded room acoustics and to events that simulation does not cover, which is why they enter only in the final curriculum stage (Section 4.1).

Hyperparameter search. Loss weights (the relative weighting of invariance and equivariance terms, and for SIE the variance and covariance regularization coefficients) are tuned per method with a Weights & Biases Bayesian sweep. Each candidate trains for up to 1k pretraining epochs and is scored periodically with the composite metric of Section 4.1, which is both the search objective and the signal for Hyperband early stopping of unpromising runs. The validation set contains only synthetic mixtures built from held-out RIRs and validation clips, so its labels are simulator metadata rather than annotations. Sphere360 is excluded from validation because it has no labels; its 25% share of training batches never overlaps with the validation data. The same score selected the Conformer backbone (Section 3.3), which outperformed the other PSELDnets encoders and trained more stably.

A.3 Representation Statistics During Pretraining

The online probe of Section 4.1 reports what is *decodable* from the representation, which is not the same question as whether the representation is degenerate. A probe can lose accuracy for many reasons, it needs labelled and rotated data, and it is refreshed only every few epochs, so a run that begins to collapse is caught late. Alongside it we therefore track two second-order statistics of the representation itself, which require no labels and are computed on every step. Both are evaluated under `no_grad` on the current training minibatch, gathered across devices so that the batch is the global one, and neither contributes gradient to f_θ: they are used to stop and select runs, not to train them. They are statistics of training batches rather than of a held-out split, which is what makes them cheap enough to compute continuously.

Definitions. Let $Z \in \mathbb{R}^{B\times D}$ collect the D-dimensional outputs of one such layer for a global batch of B clips. The first statistic measures how much of the representation is actually being used. Writing $\hat{z}_i = z_i/\|z_i\|_2$ for the ℓ_2-normalized rows, so that scale is factored out and only direction remains,

$$\bar{\sigma}(Z) = \frac{1}{D}\sum_{j=1}^{D} \mathrm{std}_i\left(\hat{Z}_{ij}\right), \tag{4}$$

the mean over dimensions of the per-dimension standard deviation across the batch. The second measures whether those dimensions are redundant. Let $\tilde{Z}$ be Z standardized per dimension to zero mean and unit variance, so that $C = \tilde{Z}^\top\tilde{Z}/(B-1)$ is the Pearson correlation matrix of the D dimensions; then

$$\bar{c}(Z) = \frac{1}{D(D-1)}\sum_{j\neq k} C_{jk}. \tag{5}$$

Note that $\bar{c}$ is a *signed* mean of the off-diagonal entries, not an absolute or squared one; the consequence is discussed at the end of this section. The two statistics normalize differently on purpose: $\bar{\sigma}$ is computed on ℓ_2-normalized rows because collapse is a statement about direction, whereas $\bar{c}$ is computed on standardized columns because redundancy is a statement about which dimensions co-vary.

Where they are measured. Both statistics are applied at up to three depths, for each view: the backbone representation $z = f_\theta(x)$, which is the quantity actually probed and transferred; the projector output; and, for the split methods, the output of the predictor ρ_R. Coverage differs by method (Table 7). SIE and SimCLR-EquiMod report all three; because their projector tensor is the concatenation of the invariant and equivariant heads, $\bar{c}$ at that depth also measures redundancy *between* the two branches, which is the property the split design is supposed to maintain. seq-JEPA has neither a projector nor a predictor on the path being measured, so only the representation statistics exist for it; there, one view is the online encoder output pooled over all L observations in the sequence and the other is the EMA target encoder output.

Ideal values. For $\bar{\sigma}$, the reference point is the isotropic one. If unit-norm representations were spread uniformly over the sphere S^{D-1}, every coordinate would have variance $1/D$, so the target value is

$$\bar{\sigma}^\star = 1/\sqrt{D}, \tag{6}$$

which is ≈ 0.044 for the $D = 512$ backbone and 0.016 to 0.022 for the 2048- to 4096-dimensional projector outputs used here. A healthy run sits within roughly 10–20% of that value and, more importantly, stays flat: the trend over training is more diagnostic than any single reading. Two failures sit on either side of the target. $\bar{\sigma} \to 0$ is complete collapse, every clip mapping to the same direction. $\bar{\sigma}$ substantially *above* $1/\sqrt{D}$ is the opposite and less obvious failure, variance concentrated in a few coordinates while the remainder are inert, which is dimensional collapse presenting as a healthy average.

For $\bar{c}$, the ideal value is 0: dimensions that are mutually uncorrelated carry non-redundant information and the representation occupies its full width. A sustained $|\bar{c}|$ above roughly 0.1 indicates growing redundancy, and drift toward 1 means the dimensions are duplicating one another, which is collapse in its informational form and can occur while $\bar{\sigma}$ still looks acceptable. The two are therefore read together: $\bar{\sigma}$ on target with $\bar{c}$ large is a wide but redundant representation, while $\bar{\sigma}$ falling with $\bar{c}$ near zero is an isotropic but shrinking one. The predictor statistics carry the same targets but

a sharper reading. A predictor whose $\bar{\sigma}$ decays while the encoder's holds has learned the constant map: it is ignoring R, which empties the equivariance term of content while the total loss continues to fall. That failure is invisible in the loss curve, which is why the predictor is instrumented separately.

Table 7: The representation statistics logged during pretraining, their ideal values and their failure signatures. D is the width of the layer being measured (512 for the backbone, 2048–4096 for the concatenated projector outputs). Both are computed on the global training batch under `no_grad` and contribute no gradient.

Statistic	Measured on	Ideal	Failure signature	Reported for
$\bar{\sigma}(z)$	Representation	$1/\sqrt{D}$	$\to 0$; or $\gg 1/\sqrt{D}$	All three
$\bar{c}(z)$	Representation	0	$\lvert\cdot\rvert$ grows toward 1	All three
$\bar{\sigma}$	Projector output	$1/\sqrt{D}$	$\to 0$	SIE, SimCLR-EquiMod
$\bar{c}$	Projector output	0	$\lvert\cdot\rvert$ grows; inv./equi. merge	SIE, SimCLR-EquiMod
$\bar{\sigma}$	Predictor ρ_R	$1/\sqrt{D}$	$\to 0$; ρ_R ignores R	SIE, SimCLR-EquiMod
$\bar{c}$	Predictor ρ_R	0	$\lvert\cdot\rvert$ grows toward 1	SIE, SimCLR-EquiMod

What they mean for each method. The same two numbers carry different weight across the three objectives, because the three prevent collapse differently (Table 6). For SIE, variance and covariance appear *in the loss*: a VICReg-style hinge $\max(0, 1 - \sigma_j)$ per dimension and a squared off-diagonal covariance penalty, both on the projector output (Bardes et al., 2022). The projector statistics there largely restate what the objective is already driving, and confirm that the regularizer is active rather than that the representation is good. The informative pair for SIE is the backbone statistics, which no loss term touches directly and which show whether regularization applied at the projector propagates back into z. For SimCLR-EquiMod there is no variance or covariance term at all; collapse is prevented by in-batch negatives, so both statistics are purely diagnostic and a value near $1/\sqrt{D}$ is meaningful precisely because nothing is pushing it there. The failure to watch for is uniform-but-low-rank, since InfoNCE can spread embeddings over a low-dimensional subspace of the sphere, which appears as $\bar{\sigma}$ on target with $|\bar{c}|$ elevated. For seq-JEPA the loss is a cosine similarity to an EMA target with no negatives, no variance term and no projector, so collapse is prevented by the encoder/target asymmetry alone (Grill et al., 2020) and these statistics are the *only* collapse detector available. The statistic on the EMA target is the critical one: if it decays, the asymmetry has failed and the two encoders are converging on a constant. Because the observation-side statistic pools all L views of a sequence, agreement between the two also indicates that distinct observations are not being encoded identically.

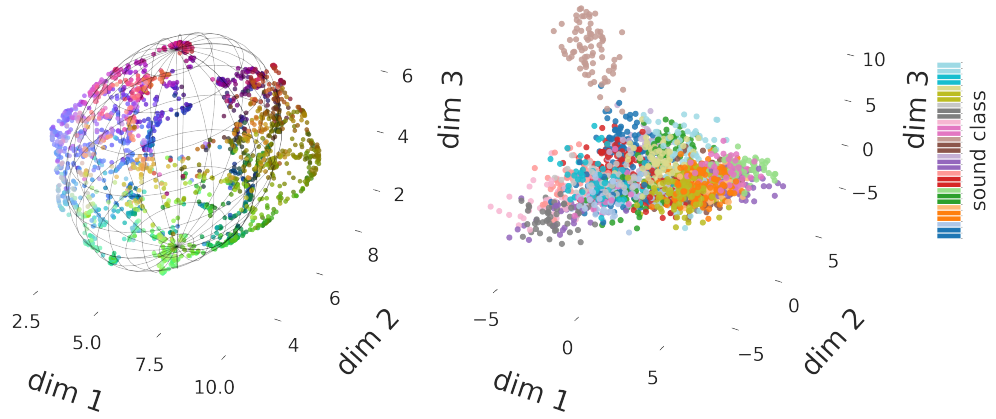
How they are used, and what they do not show. These statistics gate runs rather than rank them. A run whose $\bar{\sigma}$ decays or whose $|\bar{c}|$ climbs is stopped before the composite probe score has had time to register the problem, since they are computed every step while the probe is refreshed every few epochs. Two limitations should be stated. First, because $\bar{c}$ averages signed off-diagonal entries, a value near zero can mask large correlations of opposite sign; it bounds net rather than worst-case redundancy, and is best read as a trend rather than a level. Second, both are second-order batch statistics, so full variance and zero correlation are necessary but not sufficient for a useful representation: a representation can be isotropic and decorrelated while encoding nothing that matters for sound events or direction. That is exactly why the supervised online probe is retained alongside them, and why neither signal is used on its own.

A.4 Evaluation Head Training

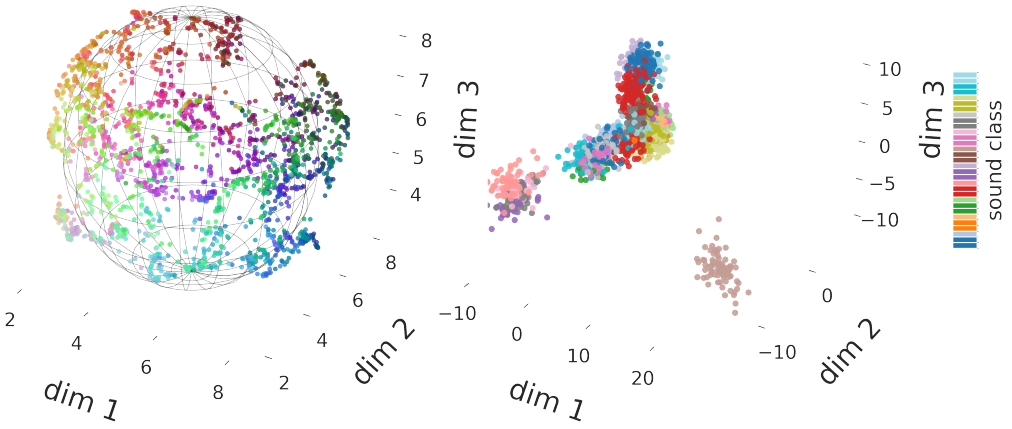
Probes are trained on frozen features from single-source synthetic training and evaluation splits with a fixed metadata seed; these splits supply the class and DoA labels used to fit the probes, and the supervised per-task baseline of Section 4.2 is trained and validated on the same data. Backbone and projectors remain frozen throughout except under the FT protocol of Section 4.2; only the task-specific probe or decoder head is ever trained from scratch

A.5 Latent Geometry Projections

Figure 2 asks not whether spatial information is decodable but whether it is organized, and is produced as follows. For the trajectories, we take held-out clips, sweep each through a full 360° of azimuth rotation by applying D_R to the waveform, embed every rotated version with the frozen encoder, and project the resulting latents with PCA. A representation that tracks the group action should trace a smooth curve that closes as azimuth returns to its starting value. Each model is given its own PCA basis, so the axes are not comparable across panels; what is comparable is the extent of a trajectory relative to the distance between clips, which is why the panels are read that way in Section 5.1. For the decomposition, we fit a linear regression from seq-JEPA's latent to source direction, treat the subspace it spans as the DoA-predictive component, and take its orthogonal complement as the residual. The former is visualized with UMAP and colored by direction, the latter with LDA over sound classes, which tests whether semantic structure survives once the linear spatial component is removed.



(a) seq-JEPA



(b) GRAM-T

Figure 3: The decomposition of Figure 2–(b) shown at a larger size for **(a)** seq-JEPA and **(b)** GRAM-T. Within each panel, the left plot is the DoA-predictive subspace under UMAP, colored by source location mapping (azimuth, elevation) to RGB. The right plot is the orthogonal residual under LDA, colored by sound class. Each projection uses its own basis, so axes are not comparable across panels or with Figure 2.

Table 8: Spatial audio–text retrieval with the seq-JEPA encoder *fine-tuned* alongside the captioning heads, against S-CLAP. R@1, R@5 and R@10 denote recall at the corresponding rank, and Dir. Acc. directional accuracy. The frozen-encoder results of Table 4 are the ones reported in the main text.

	Text-to-Audio			Audio-to-Text			Dir. Acc. ↑
Method	**R@1 ↑**	**R@5 ↑**	**R@10 ↑**	**R@1 ↑**	**R@5 ↑**	**R@10 ↑**	
Spatial-CLAP	15.8	38.8	51.9	14.7	38.3	51.1	67.9
seq-JEPA-CLAP	**29.7**	**58.0**	**69.7**	**30.2**	**59.4**	**70.6**	**83.4**

A.6 Fine-tuning the Encoder for Language Alignment

Section 5.3 keeps the encoder fully frozen, so that any difference against S-CLAP is attributable to the representation alone. For completeness we also report what happens when the encoder is fine-tuned together with the captioning heads. Retrieval improves substantially: R@1 rises from 17.8 to 29.7 (text-to-audio) and from 18.6 to 30.2 (audio-to-text), and directional accuracy from 78.8 to 83.4 (Table 8, which also adds rank 10). Captioning, however, does not follow: every metric in Table 9 is slightly lower than under the frozen protocol, including the two that score direction (spatial accuracy 0.657 to 0.614, DW-SBERT 0.228 to 0.215). Adapting the encoder therefore helps align the two modalities in a shared retrieval space, but does not make the captions themselves better, and it removes the guarantee that the comparison measures the pretrained representation rather than the adaptation. We report the frozen protocol in the main text for that reason.

Table 9: Spatial audio captioning when the seq-JEPA encoder is fine-tuned with the captioning heads, against S-CLAP and our frozen encoder (Table 5). Fine-tuning lowers every captioning metric relative to the frozen encoder.

Metric	**S-CLAP**	**Ours (frozen)**	**Ours (fine-tuned)**
BLEU-4	0.1541	**0.1687**	0.1626
ROUGE-L	0.3845	**0.4110**	0.3971
METEOR	0.1885	**0.1994**	0.1937
CIDEr	0.7007	**0.7447**	0.6633
SPICE	0.2254	**0.2389**	0.2270
SPIDEr	0.4630	**0.4918**	0.4452
BERTScore	0.9048	**0.9075**	0.9050
SBERT	0.5666	0.5677	0.5560
DW-SBERT	0.1880	**0.2281**	0.2148
Spatial Acc.	0.4590	**0.6570**	0.6140