JEPA-MUSIC: SELF-SUPERVISED PREDICTIVE SUBSPACE LEARNING FOR MULTIPLE SIGNAL CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Direction-of-arrival (DoA) estimation is fundamental to modern wireless communication and sensing systems. While the classical multiple signal classification (MUSIC) algorithm offers high angular resolution and clear interpretability, its performance degrades under limited snapshots, low signal-to-noise ratios (SNRs), and coherent sources due to empirical covariance breakdown. Conversely, data-driven methods mitigate such degradation but typically demand large DoA-labeled datasets and yield less transparent inference. To bridge this divide, we propose JEPA-MUSIC, a model-based self-supervised learning framework that integrates representation learning with subspace-based MUSIC inference. Specifically, JEPA-MUSIC constructs paired observations from the same signal scene and employs a tailored joint-embedding predictive architecture (JEPA) to learn to recover a structured, MUSIC-compatible covariance without requiring ground-truth DoA labels. During inference, this recovered covariance directly feeds into standard MUSIC for spatial-spectrum search, preserving its rigorous model-based interpretability. Extensive evaluations across diverse sensing conditions demonstrate that JEPA-MUSIC substantially outperforms classical MUSIC in challenging regimes and matches or surpasses strong data-driven baselines, while achieving higher subspace fidelity, lower tail failure rates, and superior sample efficiency.

1 INTRODUCTION

The rapid development of modern wireless systems has enabled vast numbers of devices to operate simultaneously in increasingly dense signal environments (Kalør et al., 2025). In such settings, reliable operation depends on spatially resolving concurrent signals and determining the directions from which they arrive. Direction-of-arrival (DoA) estimation addresses this need by estimating source directions from mixed signals received by antenna or sensor arrays (Molaei et al., 2024). The resulting spatial information supports source separation, localization, and tracking (Kulkarni et al., 2025), making DoA estimation widely applicable to radar and sonar systems (Molaei et al., 2025; Bekkerman & Tabrikian, 2006; Ma et al., 2025), acoustic sensing (Dwivedi et al., 2025; Sheng & Hu, 2005; Evers et al., 2020), and autonomous navigation (Xu et al., 2023).

Existing DoA estimation methods have developed along two main directions: model-based and data-driven approaches. Among model-based methods, multiple signal classification (MUSIC) is the most established and widely used high-resolution DoA estimator (Schmidt, 1986). It decomposes the spatial covariance of array observations into signal and noise subspaces and identifies source directions through *peaks* in an interpretable spatial spectrum. When the assumed signal model holds and the covariance estimate is sufficiently accurate, MUSIC can resolve closely spaced sources with high angular resolution.

However, applying MUSIC reliably in practical environments remains challenging. One major difficulty is that the true spatial covariance is unknown and must be estimated from a limited number of noisy observations. When only a small number of observations are available or the signal-to-noise ratio (SNR) is low, the resulting estimate can be inaccurate, making it difficult for MUSIC to separate signals from noise (Shmuel et al., 2025; Schmidt, 1986; Merkofer et al., 2024). In addition, multipath propagation can produce highly correlated arrivals, reducing the effective rank of the signal covariance and making distinct source directions difficult to resolve. Consequently, true source directions may

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.

be missed or incorrectly merged, while spurious *peaks* may appear in the MUSIC spatial spectrum, resulting in large DoA estimation errors (Shan et al., 1985; Merkofer et al., 2024).

In contrast, data-driven methods learn mappings from noisy array observations to DoA estimates (Lin et al., 2026; Su et al., 2026; Papageorgiou et al., 2021; Elbir, 2020; Chakrabarty & Habets, 2019; Liu et al., 2018; Wu et al., 2019). This reduces their reliance on an exact analytical signal model and allows them to capture noise and propagation effects represented in the training data. However, this flexibility comes at a cost. Direct DoA prediction without explicit subspace decomposition or spatial-spectrum analysis offers limited physical interpretability, while dependence on the training distribution can hinder generalization to unseen conditions. Importantly, these methods typically require large DoA-labeled datasets for training, yet obtaining accurate labels in practical environments often requires external reference equipment, making large-scale data collection costly.

This work proposes a model-based learning approach to address the above challenges. Motivated by the covariance bottleneck of MUSIC, we aim to learn a more reliable covariance representation for limited and noisy array observations, enabling MUSIC to better distinguish the signal and noise subspaces and estimate source directions through an interpretable spatial spectrum. This formulation allows the model to be trained using covariance information from other observations of the same signal scene, thus bypassing the need to acquire ground-truth DoA labels. To this end, we first propose a paired-observation construction strategy, which samples the same signal scene under different snapshot counts and SNR conditions to form a limited, noisy context observation and a higher-quality target observation. We then develop a tailored joint-embedding predictive architecture (JEPA) that predicts the target representation from the context observation to guide self-supervised MUSIC-compatible covariance learning. As a result, the proposed method, termed JEPA-MUSIC, combines the flexibility of data-driven learning with the interpretability of model-based inference, achieving state-of-the-art (SOTA) DoA estimation performance with high training-sample efficiency. Our main contributions are summarized as follows:

- We propose JEPA-MUSIC, a novel model-based learning framework that learns reliable MUSIC-compatible covariance and signal-subspace representations through a tailored JEPA. By seamlessly integrating learned covariance recovery with the standard MUSIC inference procedure, JEPA-MUSIC directly addresses the covariance bottleneck of MUSIC while preserving its model-based interpretability.
- We develop a paired-observation strategy that enables self-supervised representation learning for JEPA-MUSIC without ground-truth DoA labels. We theoretically derive upper bounds relating covariance recovery error to perturbations in both the noise subspace and the MUSIC pseudospectrum denominator, thereby directly bridging the self-supervised training objective and the subspace quantities governing MUSIC inference.
- We comprehensively evaluate JEPA-MUSIC across diverse benchmarks. Experimental results demonstrate that our method substantially outperforms classical MUSIC and surpasses competitive data-driven baselines, particularly in challenging regimes with low SNR, limited snapshots, and coherent sources, while exhibiting strong cross-distribution generalization and superior training-sample efficiency.

2 Related Work

Supervised Learning for DoA Estimation. Deep learning has been widely adopted for DoA estimation, broadly falling into two categories: end-to-end approaches that directly map array measurements or sample covariances to source directions (Lin et al., 2026; Papageorgiou et al., 2021), and hybrid model-based architectures that learn intermediate representations, such as discretized spectrum regions (Elbir, 2020), reconstructed covariance matrices (Barthelme & Utschick, 2021; Merkofer et al., 2024), or surrogate autocorrelations via differentiable subspace estimation (Shmuel et al., 2025). While these methods enhance robustness under low SNR and few snapshots, they universally rely on supervised training with ground-truth DoA labels (Lin et al., 2026; Su et al., 2026; Papageorgiou et al., 2021; Merkofer et al., 2024; Shmuel et al., 2025) or precomputed synthetic spectra (Elbir, 2020), limiting their real-world applicability. In contrast, JEPA-MUSIC operates fully self-supervised, leveraging paired scene observations to learn structured covariance recovery without any angle annotations or proxy targets.

Self-Supervised Predictive Representation Learning. Recent wireless self-supervised methods extract representations directly from unannotated observations via denoising (Liu et al., 2026b), multimodal alignment (Chu et al., 2026), and channel-aware modeling (Liu et al., 2026a). However, their objectives primarily target signal recognition, localization, or general radio understanding rather than multi-source DoA estimation. In parallel, JEPAs learn latent representations through context-to-target prediction (Assran et al., 2023; Nam et al., 2026; Chen et al., 2026; Li et al., 2026; Huang et al., 2026). Nevertheless, generic JEPAs impose no covariance structure and do not guarantee preservation of the MUSIC-relevant signal subspace. JEPA-MUSIC bridges this gap by tailoring the predictive architecture and learning objectives to paired array observations, reconstructing a structured, MUSIC-compatible covariance while explicitly aligning its signal subspace with that of the target observation.

3 Problem Formulation and Preliminaries

Notation. Bold lowercase and uppercase letters denote vectors and matrices, respectively. For example, random scalars, vectors, and matrices are denoted by y, $\mathbf{y}$, and $\mathbf{Y}$, with realizations y, $\boldsymbol{y}$, and $\boldsymbol{Y}$, respectively. $\mathbb{C}^{M\times N}$ denotes the set of complex $M \times N$ matrices. $(\cdot)^T$, $(\cdot)^H$, $\mathrm{Tr}(\cdot)$, $\mathbb{E}[\cdot]$, and $\|\cdot\|_{\mathrm{F}}$ denote transpose, Hermitian transpose, trace, expectation, and the Frobenius norm, respectively. $\boldsymbol{I}_M$ denotes the $M \times M$ identity matrix. $\mathrm{Re}(\cdot)$ and $\mathrm{Im}(\cdot)$ denote the real and imaginary parts, respectively, and $j = \sqrt{-1}$ is the imaginary unit.

3.1 Signal Model and DoA Estimation

We consider a general wireless scenario with D signal sources observed by an M-antenna uniform linear array (ULA). A source may correspond, for example, to an active wireless device or a signal reflected by an object in the environment. The direction from which the signal associated with source d, $d = 1, \ldots, D$, arrives at the array is denoted by θ_d. We collect all source directions in $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_D]^T$.

At observation instant n, all antennas simultaneously measure the incoming signals, forming a *snapshot* $\mathbf{y}(n) = [\mathsf{y}_1(n), \ldots, \mathsf{y}_M(n)]^T \in \mathbb{C}^M$ (Van Trees, 2002), where $\mathsf{y}_m(n)$ is the signal measured by antenna m. Under the standard narrowband far-field model (Krim & Viberg, 1996; Stoica & Moses, 2005), a signal arriving from direction θ_d produces a characteristic phase pattern across the array. This pattern is represented by the steering vector, defined as:

$$\boldsymbol{a}(\theta_d) = \left[1, e^{-j\pi \sin\theta_d}, \ldots, e^{-j\pi(M-1)\sin\theta_d}\right]^T \in \mathbb{C}^M, \tag{1}$$

where $j = \sqrt{-1}$, and the complex exponential terms encode the direction-dependent relative phase shifts across the antennas. A detailed interpretation of the steering vector is provided in Appendix A.1. Using this representation, the received array snapshot at observation instant n can be written as:

$$\mathbf{y}(n) = \sum_{d=1}^{D} \boldsymbol{a}(\theta_d)\mathsf{s}_d(n) + \mathbf{w}(n), \tag{2}$$

where $\mathsf{s}_d(n)$ is the signal associated with source d, and $\mathbf{w}(n) \in \mathbb{C}^M$ denotes additive noise.

We collect N snapshots over a short interval during which the source directions remain unchanged while the source signals and noise may vary. Stacking these snapshots gives:

$$\mathbf{Y} = [\mathbf{y}(1), \ldots, \mathbf{y}(N)] = \boldsymbol{A}(\boldsymbol{\theta})\mathbf{S} + \mathbf{W} \in \mathbb{C}^{M\times N}, \tag{3}$$

where $\boldsymbol{A}(\boldsymbol{\theta}) = [\boldsymbol{a}(\theta_1), \ldots, \boldsymbol{a}(\theta_D)] \in \mathbb{C}^{M\times D}$ is the steering matrix. The matrices $\mathbf{S} \in \mathbb{C}^{D\times N}$ and $\mathbf{W} \in \mathbb{C}^{M\times N}$ collect the source signals and noise, respectively.

The objective of DoA estimation is to recover the unknown source directions $\boldsymbol{\theta}$ from the mixed and noisy observations $\mathbf{Y}$. Neither the source signals in $\mathbf{S}$ nor the noise realization in $\mathbf{W}$ is known to the estimator.

3.2 MUSIC Algorithm and Its Covariance Bottleneck

MUSIC estimates source directions from the spatial relationships across the antennas (Schmidt, 1986). As illustrated in Figure 1, it first summarizes the received snapshots through their spatial covariance:

$$\boldsymbol{R}_{\mathrm{y}} = \mathbb{E}\big[\mathbf{y}(n)\mathbf{y}^H(n)\big] \in \mathbb{C}^{M\times M}, \tag{4}$$

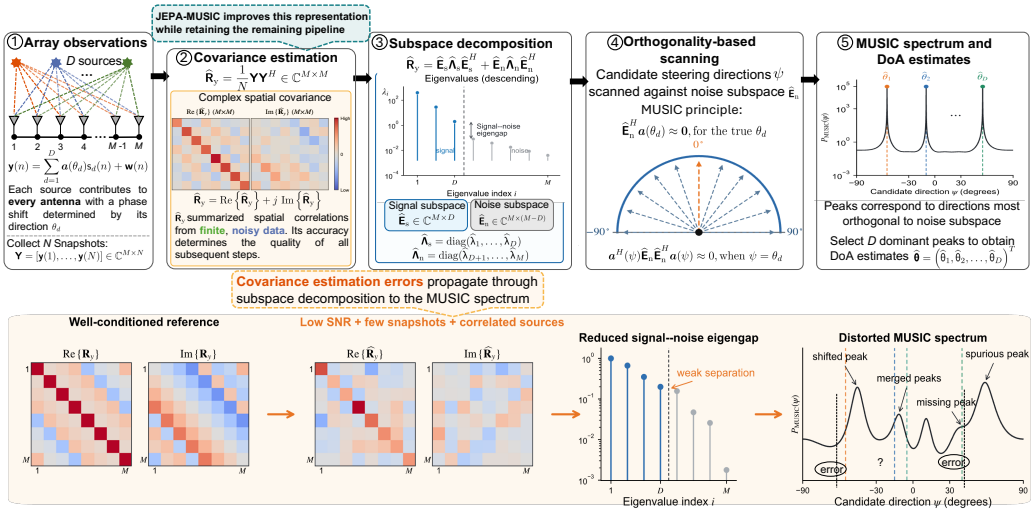

Figure 1: Classical MUSIC pipeline and its covariance bottleneck. Limited snapshots, low SNR, and source coherence can compromise subspace estimation and distort the resulting MUSIC spectrum.

where the expectation is taken over source-signal and noise realizations. Because the exact expectation is unavailable from finite observations, MUSIC uses the sample covariance:

$$\widehat{\mathbf{R}}_{\mathrm{y}} = \frac{1}{N}\mathbf{Y}\mathbf{Y}^H \in \mathbb{C}^{M\times M}. \tag{5}$$

Each entry of $\widehat{\mathbf{R}}_{\mathrm{y}}$ captures the relationship between a pair of antenna measurements and therefore preserves the direction-dependent spatial phase patterns.

Increasing N generally provides a more stable covariance estimate, but also requires a longer observation interval, which can reduce responsiveness in time-varying wireless environments. Reliable estimation from a limited number of snapshots is therefore important in practice.

Assuming that the number of sources $D < M$ is known, MUSIC applies an eigendecomposition (Abdi, 2007) to the sample covariance:

$$\widehat{\mathbf{R}}_{\mathrm{y}} = \widehat{\mathbf{E}}_{\mathrm{s}}\widehat{\mathbf{\Lambda}}_{\mathrm{s}}\widehat{\mathbf{E}}_{\mathrm{s}}^H + \widehat{\mathbf{E}}_{\mathrm{n}}\widehat{\mathbf{\Lambda}}_{\mathrm{n}}\widehat{\mathbf{E}}_{\mathrm{n}}^H, \tag{6}$$

where $\widehat{\mathbf{E}}_{\mathrm{s}} \in \mathbb{C}^{M\times D}$ contains the eigenvectors associated with the D largest eigenvalues and forms the estimated signal subspace. The remaining eigenvectors form the estimated noise subspace $\widehat{\mathbf{E}}_{\mathrm{n}} \in \mathbb{C}^{M\times(M-D)}$. The diagonal matrices $\widehat{\mathbf{\Lambda}}_{\mathrm{s}}$ and $\widehat{\mathbf{\Lambda}}_{\mathrm{n}}$ contain the corresponding eigenvalues.

Under the ideal MUSIC model, the steering vector of each true source is orthogonal to the noise subspace. For the estimated covariance, this relationship is approximated by

$$\widehat{\mathbf{E}}_{\mathrm{n}}^H \boldsymbol{a}(\theta_d) \approx \mathbf{0}, \qquad d = 1, \ldots, D. \tag{7}$$

MUSIC therefore evaluates candidate directions ψ using the pseudospectrum

$$\mathrm{P}_{\mathrm{MUSIC}}(\psi) = \frac{1}{\boldsymbol{a}^H(\psi)\widehat{\mathbf{E}}_{\mathrm{n}}\widehat{\mathbf{E}}_{\mathrm{n}}^H\boldsymbol{a}(\psi)}, \tag{8}$$

and selects its D dominant peaks as the estimated source directions $\widehat{\boldsymbol{\theta}}$. Further details on the eigendecomposition and signal-noise subspace partition are provided in Appendix A.2.

It can be observed that, once $\widehat{\mathbf{R}}_{\mathrm{y}}$ is obtained, the remaining MUSIC operations follow fixed model-based procedures. Its central challenge therefore lies in obtaining a covariance representation that preserves reliable signal–noise subspace separation. The main sources of degradation can be summarized as follows:

- **Limited snapshots** increase sample-covariance estimation error and reduce subspace accuracy.

- **Low SNR** makes signal–noise subspace separation less reliable under finite observations.
- **Source coherence** can make the source covariance rank deficient, preventing the signal subspace from spanning all source steering vectors.

These effects can produce shifted, merged, missing, or spurious spectrum peaks, as illustrated in the lower part of Figure 1 and analyzed in detail in Appendix A.3. These failure modes motivate learning a more reliable MUSIC-compatible covariance representation while retaining standard MUSIC inference.

4 METHODOLOGY

The proposed JEPA-MUSIC addresses the covariance bottleneck of MUSIC by learning to enhance the covariance representation obtained from limited and noisy array observations. The framework follows a self-supervised cross-view predictive paradigm, in which the context observation is used to recover covariance and signal-subspace information represented more reliably by a higher-quality target observation of the same scene. As illustrated in Figure 2, JEPA-MUSIC consists of three core components: (1) paired observations of the same scene, (2) a JEPA architecture that predicts the target representation from the context observation, and (3) a structured covariance head that produces MUSIC-compatible outputs, trained with covariance and subspace alignment objectives. At inference, the enhanced covariance is directly processed by standard MUSIC to obtain the spatial spectrum and final DoA estimates. We then elaborate each module as follows.

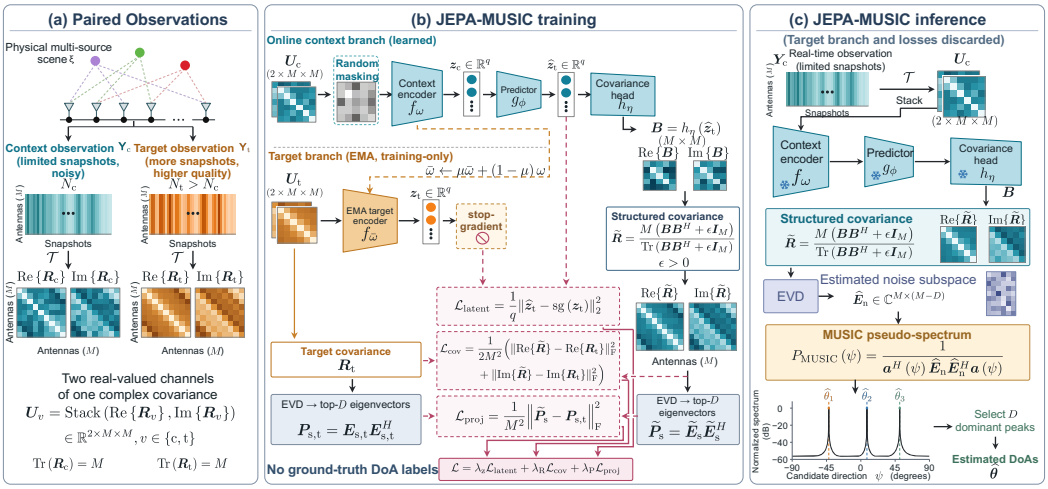

Figure 2: Overview of JEPA-MUSIC. (a) Training data: paired observations of the same scene without ground-truth DoA labels, (b) Model training: self-supervised cross-view prediction with covariance and signal-subspace alignment losses, and (c) Inference: learned covariance enhancement followed by standard MUSIC.

4.1 PAIRED OBSERVATION CONSTRUCTION

We first describe how JEPA-MUSIC derives self-supervised training signals without externally provided DoA labels. Specifically, the reliability of covariance and subspace estimates depends on the quality of the received array observations: more snapshots and a higher SNR generally yield more reliable estimates, whereas limited and noisy observations lead to greater estimation uncertainty. This motivates us to learn a mapping from a lower-quality observation to the spatial information represented more reliably by a higher-quality observation of the same scene. To this end, we construct paired context and target observations, using the higher-quality target as a training reference for covariance enhancement from the limited and noisy context.

Formally, let ξ denote the latent physical scene shared by a pair of observations. Conditioned on ξ, we sample a context observation and a target observation:

$$\mathbf{Y}_{\mathrm{c}} \sim p_{\mathrm{c}}(\cdot \mid \xi), \qquad \mathbf{Y}_{\mathrm{t}} \sim p_{\mathrm{t}}(\cdot \mid \xi), \tag{9}$$

where $\mathbf{Y}_{\mathrm{c}} \in \mathbb{C}^{M \times N_{\mathrm{c}}}$ and $\mathbf{Y}_{\mathrm{t}} \in \mathbb{C}^{M \times N_{\mathrm{t}}}$ denote the context and target observations, respectively, and $p_{\mathrm{c}}(\cdot \mid \boldsymbol{\xi})$ and $p_{\mathrm{t}}(\cdot \mid \boldsymbol{\xi})$ describe their conditional sampling distributions. Both observations correspond to the same source directions. The context observation represents the limited and noisy measurement available at inference, whereas the target observation is obtained with more snapshots ($N_{\mathrm{t}} > N_{\mathrm{c}}$) and a higher SNR and serves as a higher-quality reference used only during training.

In practice, such pairs can be constructed from repeated array measurements while the underlying scene remains approximately unchanged. This construction avoids reliance on external DoA-labeling systems, addressing the limitations discussed in Section 2. Practical construction of the paired observations is detailed in Appendix B.1.

To prepare the inputs for the context and target branches, both observations are converted into trace-normalized covariance representations:

$$\boldsymbol{R}_{\mathrm{c}} = \mathcal{T}(\boldsymbol{Y}_{\mathrm{c}}), \qquad \boldsymbol{R}_{\mathrm{t}} = \mathcal{T}(\boldsymbol{Y}_{\mathrm{t}}), \qquad \boldsymbol{R}_{\mathrm{c}}, \boldsymbol{R}_{\mathrm{t}} \in \mathbb{C}^{M \times M}, \tag{10}$$

where $\mathcal{T}(\cdot)$ computes the sample covariance according to Equation (5) and normalizes its trace to M. To enable processing by real-valued neural networks, each complex covariance is then represented by two real-valued channels:

$$\boldsymbol{U}_v = \mathrm{Stack}(\mathrm{Re}\{\boldsymbol{R}_v\}, \mathrm{Im}\{\boldsymbol{R}_v\}) \in \mathbb{R}^{2 \times M \times M}, \qquad v \in \{\mathrm{c}, \mathrm{t}\}, \tag{11}$$

where $\mathrm{Stack}(\cdot, \cdot)$ stacks the real and imaginary components along the channel dimension. The resulting representations $\boldsymbol{U}_{\mathrm{c}}$ and $\boldsymbol{U}_{\mathrm{t}}$ are passed to the context and target branches, respectively.

4.2 Self-Supervised Cross-View Prediction

The context branch comprises an encoder f_ω and a predictor g_ϕ. During training, a random subset of the scalar entries in the two-channel context representation is masked before encoding. The context branch produces

$$\boldsymbol{z}_{\mathrm{c}} = f_\omega(\widetilde{\boldsymbol{U}}_{\mathrm{c}}), \qquad \widehat{\boldsymbol{z}}_{\mathrm{t}} = g_\phi(\boldsymbol{z}_{\mathrm{c}}), \qquad \boldsymbol{z}_{\mathrm{c}}, \widehat{\boldsymbol{z}}_{\mathrm{t}} \in \mathbb{R}^q, \tag{12}$$

where $\widetilde{\boldsymbol{U}}_{\mathrm{c}}$ is the masked context representation, q is the latent dimension, and $\widehat{\boldsymbol{z}}_{\mathrm{t}}$ is the predicted target embedding. Meanwhile, the target branch encodes the unmasked target observation:

$$\boldsymbol{z}_{\mathrm{t}} = f_{\bar{\omega}}(\boldsymbol{U}_{\mathrm{t}}) \in \mathbb{R}^q. \tag{13}$$

Specifically, the target encoder parameters follow the context encoder through an exponential moving average (EMA) (Assran et al., 2023). Details of the masking strategy and EMA update are deferred to Appendix B.2.

Because the target observation provides a higher-quality observation of the same scene, predicting $\boldsymbol{z}_{\mathrm{t}}$ from the limited context observation encourages the predicted embedding to capture the spatial information represented more reliably by the target. However, $\widehat{\boldsymbol{z}}_{\mathrm{t}}$ remains an unconstrained latent vector and does not satisfy the algebraic structure required of a covariance matrix. We therefore introduce a structured covariance head that converts it into a valid representation for standard MUSIC inference.

4.3 MUSIC-Compatible Covariance Construction

As illustrated in Figure 2(b), the covariance head h_η maps the predicted target embedding to an unconstrained complex factor

$$\boldsymbol{B} = h_\eta(\widehat{\boldsymbol{z}}_{\mathrm{t}}) \in \mathbb{C}^{M \times M}. \tag{14}$$

Although $\boldsymbol{B}$ has the desired matrix dimension, it need not be Hermitian or positive semidefinite and therefore cannot be used directly as a covariance matrix. We instead treat $\boldsymbol{B}$ as an intermediate factor and construct

$$\widetilde{\boldsymbol{R}} = \frac{M\left(\boldsymbol{B}\boldsymbol{B}^H + \epsilon \boldsymbol{I}_M\right)}{\mathrm{Tr}\left(\boldsymbol{B}\boldsymbol{B}^H + \epsilon \boldsymbol{I}_M\right)}, \qquad \epsilon > 0, \tag{15}$$

where $\epsilon \boldsymbol{I}_M$ is a diagonal-loading term. This construction provides the following benefits. (i) The product $\boldsymbol{B}\boldsymbol{B}^H$ guarantees a Hermitian positive-semidefinite covariance, enabling the eigendecomposition required to construct the MUSIC signal and noise subspaces. (ii) The diagonal loading $\epsilon \boldsymbol{I}_M$ makes the covariance positive definite, preventing singular outputs and improving numerical stability. (iii) Trace normalization enforces $\mathrm{Tr}(\widetilde{\boldsymbol{R}}) = M$, placing the predicted and target covariances on the same scale without changing their eigenspaces. Consequently, every predicted covariance can be processed directly by standard MUSIC. The detailed justification is provided in Appendix B.3.

4.4 Self-Supervised Covariance and Subspace Learning

As illustrated in Figure 2, JEPA-MUSIC uses the higher-quality target observation to supervise covariance recovery from the limited context observation. This learning process must satisfy three requirements: (i) predicting the target representation, (ii) recovering the target-view spatial covariance, and (iii) preserving the signal subspace used by MUSIC. We therefore introduce three complementary loss terms.

Latent representation loss. Let $\mathrm{sg}(\cdot)$ denote the stop-gradient operation, which prevents gradients from propagating through the target representation. The latent prediction loss is defined as:

$$\mathcal{L}_{\mathrm{latent}} = \frac{1}{q} \left\| \widehat{\boldsymbol{z}}_{\mathrm{t}} - \mathrm{sg}(\boldsymbol{z}_{\mathrm{t}}) \right\|_2^2 . \tag{16}$$

Since $\boldsymbol{z}_{\mathrm{t}}$ is obtained from the higher-quality target observation, minimizing this loss trains the context branch to predict the target representation from the limited context observation. This constitutes the self-supervised cross-view learning objective of JEPA.

Covariance alignment loss. Latent prediction alone does not constrain the covariance decoded from $\widehat{\boldsymbol{z}}_{\mathrm{t}}$. We therefore use the higher-quality target covariance $\boldsymbol{R}_{\mathrm{t}}$ to supervise the covariance $\widetilde{\boldsymbol{R}}$ recovered from the context observation:

$$\mathcal{L}_{\mathrm{cov}} = \frac{1}{2M^2} \Big(\Big\| \mathrm{Re}\Big\{ \widetilde{\boldsymbol{R}} \Big\} - \mathrm{Re}\{\boldsymbol{R}_{\mathrm{t}}\} \Big\|_{\mathrm{F}}^2 + \Big\| \mathrm{Im}\Big\{ \widetilde{\boldsymbol{R}} \Big\} - \mathrm{Im}\{\boldsymbol{R}_{\mathrm{t}}\} \Big\|_{\mathrm{F}}^2 \Big). \tag{17}$$

Here, the real and imaginary components are aligned separately, consistent with the two-channel representation in Equation (11). Since each channel of an $M \times M$ covariance contains M^2 entries, the factor $1/(2M^2)$ averages the squared error over all $2M^2$ real-valued components.

Signal-subspace alignment loss. The entry-wise covariance loss does not explicitly target the eigenspace used by MUSIC. We therefore directly align the predicted and target signal subspaces. Let $\widetilde{\boldsymbol{E}}_{\mathrm{s}}, \boldsymbol{E}_{\mathrm{s,t}} \in \mathbb{C}^{M \times D}$ contain orthonormal eigenvectors associated with the D largest eigenvalues of $\widetilde{\boldsymbol{R}}$ and $\boldsymbol{R}_{\mathrm{t}}$, respectively. Their signal-subspace projectors can be expressed as $\widetilde{\boldsymbol{P}}_{\mathrm{s}} = \widetilde{\boldsymbol{E}}_{\mathrm{s}} \widetilde{\boldsymbol{E}}_{\mathrm{s}}^H$ and $\boldsymbol{P}_{\mathrm{s,t}} = \boldsymbol{E}_{\mathrm{s,t}} \boldsymbol{E}_{\mathrm{s,t}}^H$, respectively. As a result, we align these subspaces using

$$\mathcal{L}_{\mathrm{proj}} = \frac{1}{M^2} \left\| \widetilde{\boldsymbol{P}}_{\mathrm{s}} - \boldsymbol{P}_{\mathrm{s,t}} \right\|_{\mathrm{F}}^2 . \tag{18}$$

Projectors represent subspaces independently of eigenvector phase, ordering, or the choice of orthonormal basis. Aligning the signal-subspace projectors therefore also aligns their orthogonal complements. In other words, minimizing $\mathcal{L}_{\mathrm{proj}}$ also aligns the noise subspaces used to construct the MUSIC spectrum.

The three loss terms provide complementary self-supervision. Specifically, the latent loss establishes cross-view prediction, the covariance loss constrains the full spatial covariance and its eigenvalue structure, and the projector loss directly aligns the signal subspace used by MUSIC. Since different covariances can share the same projector, covariance alignment also constrains information left unspecified by projector alignment. Accordingly, the complete training objective is given by:

$$\mathcal{L} = \lambda_{\mathrm{z}} \mathcal{L}_{\mathrm{latent}} + \lambda_{\mathrm{R}} \mathcal{L}_{\mathrm{cov}} + \lambda_{\mathrm{P}} \mathcal{L}_{\mathrm{proj}}, \tag{19}$$

where λ_{z}, λ_{R}, and λ_{P} balance the three objectives.

At inference, the target branch and all training losses are removed, and context masking is disabled. The retained context encoder, predictor, and covariance head produce $\widetilde{\boldsymbol{R}}$, which replaces $\widehat{\boldsymbol{R}}_{\mathrm{y}}$ in the standard MUSIC pipeline described in Section 3.2.

4.5 Theoretical Analysis

We theoretically analyze how the training objectives support cross-view prediction and MUSIC inference in Appendix C. In particular, Proposition 1 characterizes the optimal latent predictor as a conditional mean, explaining how target-view randomness is averaged out. Proposition 2 and Corollary 1 connect the covariance loss and target-view estimation error to bounds on the recovered subspaces and MUSIC-spectrum denominator under an eigengap condition. We further show that the projector loss directly bounds alignment with the target signal subspace and provides a complementary MUSIC-denominator bound. Together, these results link the three training objectives to representation prediction, covariance recovery, and subspace alignment.

5 EXPERIMENTS

Experimental Setup.

We evaluate JEPA-MUSIC under noncoherent and fully coherent sources (Appendix A.3), varying SNR and snapshots to address the fundamental degradation regimes discussed in Section 3.2. Unless noted otherwise, the default configuration is an $M = 8$ ULA with $D = 3$ sources, $\mathrm{SNR} = 0$ dB, and $N = N_c = 32$ snapshots. We benchmark against classical MUSIC (Schmidt, 1986), FBSS-MUSIC (Shan et al., 1985; Pillai & Kwon, 1989), DeepCNN (Papageorgiou et al., 2021), DA-MUSIC (Merkofer et al., 2024), and SubspaceNet (Shmuel et al., 2025). We generate datasets using the signal model in Section 3.1 under noncoherent and fully coherent source conditions. Each dataset comprises 4,000 training, 500 validation, and 500 test samples, shared across all models. Learning-based results are averaged over five seeds (500 epochs each). We report angular RMSE, median and 90th-percentile absolute errors (Med. and P90, in degrees), and failure rate (Fail.). Further implementation details are detailed in Appendix D.

Main Results. We structure our evaluation around three questions. **Q1.** How does JEPA-MUSIC compare with state-of-the-art (SOTA) DoA estimators under limited snapshots, low SNR, and source coherence? **Q2.** Does JEPA-MUSIC provide MUSIC with a more accurate signal subspace? **Q3.** Can retaining training-free MUSIC inference enable JEPA-MUSIC to achieve accurate DoA estimation with fewer training samples?

Table 1: DoA estimation at $\mathrm{SNR} = 0$ dB with $N = 32$ snapshots. Best results are bold.

(a) Noncoherent, $\rho = 0$, SNR $= 0$ dB

Method	RMSE↓	Med.↓	P90↓	Fail.↓
Raw MUSIC	7.466	0.609	11.018	27.60
FBSS-MUSIC	5.674	0.734	4.156	17.20
DeepCNN	1.603	0.835	2.132	1.60
DA-MUSIC	2.520	1.515	4.001	17.92
SubspaceNet	10.541	7.704	17.295	95.12
JEPA-MUSIC	**1.012**	**0.477**	**1.249**	**0.36**

(b) Fully coherent, $\rho = 1$, SNR $= 0$ dB

Method	RMSE↓	Med.↓	P90↓	Fail.↓
Raw MUSIC	11.927	5.293	20.599	84.40
FBSS-MUSIC	7.977	1.325	10.444	36.20
DeepCNN	3.420	1.400	4.398	18.20
DA-MUSIC	3.157	1.581	4.443	20.12
SubspaceNet	11.198	8.155	17.931	95.08
JEPA-MUSIC	**2.915**	**0.781**	**2.533**	**6.80**

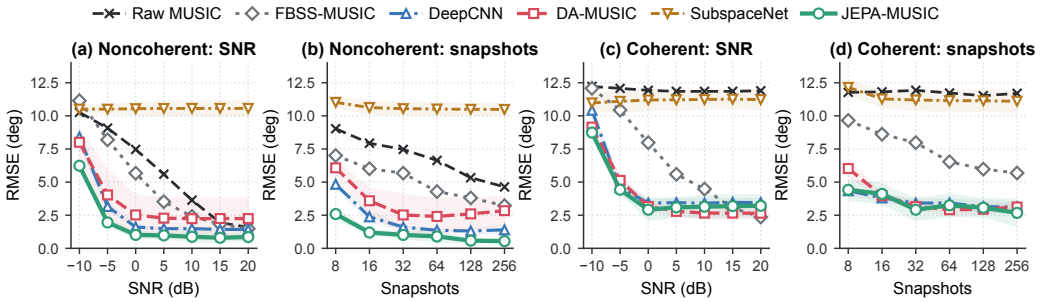

Figure 3: DoA estimation under varying SNRs and snapshot counts. SNR sweeps use 32 snapshots, while snapshot sweeps use $\mathrm{SNR} = 0$ dB.

Q1: Comparison with SOTA methods. Table 1 shows that JEPA-MUSIC achieves the best results across all four metrics in both source settings. For noncoherent sources, it reduces RMSE, P90, and failure rate by 36.9%, 41.4%, and 77.5% relative to DeepCNN, the strongest supervised baseline in this setting. Its RMSE is also 86.4% and 82.2% lower than Raw MUSIC and FBSS-MUSIC, respectively. For fully coherent sources, JEPA-MUSIC achieves an RMSE of $2.915°$, improving upon DA-MUSIC's $3.157°$ by 7.7%. Its failure rate of 6.80% is substantially lower than DeepCNN's 18.20% and DA-MUSIC's 20.12%, demonstrating gains in both estimation accuracy and reliability.

We further evaluate the trained models under varying SNRs and snapshot counts. Figures 3(a) and (b) show that JEPA-MUSIC achieves the lowest mean RMSE throughout the noncoherent sweeps. For fully coherent sources (Figure 3(c)), it also achieves the lowest mean RMSE at -10, -5, and 0 dB, highlighting its effectiveness when noise and source coherence jointly challenge MUSIC. Across snapshot counts, it remains competitive with supervised estimators and consistently outperforms classical MUSIC baselines. For fully coherent sources at higher SNRs (5–20 dB), it remains competitive with supervised estimators. Overall, these results demonstrate strong performance under limited and noisy observations.

Q2: Signal-subspace enhancement. MUSIC estimates DoAs from the subspace structure of its input covariance. We therefore compare the signal subspaces obtained from the observed and predicted covariances against the high-quality target-view reference. Table 2 shows that JEPA-MUSIC reduces the subspace error from $9.548°$ to $2.334°$ (75.6%) for noncoherent sources and from $7.527°$ to $2.164°$ (71.3%) for fully coherent sources. The improved subspace alignment is accompanied by DoA RMSE reductions of 86.4% over Raw MUSIC and 63.5% over FBSS-MUSIC in the corresponding settings.

Table 2: Signal-subspace error relative to the target-view reference and DoA RMSE at SNR $= 0$ dB with $N_c = 32$.

MUSIC input	Subspace error↓	RMSE↓
Noncoherent ($\rho = 0$)		
Raw covariance	9.548	7.466
Predicted covariance	**2.334**	**1.012**
Fully coherent ($\rho = 1$)		
FBSS covariance (see Appendix D.2)	7.527	7.977
Predicted covariance	**2.164**	**2.915**

These results support the central design of JEPA-MUSIC: the covariance learned through self-supervision enables standard MUSIC to extract a more accurate signal subspace, thereby improving DoA estimation.

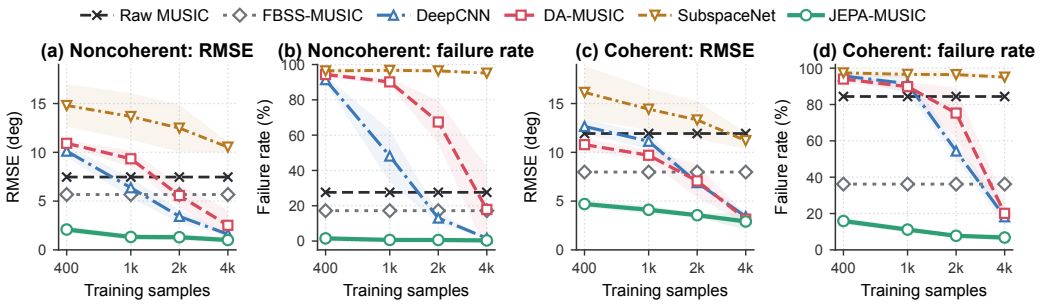

Figure 4: DoA estimation under varying numbers of training samples.

Q3: Training-sample efficiency. JEPA-MUSIC learns the covariance representation while retaining the training-free MUSIC procedure for direction estimation. We therefore examine whether the framework achieves accurate DoA estimation with fewer training samples. As shown in Figure 4, with only 400 unlabeled context–target pairs, JEPA-MUSIC achieves RMSEs of $2.088°$ and $4.701°$ for noncoherent and fully coherent sources, respectively. Both errors are less than half those of the strongest supervised baseline trained with the same number of samples in each setting. Across all evaluated training-set sizes, JEPA-MUSIC achieves the lowest mean RMSE in both source settings. These results highlight its training-sample efficiency, with particularly pronounced advantages when training data are limited.

More Experiments and Robustness Studies. We provide additional evaluations in Appendix E. Specifically, Appendix E.1 examines architectural components and training objectives through ablation studies; Appendix E.2 evaluates generalization to other source-signal distributions without retraining or fine-tuning; Appendix E.3 demonstrates transferability to the DeepMIMO dataset (Alkhateeb, 2019); and Appendix E.4 investigates SubspaceNet's performance with increased training data, providing further evidence of the training-sample efficiency of JEPA-MUSIC. Together, these results support the design choices, robustness, and generalization capabilities of our self-supervised representation learning framework.

6 Conclusion

In this work, we have proposed JEPA-MUSIC, a model-based self-supervised framework that bridges predictive representation learning with classical subspace-based DoA estimation. By exploiting cross-view predictability between paired observations of the same physical scene, JEPA-MUSIC learns to reconstruct a structured, MUSIC-compatible covariance without relying on ground-truth DoA labels. Moreover, our theoretical analysis establishes rigorous upper bounds that link covariance recovery error to perturbations in the signal-noise subspaces and the MUSIC pseudospectrum denominator, providing formal justification for the proposed self-supervised objective. Extensive evaluations confirm that JEPA-MUSIC substantially outperforms classical MUSIC in challenging regimes and rivals or surpasses strong supervised baselines, exhibiting substantial advantages in suppressing tail failure rates, enhancing subspace fidelity, and maintaining high sample efficiency.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript drafting and editing. The authors reviewed the final text and take responsibility for its content.

REFERENCES

Hervé Abdi. The eigen-decomposition: Eigenvalues and eigenvectors. *Encyclopedia of measurement and statistics*, 304, 2007.

A. Alkhateeb. DeepMIMO: A generic deep learning dataset for millimeter wave and massive MIMO applications. In *Proc. of Information Theory and Applications Workshop (ITA)*, pp. 1–8, San Diego, CA, Feb 2019.

Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *European conference on computer vision*, pp. 456–473. Springer, 2022.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

Andreas Barthelme and Wolfgang Utschick. DoA estimation using neural network-based covariance matrix reconstruction. *IEEE Signal Processing Letters*, 28:783–787, 2021. doi: 10.1109/LSP.2021.3072564.

I. Bekkerman and J. Tabrikian. Target detection and localization using MIMO radars and sonars. *IEEE Transactions on Signal Processing*, 54(10):3873–3883, 2006. doi: 10.1109/TSP.2006.879267.

Rajendra Bhatia. *Matrix Analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer, New York, NY, 1997. doi: 10.1007/978-1-4612-0653-8.

Soumitro Chakrabarty and Emanuël A. P. Habets. Multi-speaker DOA estimation using deep convolutional networks trained with noise signals. *IEEE Journal of Selected Topics in Signal Processing*, 13(1):8–21, 2019. doi: 10.1109/JSTSP.2019.2901664.

Delong Chen, Mustafa Shukor, Theo Moutakanni, Willy Chung, Lei Yu, Tejaswi Kasarla, Allen Bolourchi, Yann LeCun, and Pascale Fung. VL-JEPA: Joint embedding predictive architecture for vision-language. In *International Conference on Learning Representations*, volume 2026, pp. 89791–89807, 2026.

Yong Chu, Xun Zhou, Zenglin Xu, Hui Wang, and Yue Yu. Map as a prompt: Learning multi-modal spatial-signal foundation models for cross-scenario wireless localization. In *International Conference on Learning Representations*, 2026.

Chandler Davis and William M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970. doi: 10.1137/0707001.

Priyadarshini Dwivedi, Gyanajyoti Routray, and Rajesh M Hegde. Sparse bayesian integrated cnn framework for enhanced acoustic source localization. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Ahmet M. Elbir. DeepMUSIC: Multiple signal classification via deep learning. *IEEE Sensors Letters*, 4(4):1–4, 2020. doi: 10.1109/LSENS.2020.2980384.

Christine Evers, Heinrich W Löllmann, Heinrich Mellmann, Alexander Schmidt, Hendrik Barfuss, Patrick A Naylor, and Walter Kellermann. The locata challenge: Acoustic source localization and tracking. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1620–1643, 2020.

Lajos Hanzo, SX Ng, T Keller, and W Webb. *Quadrature amplitude modulation*. Chichester, UK: Wiley, 2004.

Hai Huang, Yann LeCun, and Randall Balestriero. LLM-JEPA: Large language models meet joint embedding predictive architectures. In *The Fourteenth International Conference on Learning Representations*, 2026.

Anders E. Kalør, Giuseppe Durisi, Sinem Coleri, Stefan Parkvall, Wei Yu, Andreas Mueller, and Petar Popovski. Wireless 6g connectivity for massive number of devices and critical services. *Proceedings of the IEEE*, 113(9):826–848, 2025. doi: 10.1109/JPROC.2024.3484529.

Hamid Krim and Mats Viberg. Two decades of array signal processing research: The parametric approach. *IEEE Signal Processing Magazine*, 13(4):67–94, 1996. doi: 10.1109/79.526899.

Saanvi Kulkarni, Avani Thakur, Samraddhi Soni, Anil Hiwale, Madhavi H Belsare, and AA Bazil Raj. A comprehensive review of direction of arrival (DoA) estimation techniques and algorithms. *Journal of Electronics and Electrical Engineering*, pp. 140–189, 2025.

Xianhang Li, Chen Huang, Chun-Liang Li, Eran Malach, Joshua M. Susskind, Vimal Thilak, and Etai Littwin. Rethinking JEPA: Compute-efficient video self-supervised learning with frozen teachers. In *The Fourteenth International Conference on Learning Representations*, 2026.

Shining Lin, Jiancheng An, Lu Gan, Victor C. M. Leung, Mehdi Bennis, Mérouane Debbah, and Tie Jun Cui. Electromagnetic neural network for direction-of-arrival estimation. *IEEE Transactions on Wireless Communications*, 25:21306–21322, 2026. doi: 10.1109/TWC.2026.3717644.

Yaoqi Liu, Jin Wang, Chunchen Wang, Hui Wang, and Chuan Shi. From denoising to de-channeling: Integrating physical channel priors into diffusion models for radio signal understanding. In *International Conference on Machine Learning*, 2026a.

Yaoqi Liu, Jin Wang, Hui Wang, and Chuan Shi. TS-DDAE: A novel temporal-spectral denoising diffusion autoencoder for wireless signal recognition model pre-training. In *International Conference on Learning Representations*, 2026b.

Zhang-Meng Liu, Chenwei Zhang, and Philip S. Yu. Direction-of-arrival estimation based on deep neural networks with robustness to array imperfections. *IEEE Transactions on Antennas and Propagation*, 66(12):7315–7327, 2018. doi: 10.1109/TAP.2018.2874430.

Xuefei Ma, Yuanhui Xiang, Haifeng Zhu, Rahim Khan, Yanni Wu, Chen Wang, Wanyan Wang, Hengliang Wu, and Zhiqiang Wu. Joint DOD and DOA estimation for bistatic MIMO sonar based on reduced-order regularized mfocuss. *Signal, Image and Video Processing*, 19(2):194, 2025.

Julian P. Merkofer, Guy Revach, Nir Shlezinger, Tirza Routtenberg, and Ruud J. G. van Sloun. DA-MUSIC: Data-driven DoA estimation via deep augmented MUSIC algorithm. *IEEE Transactions on Vehicular Technology*, 73(2):2771–2785, 2024. doi: 10.1109/TVT.2023.3320360.

Amir Masoud Molaei, Bijan Zakeri, Seyed Mehdi Hosseini Andargoli, Muhammad Ali Babar Abbasi, Vincent Fusco, and Okan Yurduseven. A comprehensive review of direction-of-arrival estimation and localization approaches in mixed-field sources scenario. *IEEE Access*, 12:65883–65918, 2024. doi: 10.1109/ACCESS.2024.3398351.

Amir Masoud Molaei, Vincent Fusco, Muhammad Ali Babar Abbasi, Mohsen Khalily, Cássio G. Lopes, and Okan Yurduseven. Efficient active azimuth-elevation DOA estimations via massive MIMO radar. *IEEE Transactions on Vehicular Technology*, 74(12):19217–19236, 2025. doi: 10.1109/TVT.2025.3586737.

Andreas F Molisch. *Wireless communications: from fundamentals to beyond 5G*. John Wiley & Sons, 2022.

Heejeong Nam, Quentin Le Lidec, Lucas Maes, Yann LeCun, and Randall Balestriero. Causal-JEPA: Learning world models through object-level latent masking. In *Forty-third International Conference on Machine Learning*, 2026.

Georgios K. Papageorgiou, Mathini Sellathurai, and Yonina C. Eldar. Deep networks for direction-of-arrival estimation in low snr. *IEEE Transactions on Signal Processing*, 69:3714–3729, 2021. doi: 10.1109/TSP.2021.3089927.

S. U. Pillai and B. H. Kwon. Forward/backward spatial smoothing techniques for coherent signal identification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(1):8–15, 1989. doi: 10.1109/29.17496.

Richard A Roberts and Clifford T Mullis. *Digital signal processing*. Addison-Wesley Longman Publishing Co., Inc., 1987.

Ralph O. 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.

T.-J. Shan, Mati Wax, and Thomas Kailath. On spatial smoothing for direction-of-arrival estimation of coherent signals. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(4):806–811, 1985. doi: 10.1109/TASSP.1985.1164649.

Xiaohong Sheng and Yu-Hen Hu. Maximum likelihood multiple-source localization using acoustic energy measurements with wireless sensor networks. *IEEE transactions on signal processing*, 53 (1):44–53, 2005.

Dor H. Shmuel, Julian P. Merkofer, Guy Revach, Ruud J. G. van Sloun, and Nir Shlezinger. SubspaceNet: Deep learning-aided subspace methods for DoA estimation. *IEEE Transactions on Vehicular Technology*, 74(3):4962–4976, 2025. doi: 10.1109/TVT.2024.3496119.

P. Stoica and R.L. Moses. *Spectral Analysis of Signals*. Pearson Prentice Hall, 2005. ISBN 9780131139565.

Yunye Su, Xianpeng Wang, Linqiang Wen, Yuehao Guo, Han Wang, and Gaojie Chen. Cascaded multi-head attention transformer framework for direction of arrival estimation. *IEEE Transactions on Cognitive Communications and Networking*, 12:10450–10465, 2026. doi: 10.1109/TCCN.2026.3712135.

Harry L. Van Trees. *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. John Wiley & Sons, 2002. doi: 10.1002/0471221104.

Liuli Wu, Zhang-Meng Liu, and Zhi-Tao Huang. Deep convolution network for direction of arrival estimation with sparse prior. *IEEE Signal Processing Letters*, 26(11):1688–1692, 2019. doi: 10.1109/LSP.2019.2945115.

He Xu, Wei Liu, Ming Jin, and Ye Tian. Positioning and contour extraction of autonomous vehicles based on enhanced DOA estimation by large-scale arrays. *IEEE Internet of Things Journal*, 10 (13):11792–11803, 2023. doi: 10.1109/JIOT.2023.3244861.

A Appendix 1: More Signal Processing Details

A.1 Interpretation of the Wireless Signal and Steering Vector

In digital wireless systems, modulation maps source information to complex-valued symbols. The source signal $\mathsf{s}_d(n)$ is represented in complex baseband form as

$$\mathsf{s}_d(n) = \mathsf{I}_d(n) + j\mathsf{Q}_d(n), \tag{20}$$

where $\mathsf{I}_d(n)$ and $\mathsf{Q}_d(n)$ denote its in-phase (I) and quadrature (Q) components, respectively. Together, they determine the magnitude and phase of the complex signal (Roberts & Mullis, 1987). Accordingly, each antenna receives a contribution from this same source signal, with a relative phase determined by the arrival direction.

To see how these phase differences are represented, Euler's formula gives

$$e^{-j\phi} = \cos\phi - j\sin\phi. \tag{21}$$

Because $|e^{-j\phi}| = 1$, multiplying a complex signal by $e^{-j\phi}$ rotates its phase by $-\phi$ while preserving its magnitude. This rotation changes its I and Q components without changing its overall magnitude.

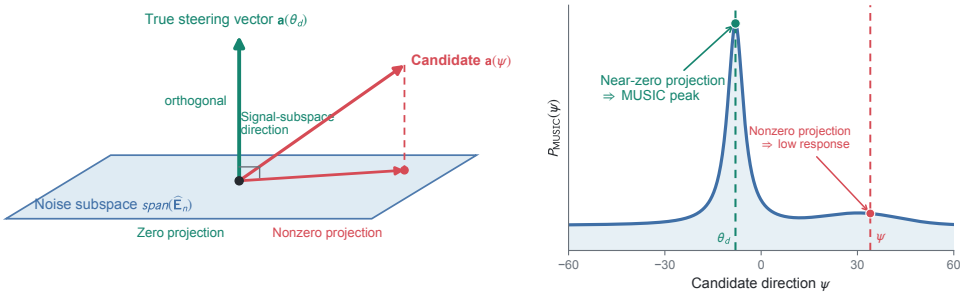

Figure 5: Geometric interpretation of MUSIC orthogonality. A true steering vector is orthogonal to the noise subspace and therefore produces a near-zero projection, resulting in a peak in the MUSIC spectrum. An incorrect candidate direction generally produces a nonzero projection and a lower spectrum response.

For the half-wavelength-spaced ULA, the steering-vector coefficient

$$a_m(\theta_d) = e^{-j\pi(m-1)\sin\theta_d} \tag{22}$$

describes the phase shift of source signal $\mathsf{s}_d(n)$ at antenna m relative to the first antenna, with θ_d measured from the array broadside. Consequently,

$$\boldsymbol{a}(\theta_d)\mathsf{s}_d(n) = \left[e^{-j\pi(1-1)\sin\theta_d}\mathsf{s}_d(n), e^{-j\pi(2-1)\sin\theta_d}\mathsf{s}_d(n), \ldots, e^{-j\pi(M-1)\sin\theta_d}\mathsf{s}_d(n)\right]^T. \tag{23}$$

This vector describes the contribution of one source across all antennas, rather than the complete received observation, which also contains other sources and noise. Since the relative phase shifts depend on θ_d, the steering vector encodes the source direction in the array measurements. The real and imaginary parts of each received sample $\mathsf{y}_m(n)$ correspond to the receiver's I and Q components.

A.2 Details of the MUSIC Eigendecomposition

The sample covariance $\widehat{\mathbf{R}}_{\mathrm{y}}$ is Hermitian and positive semidefinite. It therefore admits the eigendecomposition

$$\widehat{\mathbf{R}}_{\mathrm{y}} = \widehat{\mathbf{E}}\widehat{\boldsymbol{\Lambda}}\widehat{\mathbf{E}}^H, \tag{24}$$

where

$$\widehat{\mathbf{E}} = [\widehat{\mathbf{e}}_1, \ldots, \widehat{\mathbf{e}}_M], \qquad \widehat{\boldsymbol{\Lambda}} = \mathrm{diag}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_M). \tag{25}$$

Each eigenpair satisfies

$$\widehat{\mathbf{R}}_{\mathrm{y}}\widehat{\mathbf{e}}_m = \widehat{\lambda}_m\widehat{\mathbf{e}}_m, \qquad m = 1, \ldots, M. \tag{26}$$

The eigenvector $\widehat{\mathbf{e}}_m$ represents a spatial mode of the received signals, while $\widehat{\lambda}_m$ measures the observed power along that mode. Since $\widehat{\mathbf{R}}_{\mathrm{y}}$ is Hermitian, its eigenvalues are real and nonnegative, and its eigenvectors can be chosen to satisfy $\widehat{\mathbf{E}}^H\widehat{\mathbf{E}} = \boldsymbol{I}_M$.

MUSIC orders the eigenvalues in descending order:

$$\widehat{\lambda}_1 \geqslant \cdots \geqslant \widehat{\lambda}_D \geqslant \widehat{\lambda}_{D+1} \geqslant \cdots \geqslant \widehat{\lambda}_M. \tag{27}$$

Under the assumed D-source model, the first D spatial modes are treated as signal-dominated, while the remaining $M - D$ modes are treated as noise-dominated. Accordingly,

$$\widehat{\mathbf{E}}_{\mathrm{s}} = [\widehat{\mathbf{e}}_1, \ldots, \widehat{\mathbf{e}}_D], \qquad \widehat{\boldsymbol{\Lambda}}_{\mathrm{s}} = \mathrm{diag}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_D), \tag{28}$$

$$\widehat{\mathbf{E}}_{\mathrm{n}} = [\widehat{\mathbf{e}}_{D+1}, \ldots, \widehat{\mathbf{e}}_M], \qquad \widehat{\boldsymbol{\Lambda}}_{\mathrm{n}} = \mathrm{diag}(\widehat{\lambda}_{D+1}, \ldots, \widehat{\lambda}_M). \tag{29}$$

This partition gives $\widehat{\mathbf{E}}_{\mathrm{s}} \in \mathbb{C}^{M\times D}$ and $\widehat{\mathbf{E}}_{\mathrm{n}} \in \mathbb{C}^{M\times(M-D)}$. The two terms in Equation (6) represent the portions of the sample covariance associated with the estimated signal and noise subspaces, respectively.

Under ideal conditions, the steering vectors of the true sources lie in the signal subspace and are orthogonal to the noise subspace. As illustrated in Figure 5, a true steering vector therefore has a near-zero projection onto the noise subspace, whereas an incorrect candidate direction generally produces a nonzero projection. The corresponding projection energy is

$$\left\| \widehat{\mathbf{E}}_{\mathrm{n}}^{H} \boldsymbol{a}(\psi) \right\|_2^2 = \boldsymbol{a}^H(\psi) \widehat{\mathbf{E}}_{\mathrm{n}} \widehat{\mathbf{E}}_{\mathrm{n}}^{H} \boldsymbol{a}(\psi), \tag{30}$$

which forms the denominator of the MUSIC pseudospectrum (Equation (8)). MUSIC exploits this difference to identify the source directions. With finite and noisy observations, errors in the sample covariance perturb the eigenvalue ordering and the corresponding eigenvectors, which can weaken this orthogonality and distort the estimated MUSIC spectrum.

A.3 Effects of Nonideal Observations on MUSIC

The effects of limited snapshots, low SNR, and source coherence can be understood through their influence on the covariance eigenspaces and the MUSIC spectrum.

Limited snapshots. The sample covariance can be expressed as

$$\widehat{\mathbf{R}}_{\mathrm{y}} = \boldsymbol{R}_{\mathrm{y}} + \boldsymbol{\Delta}_N, \tag{31}$$

where $\boldsymbol{\Delta}_N$ denotes the finite-sample covariance estimation error. Since $\widehat{\mathbf{R}}_{\mathrm{y}}$ is estimated from only N snapshots, a small N can produce substantial deviations from the population covariance $\boldsymbol{R}_{\mathrm{y}}$. These deviations perturb the covariance eigenvalues and eigenvectors, reducing the accuracy of the estimated signal and noise subspaces. With independent snapshots, increasing N generally stabilizes the sample covariance and reduces its estimation error. However, at a fixed snapshot acquisition rate, collecting more snapshots also lengthens the observation interval and delays each DoA update. The snapshot number therefore introduces a trade-off between covariance accuracy and temporal responsiveness.

Low SNR. Using the observation model in Equation (3), and assuming that the source signals and noise are uncorrelated, the population covariance can be written as

$$\boldsymbol{R}_{\mathrm{y}} = \underbrace{\boldsymbol{A}(\boldsymbol{\theta}) \boldsymbol{R}_{\mathrm{s}} \boldsymbol{A}^H(\boldsymbol{\theta})}_{\text{signal component}} + \underbrace{\sigma_{\mathrm{w}}^2 \boldsymbol{I}_M}_{\text{noise component}}, \tag{32}$$

where

$$\boldsymbol{R}_{\mathrm{s}} = \mathbb{E}\left[\frac{1}{N} \mathbf{S}\mathbf{S}^H \right] \tag{33}$$

is the source covariance and σ_{w}^2 is the noise power.

Ideally, the D largest eigenvalues of $\boldsymbol{R}_{\mathrm{y}}$ correspond to source-related spatial modes, while the remaining $M - D$ eigenvalues correspond to the noise floor. Let these eigenvalues be ordered as $\lambda_1 \geqslant \cdots \geqslant \lambda_M$. Their boundary is characterized by the eigengap

$$\delta = \lambda_D - \lambda_{D+1}, \tag{34}$$

where λ_D is the weakest signal-related eigenvalue and λ_{D+1} is the largest noise-related eigenvalue.

At high SNR, the signal-related eigenvalues remain clearly distinguishable from the noise floor. At low SNR, the signal component becomes weak relative to the noise component, and finite-sample fluctuations can become comparable to the separation between the two eigenvalue groups. Consequently, covariance estimation errors can more easily perturb the eigenvalue ordering and rotate the associated eigenvectors, leading to inaccurate signal and noise subspaces and distorted MUSIC-spectrum peaks. This sensitivity is governed by the covariance perturbation relative to the eigengap, as characterized by classical subspace perturbation theory (Davis & Kahan, 1970).

Source coherence. For noncoherent sources, the signals $\mathsf{s}_1(n), \ldots, \mathsf{s}_D(n)$ vary independently of one another across snapshots, allowing $\mathbf{S}$ to contain D distinct signal components. Fully coherent sources instead carry scaled versions of the same underlying waveform. This commonly occurs when multiple propagation paths originate from the same transmitted signal.

Accordingly, the degree of source coherence can be represented by a correlation coefficient $\rho \in [0, 1]$:

$$\mathsf{s}_d(n) = \sqrt{\rho}\,\beta_d \mathsf{u}(n) + \sqrt{1-\rho}\,\mathsf{v}_d(n), \qquad d = 1, \ldots, D, \tag{35}$$

where $\mathsf{u}(n)$ is a waveform shared by all sources, $\mathsf{v}_d(n)$ is a source-specific waveform independent of the other sources, and β_d is a complex gain. When $\rho = 0$, the shared component disappears and the sources are noncoherent. When $\rho = 1$, the independent components disappear, and all sources become scaled versions of the shared waveform:

$$\mathsf{s}_d(n) = \beta_d \mathsf{u}(n). \tag{36}$$

Let $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_D]^T$ and $\mathbf{u} = [\mathsf{u}(1), \ldots, \mathsf{u}(N)]^T$. Under full coherence, the source matrix in Equation (3) becomes

$$\mathbf{S} = \boldsymbol{\beta}\mathbf{u}^T. \tag{37}$$

Although $\mathbf{S}$ contains D rows corresponding to D physical source directions, all rows are scaled versions of the same waveform. Therefore, $\mathbf{S}$ has rank one rather than rank D. Its covariance is

$$\boldsymbol{R}_{\mathrm{s}} = \mathbb{E}\left[\frac{1}{N}\mathbf{S}\mathbf{S}^H\right] = \mathbb{E}\left[\frac{\|\mathbf{u}\|_2^2}{N}\right]\boldsymbol{\beta}\boldsymbol{\beta}^H, \tag{38}$$

which also has rank one.

Consequently, although D physical directions are present, the array covariance contains only one independent signal mode. Its EVD therefore cannot recover the D eigenvectors required to represent all source steering vectors. This breaks the signal–noise subspace separation used by MUSIC and can cause source peaks to merge or disappear (Shan et al., 1985). When $0 < \rho < 1$, the sources are partially coherent. Increasing ρ weakens their independent components and makes the corresponding signal subspace more difficult to estimate.

Effect on the MUSIC spectrum. For a candidate direction ψ, the denominator of the MUSIC spectrum is the noise-subspace projection energy

$$\mathsf{q}(\psi) = \left\| \widehat{\mathbf{E}}_{\mathrm{n}}^H \boldsymbol{a}(\psi) \right\|_2^2. \tag{39}$$

Ideally, $\mathsf{q}(\theta_d)$ is close to zero at each true direction. Errors or rank deficiency in the estimated covariance can increase this value at a true direction, suppressing or shifting its spectrum peak. They can also produce a small projection at an incorrect direction, resulting in merged or spurious peaks. These mechanisms explain how the three nonideal conditions propagate from covariance estimation to the final DoA estimates.

JEPA-MUSIC is designed to mitigate these effects by learning from a higher-quality paired view, explicitly aligning the MUSIC-relevant subspace, and applying spatial smoothing when coherent sources are present.

B Appendix 2: Additional Details of JEPA-MUSIC

B.1 Practical Construction of Paired Observations

A central advantage of JEPA-MUSIC is that constructing paired context and target observations does not require external positioning sensors or calibrated optical systems for acquiring ground-truth DoA labels. Instead, the framework exploits the temporal redundancy available in multi-channel wireless measurements. This advantage concerns DoA-label acquisition; the downstream MUSIC decoder still assumes that the array manifold is known or adequately calibrated.

Specifically, within the channel coherence time, or during a quasi-static period in which the source directions and spatial second-order statistics remain approximately stationary (Molisch, 2022), an antenna array collects a sequence of consecutive baseband snapshots:

$$\boldsymbol{Y}_{\mathrm{all}} = [\boldsymbol{y}(1), \ldots, \boldsymbol{y}(N_{\mathrm{all}})] \in \mathbb{C}^{M \times N_{\mathrm{all}}}. \tag{40}$$

One practical realization is to partition this sequence into two disjoint snapshot-index sets, $\mathcal{I}_{\mathrm{c}}$ and $\mathcal{I}_{\mathrm{t}}$, to form the context and target observations:

$$\boldsymbol{Y}_{\mathrm{c}} = \big[\boldsymbol{y}(n)\big]_{n \in \mathcal{I}_{\mathrm{c}}} \in \mathbb{C}^{M \times N_{\mathrm{c}}}, \qquad \boldsymbol{Y}_{\mathrm{t}} = \big[\boldsymbol{y}(n)\big]_{n \in \mathcal{I}_{\mathrm{t}}} \in \mathbb{C}^{M \times N_{\mathrm{t}}}, \tag{41}$$

subject to

$$\mathcal{I}_{\mathrm{c}} \cap \mathcal{I}_{\mathrm{t}} = \varnothing, \qquad |\mathcal{I}_{\mathrm{c}}| = N_{\mathrm{c}}, \qquad |\mathcal{I}_{\mathrm{t}}| = N_{\mathrm{t}}. \tag{42}$$

The asymmetric fidelity between the two observations can be established along the following practical dimensions:

- **Snapshot-count asymmetry ($N_{\mathrm{t}} > N_{\mathrm{c}}$).** The target observation contains more snapshots and therefore generally provides a lower-variance estimate of the same population covariance, provided that both observations satisfy the above stationarity condition. The disjoint partition eliminates exact snapshot reuse and reduces direct sample leakage between the two predictive views. When the source and receiver-noise processes are temporally independent or sufficiently mixing, the two subsets also contain distinct observation-specific realizations.
- **SNR asymmetry ($\mathrm{SNR}_{\mathrm{t}} \geqslant \mathrm{SNR}_{\mathrm{c}}$).** When a cleaner repeated measurement of the same scene is available, it can serve as the higher-SNR target observation. Possible examples include lower-interference sensing intervals or repeated beacon and pilot intervals that preserve the same active-source configuration, relative source powers, and spatial statistics. Alternatively, when the original observations have the same physical SNR, this asymmetry can be synthesized without DoA labels by injecting controlled independent noise into $\boldsymbol{Y}_{\mathrm{c}}$ while leaving $\boldsymbol{Y}_{\mathrm{t}}$ unchanged.

Thus, the proposed pairing strategy provides a practical source of self-supervision without requiring ground-truth DoA labels.

B.2 Masking and EMA Update

Context masking. Let $\mathrm{vec}(\boldsymbol{U}_{\mathrm{c}}) \in \mathbb{R}^{2M^2}$ denote the flattened two-channel context representation. At each training iteration, we independently sample a binary mask $\mathbf{m} \in \{0,1\}^{2M^2}$ with

$$\mathsf{m}_i \sim \mathrm{Bernoulli}(1 - r_{\mathrm{mask}}), \tag{43}$$

where r_{mask} is the masking ratio. The masked context observation is constructed as

$$\widetilde{\boldsymbol{U}}_{\mathrm{c}} = \mathrm{unvec}(\boldsymbol{m} \odot \mathrm{vec}(\boldsymbol{U}_{\mathrm{c}}))\,, \tag{44}$$

where $\odot$ denotes element-wise multiplication. Masked entries are replaced with zero, and a new mask is sampled for each observation and training iteration. Masking is applied only to the context observation, while the target observation remains unmasked. It encourages the context encoder to infer the missing spatial information from the remaining covariance structure. Masking is disabled during validation and inference.

EMA target encoder. The target encoder is initialized as a copy of the context encoder, $\bar{\omega} \leftarrow \omega$, and is not updated by gradient descent. After each gradient update of the context encoder, its parameters are updated as

$$\bar{\omega} \leftarrow \mu\bar{\omega} + (1-\mu)\omega, \tag{45}$$

where $\mu \in [0,1)$ is the momentum coefficient. This exponential moving average provides a slowly evolving target representation for cross-view prediction (Assran et al., 2022; 2023). The numerical values of r_{mask} and μ are reported in Appendix D.

B.3 Validity of the Covariance Construction

We first review the structural properties of a valid spatial covariance and then show that Equation (15) satisfies these properties while preserving the eigenspaces required by MUSIC.

Specifically, standard MUSIC operates on the spatial covariance (Schmidt, 1986; Stoica & Moses, 2005):

$$\boldsymbol{R}_{\mathrm{y}} = \mathbb{E}\big[\mathbf{y}(n)\mathbf{y}^H(n)\big]\,. \tag{46}$$

Since each outer product $\mathbf{y}(n)\mathbf{y}^H(n)$ is Hermitian, we have:

$$\boldsymbol{R}_{\mathrm{y}}^H = \mathbb{E}\Big[\big(\mathbf{y}(n)\mathbf{y}^H(n)\big)^H\Big] = \boldsymbol{R}_{\mathrm{y}}. \tag{47}$$

Accordingly, for any $\boldsymbol{x} \in \mathbb{C}^M$, we have:

$$\boldsymbol{x}^H \boldsymbol{R}_{\mathrm{y}} \boldsymbol{x} = \mathbb{E}\Big[\big|\boldsymbol{x}^H \mathbf{y}(n)\big|^2\Big] \geqslant 0. \tag{48}$$

Therefore, a valid covariance matrix is necessarily Hermitian positive semidefinite. The Hermitian property guarantees real eigenvalues and orthonormal eigenvectors, allowing MUSIC to construct orthogonal signal and noise subspaces. Positive semidefiniteness guarantees nonnegative eigenvalues, which represent the powers of the corresponding spatial components.

We now verify that the predicted covariance in Equation (15) satisfies these properties. For any $\boldsymbol{B} \in \mathbb{C}^{M \times M}$, we have:

$$\big(\boldsymbol{B}\boldsymbol{B}^H\big)^H = \boldsymbol{B}\boldsymbol{B}^H \tag{49}$$

and

$$\boldsymbol{x}^H \boldsymbol{B}\boldsymbol{B}^H \boldsymbol{x} = \big\|\boldsymbol{B}^H \boldsymbol{x}\big\|_2^2 \geqslant 0. \tag{50}$$

Hence, $\boldsymbol{B}\boldsymbol{B}^H$ is Hermitian positive semidefinite. After diagonal loading, every nonzero $\boldsymbol{x} \in \mathbb{C}^M$ satisfies:

$$\boldsymbol{x}^H \left(\boldsymbol{B}\boldsymbol{B}^H + \epsilon \boldsymbol{I}_M\right) \boldsymbol{x} = \big\|\boldsymbol{B}^H \boldsymbol{x}\big\|_2^2 + \epsilon \|\boldsymbol{x}\|_2^2 > 0. \tag{51}$$

Thus, $\boldsymbol{B}\boldsymbol{B}^H + \epsilon \boldsymbol{I}_M$ is Hermitian positive definite. Its trace satisfies:

$$\mathrm{Tr}\big(\boldsymbol{B}\boldsymbol{B}^H + \epsilon \boldsymbol{I}_M\big) = \|\boldsymbol{B}\|_{\mathrm{F}}^2 + \epsilon M > 0, \tag{52}$$

so the normalization in Equation (15) remains well defined even when $\boldsymbol{B}$ is zero or rank deficient.

Finally, we obtain:

$$\mathrm{Tr}(\widetilde{\boldsymbol{R}}) = \frac{M \,\mathrm{Tr}\big(\boldsymbol{B}\boldsymbol{B}^H + \epsilon \boldsymbol{I}_M\big)}{\mathrm{Tr}(\boldsymbol{B}\boldsymbol{B}^H + \epsilon \boldsymbol{I}_M)} = M. \tag{53}$$

Adding $\epsilon \boldsymbol{I}_M$ shifts all eigenvalues by the same positive amount, while trace normalization rescales them by the same positive factor. Neither operation changes the eigenvectors. Consequently, $\widetilde{\boldsymbol{R}}$ is a valid normalized covariance whose eigenvectors can be partitioned into the signal and noise subspaces required by standard MUSIC.

C Appendix 3: Theoretical Analysis and Proofs

This appendix provides the formal assumptions, statements, and proofs for the theoretical analysis summarized in Section 4.5. The first result analyzes the latent cross-view objective in Equation (16) and characterizes how its optimal predictor emphasizes scene-consistent information while suppressing observation-specific variations. The second result then connects the learned representation to MUSIC. It shows that the covariance objective in Equation (17), together with the target-view estimation error, bounds the recovered covariance error and its propagation to the signal subspace and MUSIC-spectrum denominator. Since covariance alignment does not directly constrain the eigenspace used by MUSIC, we further show that the projector objective in Equation (18) provides complementary direct control of the signal-subspace error.

Latent cross-view prediction. We first analyze how the latent objective in Equation (16) promotes scene-consistent learning from paired observations. The following proposition characterizes its optimal predictor.

Proposition 1 (Optimal cross-view predictor)**.** *Let $\mathbf{z}_{\mathrm{c}}$ and $\mathbf{z}_{\mathrm{t}}$ denote the context and target embeddings defined in Equations* (12) *and* (13)*, respectively. To characterize the predictor independently of its neural parameterization, we treat both encoders as fixed in this analysis. Let $\boldsymbol{\xi}$ denote the scene information shared by the two views. Assume that $\mathbb{E}[\|\mathbf{z}_{\mathrm{t}}\|_2^2] < \infty$ and that the masked context representation $\widetilde{\mathbf{U}}_{\mathrm{c}}$ and target representation $\mathbf{U}_{\mathrm{t}}$ are conditionally independent given $\boldsymbol{\xi}$.*

For any measurable predictor $g : \mathbb{R}^q \to \mathbb{R}^q$, define its expected cross-view prediction error as

$$\mathcal{R}(g) = \frac{1}{q}\mathbb{E}\Big[\|g(\mathbf{z}_{\mathrm{c}}) - \mathbf{z}_{\mathrm{t}}\|_2^2\Big]. \tag{54}$$

Among all predictors for which $\mathcal{R}(g) < \infty$, this error is minimized by $g^\star$, where

$$g^\star(\mathbf{z}_{\mathrm{c}}) = \mathbb{E}[\mathbf{z}_{\mathrm{t}} \mid \mathbf{z}_{\mathrm{c}}] = \mathbb{E}[\mathbb{E}[\mathbf{z}_{\mathrm{t}} \mid \boldsymbol{\xi}] \mid \mathbf{z}_{\mathrm{c}}] \quad \textit{almost surely}. \tag{55}$$

Proof. Because $1/q$ is a positive constant, it does not affect the minimizer of $\mathcal{R}(g)$ and is omitted from the following derivation. To characterize this minimizer, we first define the conditional mean of the target embedding given the context embedding as

$$m(\mathbf{z}_\mathrm{c}) = \mathbb{E}[\mathbf{z}_\mathrm{t} \mid \mathbf{z}_\mathrm{c}]. \tag{56}$$

Using this conditional mean, the prediction error of any measurable predictor g can be separated into its deviation from the conditional mean and the remaining target variation:

$$g(\mathbf{z}_\mathrm{c}) - \mathbf{z}_\mathrm{t} = \big(g(\mathbf{z}_\mathrm{c}) - m(\mathbf{z}_\mathrm{c})\big) + \big(m(\mathbf{z}_\mathrm{c}) - \mathbf{z}_\mathrm{t}\big). \tag{57}$$

To determine how these two components contribute to the expected prediction error, we expand the squared norm and take expectations:

$$\begin{aligned}\mathbb{E}\Big[\|g(\mathbf{z}_\mathrm{c}) - \mathbf{z}_\mathrm{t}\|_2^2\Big] = \mathbb{E}\Big[\|g(\mathbf{z}_\mathrm{c}) - m(\mathbf{z}_\mathrm{c})\|_2^2\Big] + \mathbb{E}\Big[\|m(\mathbf{z}_\mathrm{c}) - \mathbf{z}_\mathrm{t}\|_2^2\Big] \\ + 2\mathbb{E}[\langle g(\mathbf{z}_\mathrm{c}) - m(\mathbf{z}_\mathrm{c}), m(\mathbf{z}_\mathrm{c}) - \mathbf{z}_\mathrm{t}\rangle]\,.\end{aligned} \tag{58}$$

The last term is a cross term. To separate the prediction error into the two squared terms above, we now show that this cross term is zero. We condition on $\mathbf{z}_\mathrm{c}$ because, once $\mathbf{z}_\mathrm{c}$ is fixed, $g(\mathbf{z}_\mathrm{c}) - m(\mathbf{z}_\mathrm{c})$ is fixed, while $m(\mathbf{z}_\mathrm{c}) - \mathbf{z}_\mathrm{t}$ contains the remaining variation in the target embedding. The law of total expectation allows us to first average over this remaining variation for each fixed $\mathbf{z}_\mathrm{c}$ and then average over $\mathbf{z}_\mathrm{c}$.

To evaluate this conditional average, we use the definition of $m(\mathbf{z}_\mathrm{c})$ and the linearity of conditional expectation:

$$\begin{aligned}\mathbb{E}[m(\mathbf{z}_\mathrm{c}) - \mathbf{z}_\mathrm{t} \mid \mathbf{z}_\mathrm{c}] &= \mathbb{E}[m(\mathbf{z}_\mathrm{c}) \mid \mathbf{z}_\mathrm{c}] - \mathbb{E}[\mathbf{z}_\mathrm{t} \mid \mathbf{z}_\mathrm{c}] \\ &= m(\mathbf{z}_\mathrm{c}) - \mathbb{E}[\mathbf{z}_\mathrm{t} \mid \mathbf{z}_\mathrm{c}] \\ &= m(\mathbf{z}_\mathrm{c}) - m(\mathbf{z}_\mathrm{c}) \\ &= \mathbf{0}.\end{aligned} \tag{59}$$

Here, $\mathbb{E}[m(\mathbf{z}_\mathrm{c}) \mid \mathbf{z}_\mathrm{c}] = m(\mathbf{z}_\mathrm{c})$ because $m(\mathbf{z}_\mathrm{c})$ is completely determined once $\mathbf{z}_\mathrm{c}$ is given. Equation (59) therefore states that the deviation of $\mathbf{z}_\mathrm{t}$ from its conditional mean has zero average for each fixed context embedding.

Having established this zero conditional mean, we apply the law of total expectation to the cross term:

$$\begin{aligned}&\mathbb{E}[\langle g(\mathbf{z}_\mathrm{c}) - m(\mathbf{z}_\mathrm{c}), m(\mathbf{z}_\mathrm{c}) - \mathbf{z}_\mathrm{t}\rangle] \\ &= \mathbb{E}\Big[\mathbb{E}\Big[\langle g(\mathbf{z}_\mathrm{c}) - m(\mathbf{z}_\mathrm{c}), m(\mathbf{z}_\mathrm{c}) - \mathbf{z}_\mathrm{t}\rangle \;\Big|\; \mathbf{z}_\mathrm{c}\Big]\Big] \\ &= \mathbb{E}[\langle g(\mathbf{z}_\mathrm{c}) - m(\mathbf{z}_\mathrm{c}), \mathbb{E}[m(\mathbf{z}_\mathrm{c}) - \mathbf{z}_\mathrm{t} \mid \mathbf{z}_\mathrm{c}]\rangle] \\ &= 0.\end{aligned} \tag{60}$$

The second equality follows because $g(\mathbf{z}_\mathrm{c}) - m(\mathbf{z}_\mathrm{c})$ is fixed when conditioning on $\mathbf{z}_\mathrm{c}$.

Having shown that the cross term vanishes, we substitute Equation (60) into Equation (58) to obtain

$$\mathbb{E}\Big[\|g(\mathbf{z}_\mathrm{c}) - \mathbf{z}_\mathrm{t}\|_2^2\Big] = \mathbb{E}\Big[\|g(\mathbf{z}_\mathrm{c}) - m(\mathbf{z}_\mathrm{c})\|_2^2\Big] + \mathbb{E}\Big[\|m(\mathbf{z}_\mathrm{c}) - \mathbf{z}_\mathrm{t}\|_2^2\Big]\,. \tag{61}$$

The second term is independent of g, whereas the first term is nonnegative. The expected prediction error is therefore minimized when

$$g^\star(\mathbf{z}_\mathrm{c}) = m(\mathbf{z}_\mathrm{c}) = \mathbb{E}[\mathbf{z}_\mathrm{t} \mid \mathbf{z}_\mathrm{c}] \tag{62}$$

almost surely.

We next express the optimal prediction in terms of the shared scene information $\boldsymbol{\xi}$. By assumption, $\widetilde{\mathbf{U}}_\mathrm{c}$ and $\mathbf{U}_\mathrm{t}$ are conditionally independent given $\boldsymbol{\xi}$. Because the fixed encoders are deterministic functions of these representations, $\mathbf{z}_\mathrm{c}$ and $\mathbf{z}_\mathrm{t}$ are also conditionally independent given $\boldsymbol{\xi}$. This conditional

independence implies that once ξ is given, additionally observing $\mathbf{z}_c$ does not change the conditional mean of $\mathbf{z}_t$:

$$\mathbb{E}[\mathbf{z}_t \mid \xi, \mathbf{z}_c] = \mathbb{E}[\mathbf{z}_t \mid \xi]. \tag{63}$$

To relate this result back to the context-conditioned prediction, we apply the law of iterated expectation and then use Equation (63):

$$\begin{aligned}\mathbb{E}[\mathbf{z}_t \mid \mathbf{z}_c] &= \mathbb{E}[\mathbb{E}[\mathbf{z}_t \mid \xi, \mathbf{z}_c] \mid \mathbf{z}_c] \\ &= \mathbb{E}[\mathbb{E}[\mathbf{z}_t \mid \xi] \mid \mathbf{z}_c]\,. \end{aligned} \tag{64}$$

Together with $g^\star(\mathbf{z}_c) = \mathbb{E}[\mathbf{z}_t \mid \mathbf{z}_c]$, this proves the proposition. □

Explanation. Proposition 1 explains why the target observation can provide supervision even when it contains noise. Under the stated assumptions, the optimal predictor learns to predict the average target representation associated with the scene from the context observation. Random variations specific to individual target observations are averaged out, allowing cross-view prediction to learn stable scene information without requiring a noise-free target.

Covariance-loss control. We next analyze the covariance objective in Equation (17). Recall that $\widetilde{\boldsymbol{R}}$ is the predicted covariance constructed in Equation (15), while $\boldsymbol{R}_t$ is the normalized target covariance defined in Equation (10). Let $\boldsymbol{R}$ denote the trace-normalized ideal covariance of the underlying scene.

Although training aligns $\widetilde{\boldsymbol{R}}$ with $\boldsymbol{R}_t$, the desired reference is $\boldsymbol{R}$. Their difference can be decomposed as

$$\widetilde{\boldsymbol{R}} - \boldsymbol{R} = \left(\widetilde{\boldsymbol{R}} - \boldsymbol{R}_t\right) + (\boldsymbol{R}_t - \boldsymbol{R})\,. \tag{65}$$

Taking the spectral norm and applying the triangle inequality gives

$$\left\|\widetilde{\boldsymbol{R}} - \boldsymbol{R}\right\|_2 \leqslant \underbrace{\left\|\widetilde{\boldsymbol{R}} - \boldsymbol{R}_t\right\|_2}_{\text{model prediction error}} + \underbrace{\|\boldsymbol{R}_t - \boldsymbol{R}\|_2}_{\text{target-view estimation error}}\,. \tag{66}$$

The first term is the model's covariance prediction error, whereas the second is the target-view estimation error.

We now relate the first term to the covariance loss evaluated on the same observation pair. For a complex matrix, its squared Frobenius norm equals the sum of the squared Frobenius norms of its real and imaginary components. Therefore, we have:

$$\begin{aligned}\left\|\widetilde{\boldsymbol{R}} - \boldsymbol{R}_t\right\|_F^2 &= \left\|\mathrm{Re}\left\{\widetilde{\boldsymbol{R}}\right\} - \mathrm{Re}\{\boldsymbol{R}_t\}\right\|_F^2 + \left\|\mathrm{Im}\left\{\widetilde{\boldsymbol{R}}\right\} - \mathrm{Im}\{\boldsymbol{R}_t\}\right\|_F^2 \\ &= 2M^2\mathcal{L}_{\mathrm{cov}}, \end{aligned} \tag{67}$$

where the last equality follows directly from Equation (17). Since the spectral norm of a matrix is bounded by its Frobenius norm, taking the square root yields

$$\left\|\widetilde{\boldsymbol{R}} - \boldsymbol{R}_t\right\|_2 \leqslant \left\|\widetilde{\boldsymbol{R}} - \boldsymbol{R}_t\right\|_F = \sqrt{2}\,M\sqrt{\mathcal{L}_{\mathrm{cov}}}. \tag{68}$$

Explanation. As described in Section 4.1, the target observation uses more snapshots and/or a higher SNR to provide a more reliable covariance estimate. Its remaining discrepancy from the ideal covariance is captured by $\|\boldsymbol{R}_t - \boldsymbol{R}\|_2$. Equations (66) and (68) show that the covariance loss and target-view estimation error jointly bound the covariance recovery error. A small covariance loss ensures that the prediction closely matches the target covariance. If the target covariance is also close to the ideal covariance, the prediction is consequently close to the ideal covariance as well.

We next quantify how this recovery error affects the signal subspace and MUSIC-spectrum denominator.

Proposition 2 (Covariance-to-MUSIC error propagation)**.** *Let $\boldsymbol{R}, \widetilde{\boldsymbol{R}} \in \mathbb{C}^{M\times M}$ be the Hermitian ideal and predicted covariances, respectively. For $1 \leqslant D < M$, let the eigenvalues of $\boldsymbol{R}$ be ordered in descending order and define the signal–noise eigengap as*

$$\delta = \lambda_D(\boldsymbol{R}) - \lambda_{D+1}(\boldsymbol{R}) > 0. \tag{69}$$

Let $\boldsymbol{P}_{\mathrm{s}}$ and $\widetilde{\boldsymbol{P}}_{\mathrm{s}}$ be the orthogonal projectors onto the subspaces spanned by the D dominant eigenvectors of $\boldsymbol{R}$ and $\widetilde{\boldsymbol{R}}$, respectively. Their noise-subspace projectors are $\boldsymbol{P}_{\mathrm{n}} = \boldsymbol{I}_M - \boldsymbol{P}_{\mathrm{s}}$ and $\widetilde{\boldsymbol{P}}_{\mathrm{n}} = \boldsymbol{I}_M - \widetilde{\boldsymbol{P}}_{\mathrm{s}}$. If the covariance recovery error satisfies

$$\left\|\widetilde{\boldsymbol{R}} - \boldsymbol{R}\right\|_2 < \frac{\delta}{2}, \tag{70}$$

then the signal- and noise-subspace errors are bounded by

$$\left\|\widetilde{\boldsymbol{P}}_{\mathrm{s}} - \boldsymbol{P}_{\mathrm{s}}\right\|_2 = \left\|\widetilde{\boldsymbol{P}}_{\mathrm{n}} - \boldsymbol{P}_{\mathrm{n}}\right\|_2 \leqslant \frac{2\left\|\widetilde{\boldsymbol{R}} - \boldsymbol{R}\right\|_2}{\delta}. \tag{71}$$

For any candidate direction ψ, define the reference and predicted MUSIC-spectrum denominators as

$$d(\psi) = \boldsymbol{a}^H(\psi)\boldsymbol{P}_{\mathrm{n}}\boldsymbol{a}(\psi), \qquad \widetilde{d}(\psi) = \boldsymbol{a}^H(\psi)\widetilde{\boldsymbol{P}}_{\mathrm{n}}\boldsymbol{a}(\psi). \tag{72}$$

The corresponding denominator error satisfies

$$\left|\widetilde{d}(\psi) - d(\psi)\right| \leqslant \frac{2\|\boldsymbol{a}(\psi)\|_2^2\|\widetilde{\boldsymbol{R}} - \boldsymbol{R}\|_2}{\delta}. \tag{73}$$

Proof. Assume that Condition (70) holds. Define the covariance recovery error matrix and its spectral norm as

$$\boldsymbol{\Delta} = \widetilde{\boldsymbol{R}} - \boldsymbol{R}, \qquad \varepsilon_{\mathrm{R}} = \|\boldsymbol{\Delta}\|_2. \tag{74}$$

The assumed condition gives $\varepsilon_{\mathrm{R}} < \delta/2$. We use this inequality to establish the subspace and MUSIC-denominator error bounds. For convenience, write $\lambda_i = \lambda_i(\boldsymbol{R})$ and $\widetilde{\lambda}_i = \lambda_i(\widetilde{\boldsymbol{R}})$.

We first determine whether the predicted covariance retains a positive signal–noise eigengap. Weyl's inequality for Hermitian matrices (Bhatia, 1997) bounds the change in each eigenvalue by ε_{R}. In particular,

$$\widetilde{\lambda}_D \geqslant \lambda_D - \varepsilon_{\mathrm{R}}, \qquad \widetilde{\lambda}_{D+1} \leqslant \lambda_{D+1} + \varepsilon_{\mathrm{R}}. \tag{75}$$

Subtracting the second bound from the first gives

$$\widetilde{\lambda}_D - \widetilde{\lambda}_{D+1} \geqslant \lambda_D - \lambda_{D+1} - 2\varepsilon_{\mathrm{R}} = \delta - 2\varepsilon_{\mathrm{R}} > 0, \tag{76}$$

where the final inequality follows from $\varepsilon_{\mathrm{R}} < \delta/2$. Thus, the D largest eigenvalues of $\widetilde{\boldsymbol{R}}$ remain separated from the remaining eigenvalues, so its dominant D-dimensional eigenspace is uniquely defined.

To bound the difference between the reference and predicted signal subspaces, we apply the Davis–Kahan theorem. This requires a separation between the eigenvalues associated with the dominant eigenspace of $\boldsymbol{R}$ and those associated with the complementary eigenspace of $\widetilde{\boldsymbol{R}}$. The smallest such separation is $\lambda_D - \widetilde{\lambda}_{D+1}$. Using the upper bound on $\widetilde{\lambda}_{D+1}$, we obtain

$$\lambda_D - \widetilde{\lambda}_{D+1} \geqslant \lambda_D - \lambda_{D+1} - \varepsilon_{\mathrm{R}} = \delta - \varepsilon_{\mathrm{R}} > 0. \tag{77}$$

Let $\boldsymbol{E}_{\mathrm{s}}$ and $\widetilde{\boldsymbol{E}}_{\mathrm{s}}$ be orthonormal bases for the ranges of $\boldsymbol{P}_{\mathrm{s}}$ and $\widetilde{\boldsymbol{P}}_{\mathrm{s}}$, respectively. Let $\boldsymbol{\Theta}(\widetilde{\boldsymbol{E}}_{\mathrm{s}}, \boldsymbol{E}_{\mathrm{s}})$ denote the diagonal matrix of principal angles between these subspaces, and let $\sin\boldsymbol{\Theta}$ denote the diagonal matrix of their sines. Applying the Davis–Kahan sin-Θ theorem with the separation in Equation (77) gives (Davis & Kahan, 1970)

$$\left\|\sin\boldsymbol{\Theta}(\widetilde{\boldsymbol{E}}_{\mathrm{s}}, \boldsymbol{E}_{\mathrm{s}})\right\|_2 \leqslant \frac{\varepsilon_{\mathrm{R}}}{\delta - \varepsilon_{\mathrm{R}}}. \tag{78}$$

Because $\varepsilon_{\mathrm{R}} < \delta/2$, the denominator satisfies $\delta - \varepsilon_{\mathrm{R}} > \delta/2$. Consequently,

$$\frac{\varepsilon_{\mathrm{R}}}{\delta - \varepsilon_{\mathrm{R}}} \leqslant \frac{2\varepsilon_{\mathrm{R}}}{\delta}. \tag{79}$$

To express this subspace bound using projectors, we use the fact that the operator norm of the difference between two equal-rank orthogonal projectors equals the sine of their largest principal angle (Bhatia, 1997). Therefore,

$$\left\|\widetilde{\boldsymbol{P}}_{\mathrm{s}} - \boldsymbol{P}_{\mathrm{s}}\right\|_2 = \left\|\sin\boldsymbol{\Theta}(\widetilde{\boldsymbol{E}}_{\mathrm{s}}, \boldsymbol{E}_{\mathrm{s}})\right\|_2 \leqslant \frac{2\|\widetilde{\boldsymbol{R}} - \boldsymbol{R}\|_2}{\delta}. \tag{80}$$

The noise projectors are the orthogonal complements of the signal projectors. Substituting their definitions gives

$$\begin{aligned}\widetilde{\boldsymbol{P}}_{\mathrm{n}} - \boldsymbol{P}_{\mathrm{n}} &= \left(\boldsymbol{I}_M - \widetilde{\boldsymbol{P}}_{\mathrm{s}}\right) - \left(\boldsymbol{I}_M - \boldsymbol{P}_{\mathrm{s}}\right) \\ &= -\left(\widetilde{\boldsymbol{P}}_{\mathrm{s}} - \boldsymbol{P}_{\mathrm{s}}\right). \end{aligned} \tag{81}$$

Since a matrix and its negative have the same operator norm, the signal- and noise-projector errors are equal. Combining this equality with the signal-projector bound proves Equation (71).

We next use the noise-projector bound to control the MUSIC-spectrum denominator. For any candidate direction ψ, the definitions of $d(\psi)$ and $\widetilde{d}(\psi)$ give

$$\left|\widetilde{d}(\psi) - d(\psi)\right| = \left|\boldsymbol{a}^H(\psi)\left(\widetilde{\boldsymbol{P}}_{\mathrm{n}} - \boldsymbol{P}_{\mathrm{n}}\right)\boldsymbol{a}(\psi)\right|. \tag{82}$$

The quadratic-form inequality $|\boldsymbol{u}^H\boldsymbol{H}\boldsymbol{u}| \leqslant \|\boldsymbol{u}\|_2^2\|\boldsymbol{H}\|_2$ bounds this difference in terms of the noise-projector error. Applying it and then Equation (71) yields

$$\left|\widetilde{d}(\psi) - d(\psi)\right| \leqslant \|\boldsymbol{a}(\psi)\|_2^2 \left\|\widetilde{\boldsymbol{P}}_{\mathrm{n}} - \boldsymbol{P}_{\mathrm{n}}\right\|_2 \tag{83}$$

$$\leqslant \frac{2\|\boldsymbol{a}(\psi)\|_2^2\|\widetilde{\boldsymbol{R}} - \boldsymbol{R}\|_2}{\delta}. \tag{84}$$

This proves Equation (73) and completes the proof. □

Substituting the covariance-loss bound in Equation (68) into the error decomposition in Equation (66) gives

$$\left\|\widetilde{\boldsymbol{R}} - \boldsymbol{R}\right\|_2 \leqslant \sqrt{2}\,M\sqrt{\mathcal{L}_{\mathrm{cov}}} + \|\boldsymbol{R}_{\mathrm{t}} - \boldsymbol{R}\|_2. \tag{85}$$

The first term on the right-hand side is controlled by the covariance loss, while the second measures the target-view estimation error. Proposition 2 requires the total covariance recovery error on the left-hand side to be smaller than $\delta/2$. Therefore, the proposition applies whenever the upper bound in Equation (85) is smaller than $\delta/2$, leading to the following corollary.

Corollary 1 (Loss-to-MUSIC error control)**.** *Retain the notation of Proposition 2, and assume that $\boldsymbol{R}$ and $\widetilde{\boldsymbol{R}}$ are Hermitian and that the signal–noise eigengap of $\boldsymbol{R}$ satisfies $\delta > 0$. If the loss-based covariance-error bound satisfies*

$$\sqrt{2}\,M\sqrt{\mathcal{L}_{\mathrm{cov}}} + \|\boldsymbol{R}_{\mathrm{t}} - \boldsymbol{R}\|_2 < \frac{\delta}{2}, \tag{86}$$

then the signal- and noise-subspace errors satisfy

$$\begin{aligned}\left\|\widetilde{\boldsymbol{P}}_{\mathrm{s}} - \boldsymbol{P}_{\mathrm{s}}\right\|_2 &= \left\|\widetilde{\boldsymbol{P}}_{\mathrm{n}} - \boldsymbol{P}_{\mathrm{n}}\right\|_2 \\ &\leqslant \frac{2}{\delta}\left(\sqrt{2}\,M\sqrt{\mathcal{L}_{\mathrm{cov}}} + \|\boldsymbol{R}_{\mathrm{t}} - \boldsymbol{R}\|_2\right). \end{aligned} \tag{87}$$

Furthermore, for any candidate direction ψ, the corresponding MUSIC-spectrum-denominator error satisfies

$$\left|\widetilde{d}(\psi) - d(\psi)\right| \leqslant \frac{2\|\boldsymbol{a}(\psi)\|_2^2}{\delta}\left(\sqrt{2}\,M\sqrt{\mathcal{L}_{\mathrm{cov}}} + \|\boldsymbol{R}_{\mathrm{t}} - \boldsymbol{R}\|_2\right). \tag{88}$$

Proof. We first bound the covariance recovery error $\|\widetilde{\boldsymbol{R}} - \boldsymbol{R}\|_2$ appearing in Proposition 2. The error decomposition in Equation (66) gives

$$\left\|\widetilde{\boldsymbol{R}} - \boldsymbol{R}\right\|_2 \leqslant \left\|\widetilde{\boldsymbol{R}} - \boldsymbol{R}_{\mathrm{t}}\right\|_2 + \|\boldsymbol{R}_{\mathrm{t}} - \boldsymbol{R}\|_2. \tag{89}$$

The first term is the model prediction error relative to the target covariance. According to Equation (68), this term is bounded by the covariance loss:

$$\left\|\widetilde{\boldsymbol{R}} - \boldsymbol{R}_{\mathrm{t}}\right\|_2 \leqslant \sqrt{2}\,M\sqrt{\mathcal{L}_{\mathrm{cov}}}. \tag{90}$$

Substituting this bound into the covariance-error decomposition yields

$$\left\|\widetilde{\boldsymbol{R}}-\boldsymbol{R}\right\|_2 \leqslant \sqrt{2}\,M\sqrt{\mathcal{L}_{\mathrm{cov}}}+\left\|\boldsymbol{R}_{\mathrm{t}}-\boldsymbol{R}\right\|_2, \tag{91}$$

which is Equation (85).

Condition (86) states that the right-hand side of this upper bound is smaller than $\delta/2$. It therefore follows that

$$\left\|\widetilde{\boldsymbol{R}}-\boldsymbol{R}\right\|_2 < \frac{\delta}{2}. \tag{92}$$

Thus, the covariance recovery error satisfies Condition (70), so Proposition 2 applies.

We first apply the projector-error result in Equation (71), which gives

$$\left\|\widetilde{\boldsymbol{P}}_{\mathrm{s}}-\boldsymbol{P}_{\mathrm{s}}\right\|_2 = \left\|\widetilde{\boldsymbol{P}}_{\mathrm{n}}-\boldsymbol{P}_{\mathrm{n}}\right\|_2 \leqslant \frac{2\left\|\widetilde{\boldsymbol{R}}-\boldsymbol{R}\right\|_2}{\delta}. \tag{93}$$

Replacing the covariance recovery error by its upper bound in Equation (85) then yields

$$\begin{aligned}\left\|\widetilde{\boldsymbol{P}}_{\mathrm{s}}-\boldsymbol{P}_{\mathrm{s}}\right\|_2 &= \left\|\widetilde{\boldsymbol{P}}_{\mathrm{n}}-\boldsymbol{P}_{\mathrm{n}}\right\|_2 \\ &\leqslant \frac{2}{\delta}\left(\sqrt{2}\,M\sqrt{\mathcal{L}_{\mathrm{cov}}}+\left\|\boldsymbol{R}_{\mathrm{t}}-\boldsymbol{R}\right\|_2\right),\end{aligned} \tag{94}$$

which proves Equation (87).

We next apply the MUSIC-denominator bound in Equation (73). For any candidate direction ψ, Proposition 2 gives

$$\left|\widetilde{d}(\psi)-d(\psi)\right| \leqslant \frac{2\|\boldsymbol{a}(\psi)\|_2^2\|\widetilde{\boldsymbol{R}}-\boldsymbol{R}\|_2}{\delta}. \tag{95}$$

Substituting Equation (85) into this inequality gives

$$\left|\widetilde{d}(\psi)-d(\psi)\right| \leqslant \frac{2\|\boldsymbol{a}(\psi)\|_2^2}{\delta}\left(\sqrt{2}\,M\sqrt{\mathcal{L}_{\mathrm{cov}}}+\left\|\boldsymbol{R}_{\mathrm{t}}-\boldsymbol{R}\right\|_2\right), \tag{96}$$

which proves Equation (88) and completes the proof. □

Explanation. Corollary 1 connects covariance learning to the quantities used by MUSIC. Under the stated condition, a smaller covariance loss and a smaller target-view estimation error tighten the bounds on the signal- and noise-subspace errors and the MUSIC-denominator error. For the same covariance recovery error, a larger signal–noise eigengap gives tighter bounds.

Projector-loss control. Proposition 2 shows that the signal-projector error also controls the noise-projector error and the MUSIC-spectrum denominator. While Corollary 1 bounds this error through covariance learning, the projector objective in Equation (18) provides a direct bound. Recall that $\boldsymbol{P}_{\mathrm{s,t}}$ denotes the target signal-subspace projector. For each observation pair, the loss definition gives

$$\left\|\widetilde{\boldsymbol{P}}_{\mathrm{s}}-\boldsymbol{P}_{\mathrm{s,t}}\right\|_2 \leqslant \left\|\widetilde{\boldsymbol{P}}_{\mathrm{s}}-\boldsymbol{P}_{\mathrm{s,t}}\right\|_{\mathrm{F}} = M\sqrt{\mathcal{L}_{\mathrm{proj}}}. \tag{97}$$

Following the same prediction-versus-target error decomposition used in Equation (66), we obtain

$$\left\|\widetilde{\boldsymbol{P}}_{\mathrm{s}}-\boldsymbol{P}_{\mathrm{s}}\right\|_2 \leqslant M\sqrt{\mathcal{L}_{\mathrm{proj}}}+\left\|\boldsymbol{P}_{\mathrm{s,t}}-\boldsymbol{P}_{\mathrm{s}}\right\|_2. \tag{98}$$

Substituting this bound into the projector-to-denominator inequality established in the proof of Proposition 2 yields

$$\left|\widetilde{d}(\psi)-d(\psi)\right| \leqslant \|\boldsymbol{a}(\psi)\|_2^2\left(M\sqrt{\mathcal{L}_{\mathrm{proj}}}+\left\|\boldsymbol{P}_{\mathrm{s,t}}-\boldsymbol{P}_{\mathrm{s}}\right\|_2\right). \tag{99}$$

Explanation. The covariance and projector objectives control the same MUSIC-relevant subspace error through different quantities. Covariance learning bounds it through the covariance recovery error and the reference eigengap. Projector learning directly bounds it through alignment with the target subspace. Both bounds retain a target-estimation error term, making the quality of the target observation relevant to both objectives.

Table 3: Shared synthetic data and evaluation settings. Training and validation settings apply to learned methods.

Parameter	Setting
Array	Half-wavelength-spaced ULA, $M = 8$
Source count	$D = 3$, known
Source symbols	Unit-average-power 16-QAM
Source coherence	$\rho = 0$ or $\rho = 1$
Context snapshots (all splits)	$N_{\mathrm{c}} = 32$
Training SNR	Uniform over $[-5, 5]$ dB
Validation / main-test SNR	0 dB
Training / validation / test samples	4,000 / 500 / 500 per dataset

D Appendix 4: More Experimental Details

This appendix supplements the experimental setup with data-generation, implementation, training, and evaluation details.

D.1 Data Generation and Paired Observations

Synthetic observations follow the signal model in Section 3.1. Table 3 summarizes the default settings shared across methods.

Source signals and noise. Source signals have equal average power and use unit-average-power 16-ary quadrature amplitude modulation (16-QAM) symbols (Hanzo et al., 2004). Noncoherent sources use independently sampled symbol sequences. Fully coherent sources share a common sequence multiplied by source-specific unit-magnitude complex coefficients. The phases of these coefficients are sampled uniformly from $[0, 2\pi)$ and remain fixed within each sample. Further explanation of source coherence is provided in Appendix A.3. Circularly symmetric complex Gaussian noise is sampled independently across antennas and snapshots.

Context–target pairs. For JEPA-MUSIC, each context observation is paired with a target containing $N_{\mathrm{t}} = 512$ snapshots at 20 dB. The two observations share source directions, powers, and coherence configuration, including the source-specific phase coefficients for coherent sources. Symbol sequences and receiver noise are sampled independently between the two observations. The target observation provides a higher-quality observation for self-supervised learning and is not required at inference.

Data splits. Training, validation, and test samples are generated independently for noncoherent and fully coherent sources, with fixed splits shared across methods. All learned methods use the same training and validation samples, but receive different supervision. Supervised baselines use ground-truth DoAs, whereas JEPA-MUSIC uses paired observations.

D.2 Model Implementations and Training

Covariance preprocessing. Raw MUSIC uses the sample covariance of the context observation. FBSS-MUSIC applies forward–backward spatial smoothing (FBSS) (Pillai & Kwon, 1989) with subarray size 6, yielding three overlapping subarrays for the eight-antenna ULA. For JEPA-MUSIC, noncoherent inputs use the full-array covariance, whereas coherent inputs use the same FBSS preprocessing as FBSS-MUSIC. This comparison therefore isolates the gain from learned covariance recovery beyond spatial smoothing. The preprocessing is applied to both context and target observations, yielding 8×8 and 6×6 covariances for the noncoherent and coherent settings, respectively. Each covariance is trace-normalized to its matrix dimension and represented by separate real and imaginary channels.

JEPA-MUSIC architecture. The context encoder processes the flattened two-channel covariance using two fully connected layers with output dimensions 192 and 96. Each layer is followed by GELU,

Table 4: Optimization settings for learned methods.

Method	Optimizer	Initial learning rate	Batch size
JEPA-MUSIC	AdamW	10^{-3}	64
DeepCNN	Adam	10^{-3}	64
DA-MUSIC	Adam	10^{-3}	16
SubspaceNet	Adam	10^{-5}	256

with layer normalization between the first GELU and the second linear layer. The predictor contains two 96-dimensional linear layers, separated by GELU and layer normalization. The covariance head uses a 192-dimensional hidden layer with GELU, followed by a linear layer producing the real and imaginary components of the complex factor in Equation (14). The structured construction uses diagonal loading $\epsilon = 10^{-3}$. The target encoder has the same architecture as the context encoder and is initialized with identical parameters. We use a masking probability of 0.1 and EMA momentum $\mu = 0.99$, following the procedures in Appendix B.2. The loss weights in Equation (19) are $\lambda_{\mathrm{z}} = 0.1$, $\lambda_{\mathrm{R}} = 10$, and $\lambda_{\mathrm{P}} = 20$.

Supervised learning baselines. We adopt the DeepCNN baseline (Papageorgiou et al., 2021) following the implementation of Merkofer et al. (Merkofer et al., 2024). It uses the real, imaginary, and phase components of the empirical covariance to predict source occupancy on a 360-point angular grid. DA-MUSIC (Merkofer et al., 2024) uses its known-source-count architecture, which processes raw snapshots through a recurrent network and learns covariance-like features and spectrum-to-angle recovery. SubspaceNet (Shmuel et al., 2025) uses eight autocorrelation lags to learn a surrogate covariance through a differentiable ESPRIT estimator. All three baselines are implemented in PyTorch and trained on the shared data splits using their respective supervised objectives.

Training. All learned methods are trained from random initialization for 500 epochs. Specifically, we use method-specific optimization settings, as summarized in Table 4, while keeping the data splits and training epoch fixed across methods. JEPA-MUSIC uses weight decay 10^{-4} and gradient-norm clipping at 1. SubspaceNet uses weight decay 10^{-9} and reduces its learning rate by a factor of 0.2 every 80 epochs (Shmuel et al., 2025). No ground-truth DoAs are used to train or select JEPA-MUSIC.

D.3 Evaluation Protocols

Evaluation metrics. For each test sample, predicted directions are matched to the ground-truth directions using Hungarian assignment with squared angular error as the matching cost. Let $e_{i,d}$ denote the resulting signed error, in degrees, for source d in test sample i. For K test samples with D sources, the angular RMSE is

$$\mathrm{RMSE} = \sqrt{\frac{1}{KD}\sum_{i=1}^{K}\sum_{d=1}^{D} e_{i,d}^2}. \tag{100}$$

Med. and P90 are the median and 90th percentile of the KD absolute errors, respectively. A test sample is considered a failure if any matched source has an absolute error exceeding 5°. Accordingly,

$$\mathrm{Fail.} = \frac{100}{K}\sum_{i=1}^{K} \mathbf{1}\left\{ \max_{1\leqslant d\leqslant D} |e_{i,d}| > 5^\circ \right\}, \tag{101}$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function. Metrics are computed separately for each training seed and then averaged over the five seeds, rather than pooling predictions across runs.

SNR and snapshot sweeps. We evaluate SNRs of $\{-10, -5, 0, 5, 10, 15, 20\}$ dB with 32 context snapshots, and snapshot counts of $\{8, 16, 32, 64, 128, 256\}$ at 0 dB. The sweeps are conducted separately for noncoherent and fully coherent sources using models trained under the corresponding source condition. Each condition contains 500 test samples.

Signal-subspace recovery. We compare the signal subspaces obtained from the observed and predicted covariances against that of the higher-quality target covariance. For the fully coherent setting, context and target observations undergo the same FBSS preprocessing, so all compared covariances are 6×6 matrices. The observed covariance is therefore the FBSS covariance before learned recovery, rather than the raw full-array covariance.

For each test sample, let $\boldsymbol{E}_{\mathrm{s}}$ contain an orthonormal basis for the D dominant eigenvectors of the covariance being evaluated, and let $\boldsymbol{E}_{\mathrm{s,t}}$ denote the corresponding target-view basis. The principal angles between their subspaces are

$$\alpha_d = \arccos\big(\sigma_d(\boldsymbol{E}_{\mathrm{s}}^H \boldsymbol{E}_{\mathrm{s,t}})\big), \qquad d = 1, \ldots, D, \tag{102}$$

where $\sigma_d(\cdot)$ denotes a singular value. Singular values are clipped to $[0, 1]$ for numerical stability. We average the D principal angles within each sample, convert them to degrees, and then average across test samples and training seeds. The target-view subspace serves as a reference, not an exact ground-truth subspace.

Training-sample efficiency. We train learned methods using 400, 1,000, 2,000, and 4,000 training samples, corresponding to 10%, 25%, 50%, and 100% of the full training set. The 500 validation and 500 test samples remain unchanged at every training-set size. Each model is trained from random initialization using the same five seeds, 500 training epochs, method-specific optimization settings described in Appendix D.2. Raw MUSIC and FBSS-MUSIC require no training, so their results remain unchanged across training-set sizes.

E Appendix 5: More Experimental Results

This appendix provides additional experiments and analyses complementing the main results.

E.1 Ablation Studies

We ablate the predictor, EMA update, and structured covariance construction of JEPA-MUSIC under noncoherent and fully coherent sources at SNR $= 0$ dB with $N_{\mathrm{c}} = 32$. All variants follow the training protocol in Appendix D.2. Removing the predictor replaces it with an identity mapping. Removing EMA directly copies the context encoder parameters to the target encoder after each update. The Hermitian-only head replaces the structured covariance construction in Equation (15) with $\widetilde{\boldsymbol{R}} = (\boldsymbol{B} + \boldsymbol{B}^H)/2$, where $\boldsymbol{B}$ is the complex matrix produced by the covariance head. This output is Hermitian but is not guaranteed to be positive semidefinite, and neither diagonal loading nor trace normalization is applied.

Table 5: Component ablations at SNR $= 0$ dB with $N_{\mathrm{c}} = 32$. Angular errors are in degrees and failure rates are in percent. RMSE is reported as mean $\pm$ standard deviation over five seeds. Best results within each source setting are bold.

Variant	RMSE↓	Med.↓	P90↓	Fail.↓
Noncoherent ($\rho = 0$)				
JEPA-MUSIC	$\mathbf{1.012} \pm 0.241$	0.477	1.249	**0.36**
w/o EMA	1.052 ± 0.394	0.476	1.278	0.40
w/o predictor	1.205 ± 0.417	**0.454**	**1.223**	0.40
Hermitian-only head	1.241 ± 0.371	0.473	1.264	0.40
Fully coherent ($\rho = 1$)				
JEPA-MUSIC	$\mathbf{2.915} \pm 0.694$	**0.781**	2.533	6.80
w/o EMA	3.093 ± 0.252	0.802	**2.500**	**6.76**
w/o predictor	3.041 ± 0.338	0.792	2.518	6.88
Hermitian-only head	3.808 ± 0.321	0.856	2.886	8.72

Table 5 shows that the full JEPA-MUSIC achieves the lowest mean RMSE among these component variants in both source settings. Removing the predictor or EMA increases mean RMSE, although some median, P90, or failure-rate metrics improve slightly. Replacing the structured covariance construction with a Hermitian-only head increases RMSE from 1.012° to 1.241° for noncoherent sources and from 2.915° to 3.808° for fully coherent sources, with higher P90 and failure rates in

both settings. These results support the component design of JEPA-MUSIC while highlighting trade-offs across error metrics.

E.2 Generalization Across Source-Signal Distributions

We evaluate JEPA-MUSIC, trained on 16-QAM signals, across different source-signal distributions, including QPSK, 64-QAM, and circularly symmetric complex Gaussian signals, without fine-tuning. Figure 6 illustrates these unit-average-power source distributions before receiver noise is added. QPSK has constant magnitude, QAM uses multiple amplitude levels, and complex Gaussian signals have continuously distributed values. These differences affect covariance estimates obtained from finite snapshots, even when the underlying source directions remain unchanged.

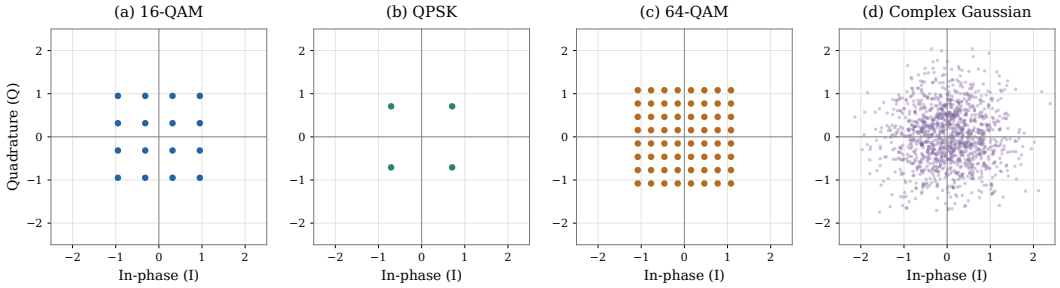

Figure 6: Unit-average-power source distributions: (a) 16-QAM, (b) QPSK, (c) 64-QAM, and (d) samples from a complex Gaussian distribution. Models are trained with 16-QAM and tested on all four distributions.

Table 6: Generalization across source-signal distributions: JEPA-MUSIC trained on 16-QAM is evaluated on 16-QAM, QPSK, 64-QAM, and complex Gaussian signals without retraining or fine-tuning. Evaluation uses 1,000 new test samples per distribution at SNR $= 0$ dB with $N_c = 32$, sharing source directions and receiver-noise realizations across distributions. Angular errors are in degrees and failure rates are in percent. RMSE is reported as mean $\pm$ standard deviation over five training seeds.

Test distribution	RMSE↓	Med.↓	P90↓	Fail.↓
Noncoherent sources ($\rho = 0$)				
16-QAM (reference)	1.168 ± 0.265	0.442	1.194	0.36
QPSK	1.153 ± 0.235	0.439	1.131	0.36
64-QAM	1.086 ± 0.238	0.450	1.203	0.36
Complex Gaussian	1.299 ± 0.292	0.468	1.242	0.54
Fully coherent sources ($\rho = 1$)				
16-QAM (reference)	3.823 ± 0.309	0.776	2.532	8.48
QPSK	3.747 ± 0.376	0.775	2.479	8.14
64-QAM	3.854 ± 0.372	0.778	2.541	8.54
Complex Gaussian	3.881 ± 0.510	0.772	2.540	7.68

Table 6 shows that JEPA-MUSIC maintains comparable performance across the evaluated distributions. For noncoherent sources, QPSK and 64-QAM yield RMSEs of $1.153°$ and $1.086°$, respectively, compared with $1.168°$ for the 16-QAM reference. Complex Gaussian signals increase RMSE to $1.299°$, while the failure rate remains below 0.6%. For fully coherent sources, RMSE ranges from $3.747°$ to $3.881°$, with similarly small changes in median and P90 errors.

These results highlight the transferability of the covariance representation learned by JEPA-MUSIC: training with 16-QAM alone yields reliable DoA estimation across distinct source-signal distributions, without distribution-specific training or model adaptation.

E.3 Transfer and Learning on DeepMIMO

We evaluate JEPA-MUSIC on the DeepMIMO dataset (Alkhateeb, 2019) using the `asu_campus_3p5` scenario under two protocols: direct transfer of synthetic-trained models without

fine-tuning, and training from scratch on DeepMIMO. Observations are constructed from dominant-path parameters for an eight-antenna ULA with three noncoherent, equal-power sources at SNR $= 0$ dB and $N_c = 32$ snapshots. Both protocols use the same 500 test samples, with learned-method results averaged over five training seeds. MUSIC-based spectral searches use 1,201 grid points over $[-90°, 90°]$, while learned internal decoding structures are retained. Together, these evaluations assess transferability without adaptation and learning effectiveness on DeepMIMO.

Table 7: Evaluation on DeepMIMO using the same 500 test samples under both protocols. Angular errors are in degrees and failure rates are in percent. RMSE is reported as mean $\pm$ standard deviation over five training seeds for learned methods. Best results within each block are bold.

Method	RMSE↓	Med.↓	P90↓	Fail.↓
Direct transfer without fine-tuning				
Raw MUSIC	18.111	1.270	29.962	54.60
FBSS-MUSIC	19.227	1.518	40.785	50.80
DeepCNN	8.184 ± 0.603	1.550	17.408	40.68
DA-MUSIC	6.448 ± 0.643	3.239	9.218	60.16
SubspaceNet	11.313 ± 0.660	6.511	19.007	92.64
JEPA-MUSIC	$\mathbf{4.569} \pm 0.239$	**0.581**	**1.759**	**7.08**
Training on DeepMIMO				
Raw MUSIC	18.111	1.270	29.962	54.60
FBSS-MUSIC	19.227	1.518	40.785	50.80
DeepCNN	2.230 ± 0.507	0.569	1.936	2.40
DA-MUSIC	2.594 ± 1.003	1.071	3.693	13.52
SubspaceNet	5.013 ± 1.476	2.663	7.658	48.44
JEPA-MUSIC	$\mathbf{1.079} \pm 0.419$	**0.498**	**1.433**	**0.16**

Table 7 shows that JEPA-MUSIC leads all four metrics under both protocols. Notably, without any retraining or fine-tuning on DeepMIMO, it achieves an RMSE of $4.569°$, reducing RMSE by 74.8% and 76.2% relative to Raw MUSIC and FBSS-MUSIC, respectively, and by 29.1% relative to DA-MUSIC, the strongest synthetic-trained learning-based baseline. Its P90 error of $1.759°$ and failure rate of 7.08% are also the lowest, demonstrating that its transfer advantage extends beyond average accuracy to estimation reliability.

When trained on DeepMIMO, JEPA-MUSIC achieves an RMSE of $1.079°$, a 51.6% reduction relative to DeepCNN, the strongest supervised baseline. It also reduces the failure rate from 2.40% to 0.16%, while attaining the lowest median and P90 errors. Together, these results support the design of JEPA-MUSIC: learning a MUSIC-compatible covariance representation through self-supervision, while retaining model-based direction estimation, enables accurate and reliable DoA estimation both with and without target-dataset training.

E.4 SubspaceNet with Increased Training Data

Table 8: Training-sample comparison between SubspaceNet and JEPA-MUSIC. All metrics are averaged over five training seeds. Angular errors are in degrees and failure rates are in percent. Best results within each source setting are bold.

Method	Samples	RMSE↓	Med.↓	P90↓	Fail.↓
Noncoherent sources ($\rho = 0$)					
SubspaceNet	4,000	10.541	7.704	17.295	95.12
SubspaceNet	45,000	2.137	1.172	3.241	8.64
JEPA-MUSIC	400	2.088	0.609	1.694	1.52
JEPA-MUSIC	1,000	1.324	0.540	1.405	0.64
JEPA-MUSIC	2,000	1.299	0.506	1.325	0.60
JEPA-MUSIC	4,000	**1.012**	**0.477**	**1.249**	**0.36**
Fully coherent sources ($\rho = 1$)					
SubspaceNet	4,000	11.198	8.155	17.931	95.08
SubspaceNet	45,000	5.269	2.394	8.156	50.04
JEPA-MUSIC	400	4.701	1.096	4.070	15.88
JEPA-MUSIC	1,000	4.110	0.895	3.191	11.16
JEPA-MUSIC	2,000	3.554	0.814	2.655	7.76
JEPA-MUSIC	4,000	**2.915**	**0.781**	**2.533**	**6.80**

To investigate SubspaceNet's relatively high errors in the main experiments, we increase its training set from 4,000 to 45,000 samples while keeping the validation and test sets, model architecture, and optimizer settings unchanged. All runs use the same five training seeds and 500 training epochs, with the same evaluation metrics as the main experiments.

Table 8 shows that expanding SubspaceNet's training set from 4,000 to 45,000 samples reduces RMSE from 10.541° to 2.137° for noncoherent sources and from 11.198° to 5.269° for fully coherent sources. This substantial improvement with more training data and optimization steps suggests that SubspaceNet's performance in the main experiments is constrained by the limited training setting.

Nevertheless, JEPA-MUSIC achieves better mean results across all four metrics with only 400 training samples. Although its noncoherent RMSE of 2.088° is only slightly lower than SubspaceNet's 2.137°, its failure rate is substantially lower at 1.52% versus 8.64%. For fully coherent sources, it achieves an RMSE of 4.701° and reduces the failure rate from 50.04% to 15.88%. As the number of training pairs increases from 400 to 4,000, all four metrics improve consistently in both settings. These results demonstrate that JEPA-MUSIC achieves accurate and reliable estimation with limited training data and benefits further from additional samples.