# SUPPORT SAMPLING FOR MONTAGE-AGNOSTIC EEG SOURCE IMAGING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

EEG source imaging (ESI) aims to reconstruct cortical activity from scalp measurements that mix signals generated by multiple sources. This represent a notable ill-posed inverse problem since distinct source configurations can produce similar EEG. Classical methods use a supplied leadfield and explicit assumptions to select a reconstruction, Bayesian methods use a leadfield to infer source uncertainty for each recording, and learned imagers are fast but often return a single estimate for a fixed electrode layout. We introduce a geometry-conditioned imager that estimates parcel waveforms and autoregressively samples sets of active parcels of variable size. By encoding EEG measurements and electrode coordinates in an internal representation, the same model processes different montages without a supplied leadfield, which it predicts directly from geometry. On held-out 256-channel synthetic data, our deterministic estimate improves on the previous state of the art, reducing median localization error by 13.4% to 9.7 mm and increasing AUPRC from 0.749 to 0.787. On real EEG with known intracranial stimulation sites, the full autoregressive model reduces median peak-to-site distance by 15.5% to 22.3 mm relative to the best-performing baseline and achieves a stimulated-parcel AUPRC of 0.090. Trained jointly on four electrode layouts, the imager also generalizes to unseen 21-, 76-, and 256-channel montages, achieving median localization errors of 10.4–16.5 mm compared with 10.6 mm across seen layouts.

## 1 INTRODUCTION

Electroencephalography (EEG) measures brain activity with millisecond resolution; however, each scalp electrode captures a composite of signals originating from multiple cortical sources. EEG source imaging (ESI) aims to reverse this mixing and recover the underlying neural activity, with applications spanning neuroscience research and the localization of epileptogenic regions (He et al., 2018; Michel & Brunet, 2019). The relationship between brain sources and scalp measurements is characterized by the leadfield (LF) matrix, which specifies how activity at each brain location manifests at each electrode. The main challenge is that the inverse problem is fundamentally ill-posed, as distinct source configurations can yield nearly identical scalp measurements.

Classical ESI methods use a leadfield to relate candidate sources to the measurements for a given electrode layout and head model. They select a solution by imposing explicit assumptions, for instance, minimum-norm methods penalize source amplitude (H"am"al"ainen & Ilmoniemi, 1994), sparse methods favor a restricted set of active sources (Gramfort et al., 2012), and beamformers construct spatial filters subject to constraints derived from the leadfield and sensor covariance (Van Veen et al., 1997). In all cases, it is the researcher who introduces the assumptions, based on neurophysiological intuition, that selects one solution among the many compatible with the EEG.

Bayesian methods also use an explicit forward model, but express source assumptions as priors and infer a distribution over possible solutions (Costa et al., 2017; Luria et al., 2024). Multi-dipole methods can quantify uncertainty in both source location and number, although they typically require a leadfield and perform inference for each recording (Sorrentino et al., 2014; Luria et al., 2024). Learned imagers instead, follow a different path: they fit an inverse mapping from simulated EEG–source pairs, allowing source patterns in the training data to shape the model's predictions (Hecker et al., 2021; Sun et al., 2022). They still inherit assumptions from the simulator, architecture, and



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.

Table 1: Properties of representative ESI approaches. 'Cross-montage' denotes applying the same method or weights to another electrode layout, without implying equal accuracy, 'support-set distribution' denotes a distribution over complete sets of active locations, including their number. It describes the model's output, not the calibration of its probabilities. △ indicates method-dependent behavior.

| Approach | Cross-montage | No supplied LF | Support-set distribution |
|---|---|---|---|
| Classical ESI | ✓ | × | × |
| Physics-based Bayesian | △ | × | △ |
| Learned ESI | △ | △ | △ |
| **Ours** | ✓ | ✓ | ✓ |

training objective: learning the mapping changes how the inverse problem is constrained, but does not remove the forward model from the overall pipeline.

Furthermore, there is an additional difficulty with learning and Bayesian ESI solvers, namely their inability to apply the same inverse procedure to different assemblies. Models such as ConvDip Hecker et al. (2021) use a fixed sensor representation, while evaluations across electrode configurations may train a separate model for each layout (Rong et al., 2025). Table 1 summarizes all of these aspects in comparison with our method.

To address these limitations, we develop a geometry-conditioned imager that uses EEG measurements and electrode coordinates without requiring a supplied leadfield at inference. Therefore, our research questions are: (i) can a geometry-conditioned imager localize cortical sources across electrode layouts without a supplied leadfield? (ii) How does autoregressive support sampling accommodate uncertainty in parcel selection?

We train on synthetic EEG–source pairs from different backend simulators, then evaluate montage transfer and support sampling separately on held-out scenarios. To probe transfer beyond simulation, we also test localization and uncertainty on real scalp EEG recorded during intracranial stimulation, using the known stimulation site as a spatial reference. In summary, our main contributions are:

- **Geometry-conditioned imaging:** We introduce an imager that reconstructs cortical activity from EEG and predicts the leadfield from electrode geometry, without a supplied forward model. Trained on synthetic data, our full autoregressive model generalizes to real EEG on Localize-MI Mikulan et al. (2020), achieving 22.3 mm median peak-to-site distance and 0.090 stimulated-parcel AUPRC. Its localization error is 15.5% lower than that of the strongest baseline, ConvDip Hecker et al. (2021).
- **Cross-montage transfer:** A single model trained across four layouts transfers without retraining to unseen 21-, 76-, and 256-channel montages, covering both new combinations of familiar electrodes and previously unseen positions.
- **Autoregressive support sampling:** We introduce a decoder that samples variable-size sets of active parcels by modeling dependencies between activations. On held-out synthetic data, it achieves lower binary-support energy scores and excess AURC than the Bayesian baseline SESAME and the evaluated learned alternatives.

## 2 PRELIMINARIES

**EEG forward and inverse problems.** Neuronal populations and the cortical surface are represented by a dense mesh of vertices $V$, which are typically grouped using a parcel scheme into $P$ anatomical parcels. At this resolution, let $\boldsymbol{S} \in \mathbb{R}^{P\times T}$ contain the parcel waveforms. We restrict all sensor-indexed quantities to electrodes marked valid by a binary mask $\boldsymbol{m}$ ($m_c = 1$), reindex the retained electrodes from 1 to $C$, and omit the mask from subsequent notation. Thus $\boldsymbol{Y} \in \mathbb{R}^{C\times T}$ contains their EEG measurements over $T$ time samples and $\boldsymbol{X} \in \mathbb{R}^{C\times 3}$ their positions in the common anatomical coordinate system. For fixed anatomy, source orientations, and an EEG reference convention, the forward model is

$$\boldsymbol{Y} = \boldsymbol{G}\boldsymbol{S} + \boldsymbol{E}, \qquad \boldsymbol{G} \in \mathbb{R}^{C\times P}, \tag{1}$$

where $\boldsymbol{G}$ is the parcel-level leadfield and $\boldsymbol{E}$ collects background activity, measurement noise, and the discrepancy introduced by parcellation. Since $P > C$, the leadfield has a nontrivial null space: distinct source fields can produce identical noiseless measurements, so reconstruction requires assumptions that distinguish among these compatible sources.

**A fixed temporal coordinate system.** Let $\boldsymbol{\Phi} \in \mathbb{R}^{D\times T}$ contain the first $D \leq T$ rows of the orthonormal type-II discrete cosine transform (DCT), so that $\boldsymbol{\Phi}\boldsymbol{\Phi}^\top = \boldsymbol{I}_D$. For a field $\boldsymbol{A} \in \mathbb{R}^{n\times T}$, its temporal coefficients and retained-band reconstruction are

$$\boldsymbol{A}_{\mathrm{D}} = \boldsymbol{A}\boldsymbol{\Phi}^\top, \qquad \widetilde{\boldsymbol{A}} = \boldsymbol{A}_{\mathrm{D}}\boldsymbol{\Phi}. \tag{2}$$

The transform acts independently on each spatial row, reducing the temporal dimension while preserving the spatial coordinates. Applying it to equation 1 gives

$$\boldsymbol{Y}_{\mathrm{D}} = \boldsymbol{G}\boldsymbol{S}_{\mathrm{D}} + \boldsymbol{E}_{\mathrm{D}}.$$

Thus, the forward relation remains exact in coefficient space, while reconstruction in time retains only the first $D$ DCT modes. Throughout, matrices and vectors are written in bold uppercase and lowercase, respectively, with hats denoting estimates and tildes denoting deterministic approximations.

**Distributions over finite supports.** For a finite catalog $[P] = \{1, \ldots, P\}$ of candidate locations, an active support is an unordered subset $Z \subseteq [P]$ with cardinality $K = |Z|$. We describe a support by a sequence of distinct indices followed by a STOP token, with autoregressive law

$$q(\boldsymbol{z}) = \prod_{\ell=1}^{K+1} q(z_\ell \mid z_{<\ell}), \qquad z_{K+1} = \textsc{stop}$$

The probability of a support is obtained by summing over all sequences that select exactly its elements:

$$p(Z) = \sum_{\boldsymbol{z}:\,\{z_1,\ldots,z_K\}=Z} q(\boldsymbol{z}) \tag{3}$$

## 3 METHOD

We normalize each nonzero EEG window by a single RMS value over its electrodes and time samples, preserving relative scalp amplitudes (throughout this section, $\boldsymbol{Y}$ denotes this normalized input). Using the temporal DCT introduced in Section 2, we retain $D$ coefficients per electrode and represent the input as the unordered set

$$\mathcal{C} = \{(\boldsymbol{Y}_{\mathrm{D},c,:}, \boldsymbol{X}_{c,:}) : c = 1, \ldots, C\}.$$

We represent the vertex-level source field using the $P = 994$-parcel scheme of DeepSIF (Sun et al., 2022), then apply the same DCT to each parcel waveform. This yields a $P \times D$ coefficient field instead of the original $V \times T$ source field, with each row still tied to an anatomical parcel. We choose $D = 64$ because the simulated waveforms are temporally compressible, while retaining all $P$ parcels avoids the peak displacement caused by further spatial truncation. Appendix E quantifies in detail both choices.

Given $\mathcal{C}$, the model predicts a coefficient field $\widehat{\boldsymbol{S}}_{\mathrm{D}} \in \mathbb{R}^{P\times D}$ and a distribution $p_\theta(Z \mid \mathcal{C})$ over active parcel supports. The field specifies a candidate waveform for each parcel, while the support distribution describes the probability of parcels active together for a given $\mathcal{C}$.

To construct the training targets, we average the activity of active vertices within each parcel, set inactive parcels to zero, and normalize each nonzero source field by its global RMS. Applying the same DCT yields the target coefficients $\boldsymbol{S}_{\mathrm{D}}$, the target support is determined from the simulated activity before averaging or temporal compression, since either operation can cancel activity.

For a sampled support $Z^{(j)}$, the corresponding reconstruction retains the selected rows of the predicted field:

$$\widehat{\boldsymbol{S}}^{(j)}_{\mathrm{D},r,:} = \mathbf{1}[r \in Z^{(j)}]\widehat{\boldsymbol{S}}_{\mathrm{D},r,:}, \tag{4}$$

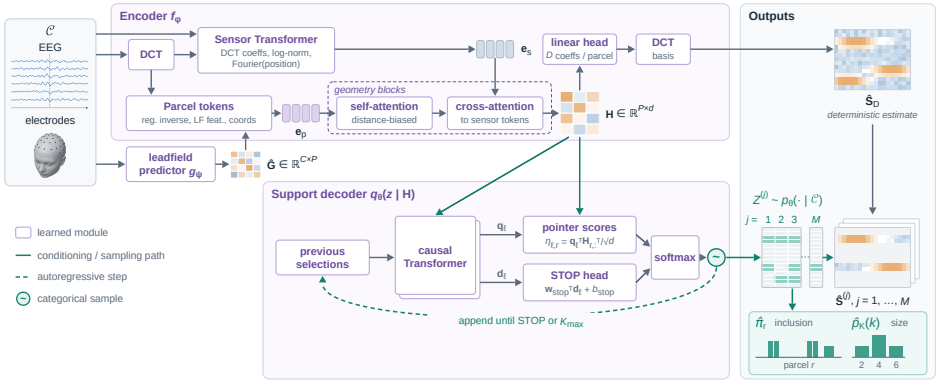



Figure 1: Overview of the model. Electrode coordinates provide the predicted leadfield used by the imager to combine EEG measurements into parcel representations $\boldsymbol{H}$ and a shared coefficient field $\widehat{\boldsymbol{S}}_{\mathrm{D}}$. The decoder samples supports conditioned on $\boldsymbol{H}$; each support selects rows of the coefficient field through equation 4. The auxiliary support head used during training is omitted.

where $\mathbf{1}[\cdot]$ is the indicator function. The shared field $\widehat{\boldsymbol{S}}_{\mathrm{D}}$ remains fixed across draws, so the sampled uncertainty concerns the composition and cardinality of the support. Temporal reconstructions follow by multiplication with $\boldsymbol{\Phi}$. For reporting, each nonzero support-restricted field is rescaled by a single factor to unit RMS, preserving relative amplitudes and waveform shapes within the draw; an all-zero field remains zero.

### 3.1 Model Architecture

The architecture comprises a geometry-conditioned leadfield predictor, a parcel encoder, and an autoregressive decoder:

$$\widehat{\mathbf{G}} = g_\psi(\mathcal{C}), \qquad (\mathbf{H}, \widehat{\mathbf{S}}_{\mathrm{D}}) = f_\phi(\mathcal{C}, \widehat{\mathbf{G}}), \qquad Z \sim p_\theta(\cdot \mid \mathcal{C}). \tag{5}$$

Here $g_\psi$ denotes the leadfield predictor and $f_\phi$ the encoder, while $p_\theta$ is the set distribution induced by the decoder. $f_\phi$ encodes $\mathcal{C}$ into learned parcel features $\boldsymbol{H} \in \mathbb{R}^{P\times d}$, which condition the autoregressive decoder. The resulting sequence distribution $q_\theta(\boldsymbol{z} \mid \boldsymbol{H})$ induces $p_\theta(\, Z \mid \mathcal{C})$ by summing over all selection orders yielding $Z$ as in Eq. 3. In the following paragraphs, we analyze more deeply the model's sub-components and implementation details.

**Montage-agnostic conditioning.** We construct sensor and parcel tokens from the EEG and electrode geometry by first projecting the EEG onto the temporal DCT basis as seen in Eq. 2. For each electrode, its DCT coefficients, their log norm, and Fourier features of its 3D position, inspired by El Ouahidi et al. (2025), are processed by a shared Transformer with masked sensor self-attention to produce $\boldsymbol{e}_s$. With no channel-index embeddings, jointly reordering EEG channels, positions, and masks simply reorders these sensor tokens. In parallel, a leadfield predictor $g_\psi$, fitted to precomputed forward models, maps electrode positions to an approximate leadfield $\widehat{\boldsymbol{G}} \in \mathbb{R}^{C\times P}$. Regularized inverse estimates computed from $\widehat{\boldsymbol{G}}$ and the projected EEG are combined with leadfield-derived features and cortical coordinates to form one token $\boldsymbol{e}_p$ per parcel. These shared operations accept different electrode counts without a supplied leadfield, within the fixed anatomy and coordinate system used in training.

**Parcel imager.** The geometry blocks refine the parcel tokens $\boldsymbol{e}_p$ through distance-biased self-attention among parcels and masked cross-attention to the sensor tokens $\boldsymbol{e}_s$. This lets each parcel integrate evidence from the available EEG channels with spatial context from other cortical locations. Their final normalized states form $\boldsymbol{H} \in \mathbb{R}^{P\times d}$, which conditions the support decoder. Two linear heads map $\boldsymbol{H}$ to parcel support logits and a field of $D$ temporal coefficients per parcel. The logits score parcel activity, while the coefficient field $\widehat{\boldsymbol{S}}_{\mathrm{D}}$ specifies a candidate waveform for every parcel after multiplication by the DCT basis. This field is the imager's shared deterministic estimate

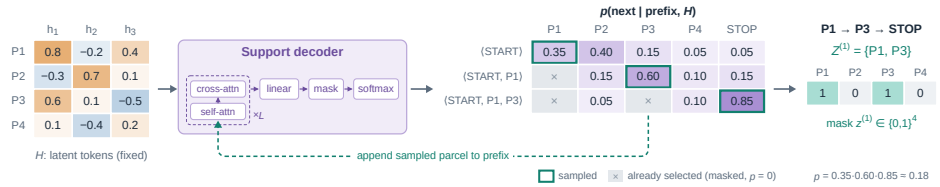



Figure 2: Autoregressive support sampling example ($P = 4$). With $\boldsymbol{H}$ fixed, the decoder selects one parcel per step from the distribution conditioned on the prefix; selected parcels are masked ($\times$) and the sampled choice (outlined) is appended until STOP. The sequence $P_1 \to P_3 \to$ STOP ($q_\theta = 0.18$) yields the support $Z^{(1)} = \{P_1, P_3\}$.

that the decoder will subsequently use to select which parcel rows to retain without changing their predicted waveforms.

**Autoregressive support selection.** The support decoder uses the internal representation $\boldsymbol{H}$ to select active parcels one at a time. At each selection step $\ell$, a causal Transformer attends to the previous selections and $\boldsymbol{H}$, producing a state $\boldsymbol{d}_\ell$ and a query $\boldsymbol{q}_\ell$. Following Vinyals et al. (2015), the query scores each parcel, while a separate head scores STOP:

$$\eta_{\ell,r} = \frac{\boldsymbol{q}_\ell^\top \boldsymbol{H}_{r,:}^\top}{\sqrt{d}}, \qquad \eta_{\ell,\text{stop}} = \boldsymbol{w}_{\text{stop}}^\top \boldsymbol{d}_\ell + b_{\text{stop}}.$$

A softmax assigns probabilities to the remaining parcels and STOP, and the next choice is sampled from this distribution. Selection ends at STOP, which is available from the first step, or after $K_{\max}$ parcels. Figure 2 illustrates these sampling process and the resulting support mask in a four-parcel example. Repeating this process $M$ times with $\boldsymbol{H}$ fixed gives alternative supports $Z^{(j)} \sim p_\theta(\cdot \mid \mathcal{C})$. Each support defines a binary mask $\mathbf{1}[r \in Z^{(j)}]$ that retains or zeros entire parcel waveforms from the shared deterministic estimate as in Eq. 4

The samples therefore vary in active parcels, while a parcel retains the same estimated waveform whenever it is selected. We estimate parcel inclusion and support-size probabilities from the sampled masks:

$$\widehat{\pi}_r = \frac{1}{M}\sum_{j=1}^{M} \mathbf{1}[r \in Z^{(j)}], \qquad \widehat{p}_K(k) = \frac{1}{M}\sum_{j=1}^{M} \mathbf{1}[|Z^{(j)}| = k].$$

Complete supports also show which parcels are selected together, information that their individual marginals do not contain.

### 3.2 Training

**Synthetic data.** We train on simulated EEG–source pairs and reserve real recordings for evaluation. SEREEGA (Krol et al., 2018) provides controlled source waveforms and spatial patterns, while The Virtual Brain (TVB) (Sanz Leon et al., 2013) generates activity through neural-mass dynamics. We vary several metrics like source locations, extents, counts, temporal dynamics, electrode configurations, background activity, and acquisition noise to provide more data differentiation and variability. Note that each parcel-level training example $(\mathcal{C}; \boldsymbol{S})$ supplies $\boldsymbol{S}_{\text{D}}$ and $Z$. Additional information about the acquisition model and sampling distributions are given in Appendix A. We also provide real ground-truth validation through the Localize-MI dataset (Mikulan et al., 2020), where intracranial stimulation provides known source locations and is used only for evaluation (Section 4).

**Imager objective.** We first train the encoder, whose parcel states $\boldsymbol{H}$ will later condition the support decoder. Its support logits $\boldsymbol{u}$ identify active parcels, while its deterministic field $\widehat{\boldsymbol{S}}_{\text{D}}$ captures their relative activity over time. Because the EEG and target souce fields use different amplitude normalizations, we compare prediction and taget after fitting one signed scale factor per example. Let $Z$ be the target set of active parcels. Its binary mask $\mathbf{1}_Z \in 0, 1^P$ has entry $(\mathbf{1}_Z)_r = 1$ when

$r \in Z$ and 0 otherwise, then, the per-example loss is

$$\mathcal{L}_{\text{img}} = \lambda_{\text{sup}}\,\text{FL}(\boldsymbol{u}, \mathbf{1}_Z) + \lambda_{\text{field}}\frac{|\widehat{\alpha}\widehat{\boldsymbol{S}}_{\text{D}} - \boldsymbol{S}_{\text{D}}|_F^2}{\max(|\boldsymbol{S}_{\text{D}}|_F^2, \varepsilon)},\ \widehat{\alpha} \quad = \frac{\langle \widehat{\boldsymbol{S}}_{\text{D}}, \boldsymbol{S}_{\text{D}} \rangle_F}{\max(|\widehat{\boldsymbol{S}}_{\text{D}}|_F^2, \varepsilon)}.$$

Here FL is focal binary cross-entropy averaged over parcels, $\lambda_{\text{sup}}$ and $\lambda_{\text{field}}$ weight the two terms, and $\varepsilon > 0$ stabilizes the ratios. The focal term downweights easy decisions, especially the many inactive parcels in a sparse target. The field term penalizes errors in relative parcel activity and waveforms after aligning global scale and sign; division by target field energy keeps examples with different source strengths comparable. We average this loss over the batch.

**Decoder objective.** We then freeze the encoder and train the decoder on its parcel states $\boldsymbol{H}$. Each target support is converted into a sequence by ordering its active parcels by decreasing energy of their target DCT coefficients, retaining at most $K_{\max}$ parcels, and appending STOP. With teacher forcing (Williams & Zipser, 1989), the decoder receives the correct preceding parcels and learns to predict the next one. The resulting loss is

$$\mathcal{L}_{\text{dec}} = -\frac{1}{\sum_i L_i}\sum_i \sum_{\ell=1}^{L_i} \log \rho_\theta \left( z_\ell^{(i)} \mid z_{<\ell}^{(i)}, \boldsymbol{H}^{(i)} \right)$$

where $\boldsymbol{z}^{(i)}$ is the target sequence, $L_i$ includes its terminal STOP token, and $\rho_\theta$ denotes the next-token probabilities computed from the decoder's parcel and STOP scores. Each term penalizes a low probability for the correct next choice; the STOP term teaches the decoder when the support is complete. We average over all valid decisions in the batch, excluding padding. This objective uses one energy-defined ordering for each target support rather than summing over its possible permutations. At inference, the decoder feeds back its sampled choices, excludes parcels already selected, and samples the next token from the remaining probabilities.

# 4 Experiments

We structure the evaluation along three axes: source reconstruction under matched acquisition, generalisation to unseen electrode montages, and predictive uncertainty over cortical supports.

## 4.1 Experimental setup

**Data and splits.** For each montage, we generate 100k synthetic scenarios with SEREEGA and 100k with TVB (Section 3.2) and train the main learned methods on both simulators. In the matched 256-channel comparison, a hash-based split places 90,018 scenarios per simulator (180,036 total) in training; the remainder is used to select checkpoints. Synthetic evaluation uses separately generated, scenario-disjoint 10k sets per simulator and montage, with different seeds and a rebalanced source-count distribution (Appendix C.7). We additionally evaluate on Localize-MI (Mikulan et al., 2020), our only real-EEG benchmark with known intracranial stimulation sites as source-location ground truth. Neither the synthetic test sets nor Localize-MI informed choices about the model or training pipeline.

**Baselines and ablations.** The classical baselines are MNE (H"am"al"ainen & Ilmoniemi, 1994), sLORETA (Pascual-Marqui et al., 1994), and dSPM (Dale et al., 2000); the learned baselines are ConvDip (Hecker et al., 2021) and DeepSIF (Sun et al., 2022). Additionally, we add SESAME (Luria et al., 2024), a sequential Monte Carlo sampler over dipole configurations, as a Bayesian baseline scored with the same distributional metrics. Our full model combines the imager with an autoregressive support decoder (Ours$^{\text{AR}}$); the imager alone (Ours$^{\text{Det}}$) isolates the decoder's contribution. To test whether that contribution depends on the imager, we train a support head on DeepSIF under the same objective as ours and add the same decoder architecture (DeepSIF$^{\text{AR}}$). Replacing autoregression with independent support decisions (Ours$^{\text{Ind}}$) then tests the role of dependencies. For detection metrics, ConvDip and the classical methods use reconstructed amplitudes as scores. Appendix D ablates the remaining design choices: the number of posterior draws, the reconstructed leadfield, the physics prior (Table 10), and the training simulators.

Table 2: **Source reconstruction at 256 channels.** On synthetic data, all methods are evaluated on the same scenario-disjoint test set, with SEREEGA and TVB results pooled. LE and SD are medians, and detection metrics are means (Section 4.2). Localize-MI comprises 61 real-EEG stimulation runs: peak–site is the median Euclidean distance from the reconstructed peak to the stimulation site (Mikulan et al., 2020), and AUPRC is averaged over runs. Classical Localize-MI values come from the benchmark's subject-specific head-model results; no parcel scores are available for AUPRC. Autoregressive predictions use $S = 50$ draws on synthetic data and $S = 100$ on Localize-MI. Learned results show mean $\pm$ standard deviation over three seeds; bold metric values mark the best in each column. $^{\dagger}$SESAME was not run on Localize-MI and thresholded scores are omitted rather than unavailable: it returns a median of 4.1 parcels against 13.1 true, which caps recall by construction and makes them a comparison of source models, not of estimation quality.

| | Synthetic test sets | | | | | | Localize-MI | |
|---|---|---|---|---|---|---|---|---|
| **Method** | LE (mm) ↓ | AUPRC ↑ | Prec. ↑ | Rec. ↑ | F1 ↑ | SD (mm) ↓ | Peak–site (mm) ↓ | AUPRC ↑ |
| MNE | 60.3 | 0.157 | 0.154 | 0.189 | 0.150 | 5.4 | 34.8 | — |
| sLORETA | 39.8 | 0.296 | 0.315 | 0.279 | 0.261 | 25.4 | 34.6 | — |
| dSPM | 35.7 | 0.156 | 0.136 | 0.333 | 0.173 | 45.7 | 28.4 | — |
| SESAME$^{\dagger}$ | 51.9 | 0.227 | — | — | — | 144.5 | — | — |
| ConvDip | $22.5_{\pm 0.9}$ | $0.468_{\pm 0.019}$ | $0.322_{\pm 0.002}$ | $0.519_{\pm 0.036}$ | $0.332_{\pm 0.007}$ | $35.3_{\pm 0.6}$ | $26.4_{\pm 0.3}$ | $0.051_{\pm 0.006}$ |
| DeepSIF | $11.2_{\pm 0.1}$ | $0.749_{\pm 0.001}$ | $0.687_{\pm 0.004}$ | $0.687_{\pm 0.003}$ | $0.642_{\pm 0.000}$ | $5.0_{\pm 0.0}$ | $57.6_{\pm 5.0}$ | $0.055_{\pm 0.014}$ |
| DeepSIF$^{\mathrm{AR}}$ | $11.9_{\pm 0.3}$ | $0.720_{\pm 0.004}$ | $0.713_{\pm 0.008}$ | $0.687_{\pm 0.006}$ | $0.659_{\pm 0.001}$ | $3.1_{\pm 0.1}$ | $62.6_{\pm 8.9}$ | $0.054_{\pm 0.016}$ |
| Ours$^{\mathrm{Det}}$ | $\mathbf{9.7_{\pm 0.0}}$ | $\mathbf{0.787_{\pm 0.005}}$ | $\mathbf{0.744_{\pm 0.013}}$ | $0.712_{\pm 0.011}$ | $0.692_{\pm 0.004}$ | $3.7_{\pm 0.1}$ | $22.9_{\pm 0.9}$ | $\mathbf{0.093_{\pm 0.010}}$ |
| **Ours**$^{\mathrm{AR}}$ | $11.1_{\pm 0.5}$ | $0.754_{\pm 0.006}$ | $0.737_{\pm 0.003}$ | $\mathbf{0.720_{\pm 0.003}}$ | $\mathbf{0.705_{\pm 0.003}}$ | $\mathbf{2.9_{\pm 0.0}}$ | $\mathbf{22.3_{\pm 1.4}}$ | $0.090_{\pm 0.011}$ |

**Model selection and reporting.** Validation loss selects checkpoints for each learned method. We report performance as the mean and standard deviation over three training seeds; deterministic inverse methods are evaluated once. For sampled predictions in the main synthetic comparisons, we fix $S = 50$ draws per observation. The draw-count analysis and full evaluation protocol are given in Appendices D.1 and C, respectively.

### 4.2 Evaluation metrics

On synthetic data, all methods are evaluated against vertex-derived truth in the same 994-parcel cortical space (Appendix C.1). Localisation error (LE, mm) is the Euclidean distance from the highest-scoring parcel to the parcel containing the strongest true vertex, measured between parcel centroids as in DeepSIF and ConvDip; distance to the nearest active parcel can mask errors for extended sources. We assess support recovery with area under the precision-recall curve (AUPRC), our primary detection metric given the sparse active support (1.3% of parcels), and with precision, recall, and F1. For the thresholded metrics, each method's F1-maximising operating point is fitted on a subset of each synthetic test set and evaluated on disjoint samples. Spatial dispersion (SD, mm) measures estimated energy away from the true support, capturing diffuse or misplaced reconstructions that peak error alone misses (Sun et al., 2022). Across observations, LE and SD are summarised by the median and detection metrics by the mean; definitions and threshold details are in Appendices C.2 and C.4. Appendix C.3 repeats the Table 2 distances along the cortical surface, which separate the banks of a sulcus and the two hemispheres; on synthetic data Ours$^{\mathrm{Det}}$ keeps the lowest LE, and the gap to the diffuse classical and ConvDip reconstructions widens.

### 4.3 Source reconstruction

Along the first axis, we evaluate source reconstruction under matched acquisition, measuring localisation and spatial extent on the same 256-channel montage. We compare our full imager-decoder model with classical and learned baselines, using its imager-only ablation to isolate the decoder's contribution and DeepSIF$^{\mathrm{AR}}$ to test whether the same decoding mechanism extends to another imager.

Table 3: **Montage generalisation.** Ours$^{\mathrm{Det}}$ is trained jointly on 32, 64, 90, and 128 channels in the upper block and only at 256 channels in the lower block. Synthetic rows pool SEREEGA and TVB, with the first row also pooling the four seen layouts. Distance is median LE on synthetic data and median peak–site distance on Localize-MI; AUPRC and F1 are means, and SD is a median. Localize-MI F1 and SD are provisional single-site proxies, not directly comparable to synthetic support scores.

| **Evaluation set** | **Distance (mm)**↓ | **AUPRC**↑ | **F1**↑ | **SD (mm)**↓ |
|---|---|---|---|---|
| *Jointly trained on 32, 64, 90, and 128 channels* | | | | |
| Seen layouts (pooled) | $10.6_{\pm 0.1}$ | $0.735_{\pm 0.005}$ | $0.638_{\pm 0.005}$ | $4.4_{\pm 0.2}$ |
| 76 channels (unseen) | $10.4_{\pm 0.1}$ | $0.747_{\pm 0.005}$ | $0.648_{\pm 0.005}$ | $4.2_{\pm 0.2}$ |
| 21 channels (unseen) | $16.5_{\pm 0.1}$ | $0.521_{\pm 0.005}$ | $0.460_{\pm 0.007}$ | $8.8_{\pm 0.4}$ |
| 256 channels (unseen) | $13.0_{\pm 0.2}$ | $0.598_{\pm 0.006}$ | $0.527_{\pm 0.007}$ | $7.8_{\pm 0.2}$ |
| Localize-MI | $22.4_{\pm 1.0}$ | $0.102_{\pm 0.004}$ | $0.039_{\pm 0.007}$ | $56.9_{\pm 1.5}$ |
| *Separately trained on 256 channels* | | | | |
| 256 channels (seen) | $9.7_{\pm 0.0}$ | $0.787_{\pm 0.005}$ | $0.692_{\pm 0.004}$ | $3.7_{\pm 0.1}$ |
| Localize-MI | $22.9_{\pm 0.9}$ | $0.093_{\pm 0.010}$ | $0.039_{\pm 0.005}$ | $53.0_{\pm 1.5}$ |

Table 2 shows that better peak localisation and parcel ranking need not coincide with better thresholded support recovery. On synthetic data, Ours$^{\mathrm{Det}}$ has the lowest median LE (9.7 mm), highest AUPRC, and highest precision, whereas Ours$^{\mathrm{AR}}$ has the highest recall and F1 and the lowest SD. DeepSIF$^{\mathrm{AR}}$ shows the same F1 and SD gains at the cost of LE and AUPRC, with recall unchanged at the reported precision, so the trade-off is not specific to our imager. On Localize-MI, both our variants have lower peak–site distances than every baseline, including classical methods using subject-specific head models, while Ours$^{\mathrm{Det}}$ has the highest mean AUPRC. The gap to DeepSIF is particularly revealing: despite comparable synthetic LE, our full model places the peak much closer to the stimulation site on real EEG.

### 4.4 Montage generalisation

With fixed-montage reconstruction established, we next test whether our montage-agnostic imager transfers to electrode layouts it has not seen during training. We train a single imager jointly on four layouts with 32, 64, 90, and 128 channels. We then evaluate it on three unseen layouts: 76 channels lie within the training density range and use known electrode positions, 21 channels use known positions below that range, and 256 channels exceed it and introduce new positions.

Table 3 shows that synthetic performance on the unseen 76-channel layout is close to that across the training layouts, indicating transfer to a new configuration of familiar electrodes. The 21- and 256-channel layouts pose harder tests through sparse coverage and new electrode positions, respectively; at 256 channels, an oracle leadfield nearly recovers the matched model's synthetic performance (Appendix D.2). On Localize-MI, however, the multi-montage imager at the unseen 256-channel layout has a peak–site distance comparable to the separately trained 256-channel imager and a slightly higher mean AUPRC. These differences do not establish superiority, but show that the model retains real-EEG localisation without 256-channel training.

### 4.5 Predictive uncertainty

The first two axes ask where each reconstruction places cortical activity; the third asks whether alternative supports sampled by the decoder express useful uncertainty. Because Localize-MI provides a stimulation site but not the full active support, we assess this question on held-out synthetic data. We measure joint predictive quality with the binary-support energy score, error ranking with excess AURC, and pooled parcelwise marginal agreement with ECE (Appendix C.5).

Table 4 shows that the decoder adds predictive information beyond a single support: Ours$^{\mathrm{AR}}$ lowers the energy score from 2.519 for the deterministic imager to 1.576, and also outperforms DeepSIF$^{\mathrm{AR}}$. Its energy-score advantage over Ours$^{\mathrm{Ind}}$ is smaller, but it has substantially lower excess AURC (55.18 versus 83.11), consistent with better identification of poorly localised cases. Ours$^{\mathrm{AR}}$ also

Table 4: **Predictive uncertainty at 256 channels.** All methods use the same 1k-scenario subset per synthetic test set. Learned results show mean ± standard deviation over three seeds; sampled rows use $S = 50$ draws. Energy score compares binary support (a point mass for Ours$^{\mathrm{Det}}$); ECE is in $10^{-3}$ units.

| **Method** | **Energy score** ↓ | **Excess AURC (mm)** ↓ | **ECE** $\times 10^{-3}$ ↓ |
|---|---|---|---|
| SESAME | 3.588 | 86.88 | 15.35 |
| DeepSIF$^{\mathrm{AR}}$ | $1.684_{\pm 0.007}$ | $64.76_{\pm 3.69}$ | $1.33_{\pm 0.27}$ |
| Ours$^{\mathrm{Det}}$ | $2.519_{\pm 0.039}$ | $64.27_{\pm 3.41}$ | N/A |
| Ours$^{\mathrm{Ind}}$ | $1.585_{\pm 0.007}$ | $83.11_{\pm 5.12}$ | $2.05_{\pm 0.10}$ |
| **Ours$^{\mathrm{AR}}$** | $\mathbf{1.576_{\pm 0.006}}$ | $\mathbf{55.18_{\pm 0.98}}$ | $\mathbf{0.63_{\pm 0.04}}$ |

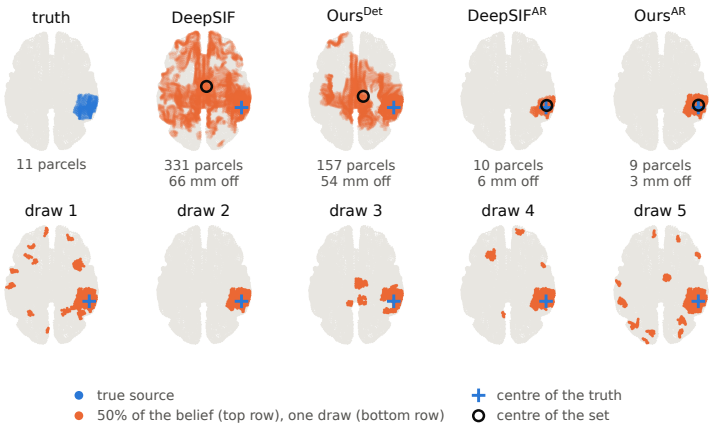



Figure 3: **Marginal mass and alternative supports.** A selected 256-channel SEREEGA example. The top row shows the true support in blue and each method's smallest parcel set holding 50% of its normalised inclusion mass in orange; labels give set size and centroid error. Blue + and black ○ mark the true and estimated set centroids. The bottom row shows five spatially diverse supports selected from 50 Ours$^{\mathrm{AR}}$ draws.

has the lowest pooled ECE, although the many inactive parcels limit what this aggregate reveals about calibration. Because the independent decoder learns different marginals, these comparisons establish the performance of the full autoregressive model rather than isolating the effect of sole dependencies alone. The decoder concentrates support mass near the true sources: the median error of the 50%-mass-set centroid falls from 32.3 mm for the imager's parcelwise head to 8.6 mm, with a closer centroid in 86% of cases. Peak error, however, rises from 9.8 to 11.1 mm. In the 200 held-out SEREEGA cases analysed for Figure 3, the two parcelwise heads need a median of 148–330 parcels to hold half their inclusion mass, versus 8–9 for the decoded distributions; the selected case shows individual draws preserving the main source region while varying secondary parcels.

## 5 Conclusions

We introduced a geometry-conditioned EEG source imager with autoregressive sampling of variable-size parcel supports, requiring no supplied leadfield at inference. Trained entirely on synthetic data, the full model achieves a median peak-to-site distance of 22.3 mm on Localize-MI, demonstrating transfer to stimulation-grounded real EEG. A single multi-montage imager also transfers without retraining to unseen electrode layouts. On held-out synthetic data, autoregressive sampling improves support F1 and lowers binary-support energy score and excess AURC relative to the deterministic imager. The evaluated scope remains limited to a fixed anatomical representation and uncertainty over support membership; waveform uncertainty and calibration on real EEG remain unassessed. Appendix F discusses these limitations and the interpretation of the baseline comparisons.

### AI USE STATEMENT

We used generative AI to assist with manuscript drafting and language editing, literature retrieval and discovery, and research ideation and execution, including feedback on mathematical arguments and experimental design. The authors reviewed the AI-assisted work and take responsibility for the final text, analyses, and claims.

### REPRODUCIBILITY STATEMENT

The main text and appendices document the model architecture and training objectives, the SEREEGA and TVB data-generation pipelines, preprocessing, experimental splits, and evaluation metrics. We report learned-model results across three training seeds and describe the protocols for held-out montages and Localize-MI. The implementation is available anonymously at `https://anonymous.4open.science/r/eeg-synthgen/`.

## REFERENCES

Facundo Costa, Hadj Batatia, Thomas Oberlin, Carlos D'Giano, and Jean-Yves Tourneret. Bayesian EEG source localization using a structured sparsity prior. *NeuroImage*, 144:142–152, 2017. doi: 10.1016/j.neuroimage.2016.08.064. URL https://doi.org/10.1016/j.neuroimage.2016.08.064.

Anders M. Dale, Arthur K. Liu, Bruce R. Fischl, Randy L. Buckner, John W. Belliveau, Jeffrey D. Lewine, and Eric Halgren. Dynamic statistical parametric mapping: Combining fMRI and MEG for high-resolution imaging of cortical activity. *Neuron*, 26(1):55–67, 2000. doi: 10.1016/S0896-6273(00)81138-1. URL https://doi.org/10.1016/S0896-6273(00)81138-1.

Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. Reve: A foundation model for eeg-adapting to any setup with large-scale pretraining on 25,000 subjects. In *Advances in Neural Information Processing Systems*, volume 38, pp. 22541–22577, 2025.

Yonatan Geifman, Guy Uziel, and Ran El-Yaniv. Bias-reduced uncertainty estimation for deep neural classifiers. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SJfb5jCqKm.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

Alexandre Gramfort, Matthieu Kowalski, and Matti H"am"al"ainen. Mixed-norm estimates for the M/EEG inverse problem using accelerated gradient methods. *Physics in Medicine and Biology*, 57(7):1937–1961, 2012. doi: 10.1088/0031-9155/57/7/1937. URL https://doi.org/10.1088/0031-9155/57/7/1937.

Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, and Matti H"am"al"ainen. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, 7, 2013. doi: 10.3389/fnins.2013.00267. URL https://doi.org/10.3389/fnins.2013.00267.

Matti S. H"am"al"ainen and Risto J. Ilmoniemi. Interpreting magnetic fields of the brain: minimum norm estimates. *Medical & Biological Engineering & Computing*, 32(1):35–42, 1994. doi: 10.1007/BF02512476. URL https://doi.org/10.1007/BF02512476.

Bin He, Abbas Sohrabpour, Emery Brown, and Zhongming Liu. Electrophysiological source imaging: A noninvasive window to brain dynamics. *Annual Review of Biomedical Engineering*, 20(1):171–196, 2018. doi: 10.1146/annurev-bioeng-062117-120853. URL https://doi.org/10.1146/annurev-bioeng-062117-120853.

Lukas Hecker, Rebekka Rupprecht, Ludger Tebartz Van Elst, and J"urgen Kornmeier. ConvDip: A convolutional neural network for better EEG source imaging. *Frontiers in Neuroscience*, 15:569918, 2021. doi: 10.3389/fnins.2021.569918. URL https://doi.org/10.3389/fnins.2021.569918.

Ben H Jansen and Vincent G Rit. Electroencephalogram and visual evoked potential generation in a mathematical model of coupled cortical columns. *Biological cybernetics*, 73(4):357–366, 1995.

Laurens R. Krol, Juliane Pawlitzki, Fabien Lotte, Klaus Gramann, and Thorsten O. Zander. SEREEGA: Simulating event-related EEG activity. *Journal of Neuroscience Methods*, 309:13–24, 2018. doi: 10.1016/j.jneumeth.2018.08.001. URL https://doi.org/10.1016/j.jneumeth.2018.08.001.

Eric Larson et al. MNE-Python, 2026. URL https://doi.org/10.5281/zenodo.22712317.

Gianvittorio Luria, Alessandro Viani, Annalisa Pascarella, Harald Bornfleth, Sara Sommariva, and Alberto Sorrentino. The SESAMEEG package: a probabilistic tool for source localization and uncertainty quantification in M/EEG. *Frontiers in Human Neuroscience*, 18:1359753, 2024. doi: 10.3389/fnhum.2024.1359753. URL https://doi.org/10.3389/fnhum.2024.1359753.

Christoph M. Michel and Denis Brunet. EEG source imaging: A practical review of the analysis steps. *Frontiers in Neurology*, 10:325, 2019. doi: 10.3389/fneur.2019.00325. URL https://doi.org/10.3389/fneur.2019.00325.

Ezequiel Mikulan, Simone Russo, Sara Parmigiani, Simone Sarasso, Flavia Maria Zauli, Annalisa Rubino, Pietro Avanzini, Anna Cattani, Alberto Sorrentino, Steve Gibbs, Francesco Cardinale, Ivana Sartori, Lino Nobili, Marcello Massimini, and Andrea Pigorini. Simultaneous human intracerebral stimulation and HD-EEG, ground-truth for source localization methods. *Scientific Data*, 7:127, 2020. doi: 10.1038/s41597-020-0467-x.

Joseph SB Mitchell, David M Mount, and Christos H Papadimitriou. The discrete geodesic problem. *SIAM Journal on Computing*, 16(4):647–668, 1987.

Roberto D. Pascual-Marqui, Christoph M. Michel, and Dietrich Lehmann. Low resolution electromagnetic tomography: a new method for localizing electrical activity in the brain. *International Journal of Psychophysiology*, 18(1):49–65, 1994. doi: 10.1016/0167-8760(84)90014-X. URL https://doi.org/10.1016/0167-8760(84)90014-X.

François Perrin, Jacques Pernier, Olivier Bertrand, and Jean Francois Echallier. Spherical splines for scalp potential and current density mapping. *Electroencephalography and clinical neurophysiology*, 72(2):184–187, 1989.

Tom Pollard, Benjamin E. Moody, Li-wei H. Lehman, Brian J. Gow, Chrystinne Fernandes, Chen Xie, Alistair Johnson, Roger G. Mark, and Thomas Heldt. PhysioNet as a global platform for biomedical research. *Nature Health*, 1:792–795, August 2026. ISSN 3005-0693. doi: 10.1038/s44360-026-00096-z. URL https://doi.org/10.1038/s44360-026-00096-z.

Sarah Reynaud, Adrien Merlini, Douraied Ben Salem, and François Rousseau. Comprehensive analysis of supervised learning methods for electrical source imaging. *Frontiers in Neuroscience*, 18:1444935, 2024. doi: 10.3389/fnins.2024.1444935. URL https://doi.org/10.3389/fnins.2024.1444935.

Jesse Rong, Rui Sun, Boney Joseph, Greg Worrell, and Bin He. Deep learning-based EEG source imaging is robust under varying electrode configurations. *Clinical Neurophysiology*, 175:2010730, 2025. doi: 10.1016/j.clinph.2025.04.009.

Paula Sanz Leon, Stuart A. Knock, M. Marmaduke Woodman, Lia Domide, Jochen Mersmann, Anthony R. McIntosh, and Viktor Jirsa. The virtual brain: a simulator of primate brain network dynamics. *Frontiers in Neuroinformatics*, 7, 2013. doi: 10.3389/fninf.2013.00010. URL https://doi.org/10.3389/fninf.2013.00010.

Gerwin Schalk. EEG Motor Movement/Imagery Dataset. *PhysioNet*, September 2009. doi: 10.13026/C28G6P. URL https://doi.org/10.13026/C28G6P. Version 1.0.0.

Gerwin Schalk, Dennis J McFarland, Thilo Hinterberger, Niels Birbaumer, and Jonathan R Wolpaw. Bci2000: a general-purpose brain-computer interface (bci) system. *IEEE Transactions on biomedical engineering*, 51(6):1034–1043, 2004.

Alberto Sorrentino, Gianvittorio Luria, and Riccardo Aramini. Bayesian multi-dipole modelling of a single topography in MEG by adaptive sequential Monte Carlo samplers. *Inverse Problems*, 30(4):045010, 2014. doi: 10.1088/0266-5611/30/4/045010. URL https://doi.org/10.1088/0266-5611/30/4/045010.

Rui Sun, Abbas Sohrabpour, Gregory A. Worrell, and Bin He. Deep neural networks constrained by neural mass models improve electrophysiological source imaging of spatiotemporal brain dynamics. *Proceedings of the National Academy of Sciences*, 119(31):e2201128119, 2022. doi: 10.1073/pnas.2201128119. URL https://doi.org/10.1073/pnas.2201128119.

Barry D. Van Veen, Wim Van Drongelen, Moshe Yuchtman, and Akifumi Suzuki. Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. *IEEE Transactions on Biomedical Engineering*, 44(9):867–880, 1997. doi: 10.1109/10.623056. URL https://doi.org/10.1109/10.623056.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. Pointer networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://papers.nips.cc/paper/5866-pointer-networks.

Ronald J Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280, 1989.

# A Synthetic Dataset

This appendix documents the complete data-generation protocol and the construction of the training and test datasets used in Section 4.

## A.1 General Overview

All synthetic data samples are produced via a unified generation pipeline, implemented in a dedicated library that supports two widely used source simulation backends: SEREEGA Krol et al. (2018) and The Virtual Brain (TVB) Sanz Leon et al. (2013). To ensure consistency across experiments, all data are simulated using the *fsaverage* anatomy, partitioned into a 994-parcel evaluation space, and consist of one-second windows sampled at $500\,\mathrm{Hz}$.

The two backends offer complementary approaches to source modeling: SEREEGA allows direct control over source waveforms,designed to mimic plausible EEG patterns without relying on a full physiological dynamical model, as well as over their spatial activation profiles. TVB, by contrast, generates source activity directly from a physiological model Jansen & Rit (1995). Rather than prescribing exact waveforms, the signals emerge from the dynamics of the simulated brain network, with selected regions made more excitable to create focal activity. We refer readers to the respective reference studies for exhaustive technical details on these simulators.

At a high level, each EEG sample is generated by first simulating focal cortical activity and cerebral background, projecting both to the scalp, and subsequently introducing acquisition disturbances. The complete process can be formalized as:

$$\boldsymbol{Y} = \mathcal{R}\left[\mathcal{A}\left(\boldsymbol{G}_0\boldsymbol{S} + \alpha\,\boldsymbol{G}_0\boldsymbol{S}_{\mathrm{bg}} + \boldsymbol{E}\right)\right], \tag{6}$$

where $\boldsymbol{S}$ and $\boldsymbol{S}_{\mathrm{bg}}$ denote the focal and background cortical activity, $\boldsymbol{G}_0$ is the pre-reference forward operator, $\alpha$ controls the background strength, and $\boldsymbol{E}$ denotes sensor noise. The operator $\mathcal{A}$ introduces optional electrode artifacts, and $\mathcal{R}$ applies EEG referencing. Figure 4 summarizes the generation pipeline which is further analyzed in the following subsection. The SEREEGA backend interfaces with the original MATLAB SEREEGA toolbox through MATLAB R2026a and the MATLAB Engine for Python (`matlabengine` $\geq$ 26.1), while the TVB backend relies on `tvb-library` $\geq 2.8$ and `tvb-data` $\geq 2.8$.

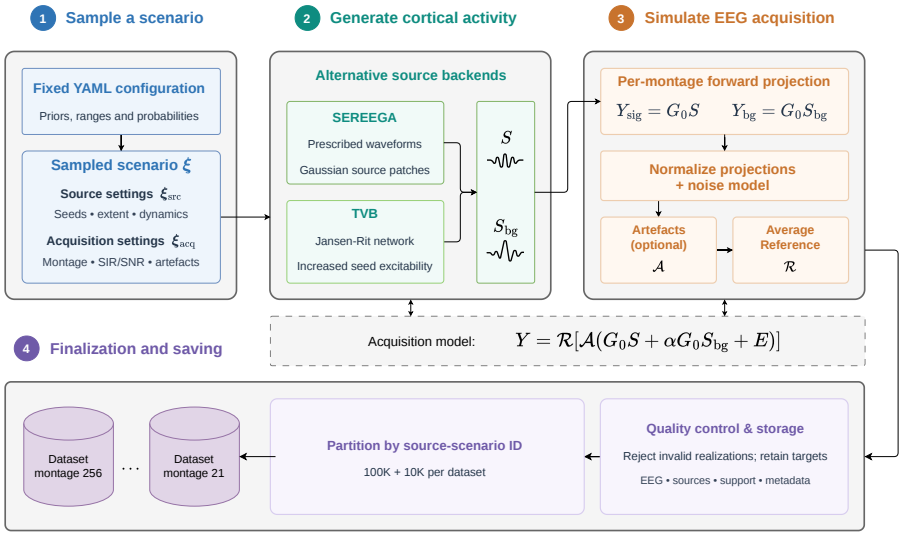



Figure 4: Overview of the synthetic data generation pipeline. The process details the simulation of focal and background cortical activity, their projection to the scalp, and the subsequent introduction of sensor noise and acquisition artifacts.

### A.2 Generation Protocol

The generation process follows a top-down approach: it begins by defining the global dataset parameters, then instantiates individual samples. The configuration phase utilizes a YAML file specifying the allowable source configurations, temporal dynamics, montages, noise conditions, and acquisition perturbations, along with their sampling probabilities and parameter ranges. This configuration remains strictly fixed for the entire dataset.

A single sample is associated with a *scenario* $\boldsymbol{\xi}$ (implemented as a Python dataclass), representing one concrete realization of these choices. While the YAML file defines the distributional bounds, the scenario records the exact selections for a specific realization: the number and location of the sources, their spatial extent and temporal parameters, the montage, and the acquisition conditions. Conceptually, this is denoted as:

$$\boldsymbol{\xi} = (\boldsymbol{\xi}_{\text{src}}, \boldsymbol{\xi}_{\text{acq}}),$$

where $\boldsymbol{\xi}_{\text{src}}$ describes the cortical source configuration and $\boldsymbol{\xi}_{\text{acq}}$ the conditions under which that source activity is observed at the scalp. The resulting source configuration $\boldsymbol{\xi}_{\text{src}}$ is passed to the selected backend $b \in \{\text{SEREEGA}, \text{TVB}\}$, which determines how the corresponding cortical activity evolves over time.

SEREEGA and TVB receive identical spatial specifications but instantiate them in fundamentally different ways. Table 5 summarizes their main simulation settings.

- **SEREEGA** assigns explicitly controlled signal classes to the samples' source patches. For each scenario:
  1. A specific temporal family is sampled for the source waveform: Oscillatory bursts (35%), Event-Related Potentials (25%), Autoregressive dynamics (20%), or Interictal-like spikes (20%).
  2. The sampled onset, frequency, and family-specific parameters in $\boldsymbol{\xi}_{\text{src}}$ instantiate the corresponding SEREEGA signal class.
  3. The resulting waveform is generated and distributed over the source patch using a Gaussian spatial profile.
  4. Independent distributed $1/f$-like cortical activity is generated to serve as cerebral background.

Table 5: Source-generation settings for the two simulation backends.

| Property | SEREEGA | TVB |
|---|---|---|
| Temporal model | Explicitly generated source waveforms | Stochastic Jansen–Rit neural-mass dynamics |
| Focal activity | Waveforms are assigned to the sampled cortical patches and spatially weighted around their seed | Seed parcels are made more excitable than the remaining network ($A = 3.5$ vs. $A = 3.25$) |
| Spatial profile | Gaussian weighting within each patch, $\sigma =$ 12–25 mm | Activity is defined at parcel level and mapped back to the corresponding cortical vertices |
| Background | Independent distributed $1/f$-like cortical activity | Activity of the same simulated neural network outside the focal source definition |
| Numerical simulation | SEREEGA MATLAB signal generators | Stochastic Heun integration, 0.5 ms step, 3 mm/ms conduction speed, 5 s warm-up |

- **TVB** generates source activity from an interacting network of Jansen–Rit neural mass models. For each scenario:
  1. Sampled source seeds are mapped to their corresponding cortical parcels.
  2. Their Jansen–Rit excitability is increased relative to the remaining network.
  3. The complete 994-region network is simulated stochastically following a 5 s warm-up phase.
  4. A one-second window is extracted, resampled to 500 Hz, and DC-corrected.
  5. Seed-parcel activity defines the focal sources, while activity from the remaining network defines the cerebral background.

Once cortical activity is generated, the second component of the scenario, $\boldsymbol{\xi}_{\mathrm{acq}}$, specifies *how that activity is observed at the scalp*. It determines the montage, cerebral-background level, sensor-noise level and model, optional electrode artifacts, and reference scheme.

Foreground and background activity are first projected independently through the leadfield for the selected montage:

$$\boldsymbol{Y}_{\mathrm{sig}} = \boldsymbol{G}_0 \boldsymbol{S}, \qquad \boldsymbol{Y}_{\mathrm{bg}} = \boldsymbol{G}_0 \boldsymbol{S}_{\mathrm{bg}}.$$

Since SEREEGA and TVB employ distinct native amplitude conventions, we normalize these raw projections prior to mixing, eliminating arbitrary absolute amplitude discrepancies. The sampled Signal-to-Interference Ratio (SIR) establishes the strength of the cerebral background relative to the focal signal, while sensor noise is calibrated independently via the Signal-to-Noise Ratio (SNR), as defined in Eq. 7:

$$\mathrm{SIR} = 20 \log_{10} \frac{\mathrm{RMS}(\boldsymbol{Y}_{\mathrm{sig}})}{\mathrm{RMS}(\alpha \boldsymbol{Y}_{\mathrm{bg}})}, \quad \mathrm{SNR} = 20 \log_{10} \frac{\mathrm{RMS}(\boldsymbol{Y}_{\mathrm{sig}})}{\mathrm{RMS}(\boldsymbol{E})}. \tag{7}$$

Both are defined relative to the foreground signal, ensuring that background and sensor noise vary independently. The complete parameter space for the acquisition pipeline, including the distribution of noise profiles and artifact injections, is detailed in Table 6. The covariance and RMS calibration statistics used for spatially correlated noise are estimated from resting recordings in the PhysioNet EEG Motor Movement/Imagery database Schalk (2009); Schalk et al. (2004); Pollard et al. (2026).

**Montage Variation.** The electrode montage forms an integral part of $\boldsymbol{\xi}_{\mathrm{acq}}$. All layouts are derived from the MNE-Python Larson et al. (2026); Gramfort et al. (2013) standard coordinate system and comprise $C \in \{21, 32, 64, 76, 90, 128, 256\}$ electrodes. Lower-density configurations are carefully subsampled to preserve whole-scalp coverage rather than naively selecting the first $C$ electrodes. Each layout possesses its own electrode coordinates and corresponding forward matrix, while the underlying anatomy and cortical source space remain constant. In the following subsection, we delineate the specific montages used for training versus those reserved for rigorous zero-shot evaluation.

Table 6: Unified parameters for the acquisition and noise pipeline ($\boldsymbol{\xi}_{\mathrm{acq}}$).

| Component | Distribution | Description |
|---|---|---|
| SIR | $\{0, 5, 10, 15, 20\}$ dB | Strength of cerebral background relative to the focal signal. |
| SNR | $\{-5, 0, 5, 10, 15, 20\}$ dB | Sensor noise strength (independent of cerebral background). |
| Noise profile | Categorical | White Gaussian (20%), $1/f$ colored noise (20%), independent AR(2) (30%), or covariance-shaped (30%). |
| Reference | Average | Applied consistently after noise and artifact generation. |
| Artifact trigger | $\mathrm{Bernoulli}(0.30)$ | 30% chance to apply a transient or persistent channel artifact. |
| $\hookrightarrow$ Ocular | Uniform (given trigger) | Gaussian transient (random onset/width) scaled to EEG RMS. |
| $\hookrightarrow$ Muscular | Uniform (given trigger) | 30–100 Hz burst on a random subset of channels. |
| $\hookrightarrow$ Line noise | Uniform (given trigger) | 50 or 60 Hz sinusoidal interference with random phase. |
| $\hookrightarrow$ Dropout | Uniform (given trigger) | A randomly selected subset of channels is zeroed out. |

### A.3 Dataset Composition and Splits

Development data were generated independently utilizing the SEREEGA and TVB backends across five core montage configurations: 32, 64, 90, 128, and 256 channels. For every combination of simulator and montage, we generated exactly 100k quality-controlled samples for the development dataset and an additional 10k samples for independent testing. The 256-channel groups are strictly used to train the matched-acquisition models, whereas montage-generalization training utilizes the pooled 32-, 64-, 90-, and 128-channel groups (see Appendix C.7).

To partition the data, a stable hash of the underlying scenario identifier strictly assigns development samples into training, validation, and internal test sets. Crucially, this split is executed at the *source scenario* level rather than independently for each simulator or montage rendering. Consequently, disparate observations originating from the same source scenario consistently reside within the same partition. This prevents data leakage, guaranteeing that the model never observes a source configuration during training under one montage and subsequently encounters a different rendering of that identical configuration during validation or testing.

For final evaluations, the 10k-sample independent test datasets are structurally completely isolated: they are not subsampled from the development corpora, but provide fresh realizations of the underlying generative distribution. To guarantee strict independence, evaluation sets were generated with distinct random configuration seeds, and we explicitly verified that the intersection of training and test scenario identifiers is perfectly empty ($\mathcal{I}_{\text{dev}} \cap \mathcal{I}_{\text{test}} = \emptyset$). These external scenarios were exclusively reserved for evaluation and never utilized for model fitting or selection.

The 21- and 76-channel layouts possess separately generated evaluation sets and are entirely held out of the montage-generalization training phase. Since their individual electrode positions intersect with those of the training layouts, they effectively test the model's robustness to novel spatial densities and combinatorial layouts; the 256-channel layout evaluates both novel combinations and entirely unseen sensor positions (Section 4.4).

### A.4 Quality Control and Reproducibility

Data generation continues sequentially until the targeted number of fully accepted samples is achieved; failed realizations are aggressively discarded. A sample is rejected if the final EEG trace contains non-finite values, if fewer than ten channels remain numerically valid post-corruption, or, for multi-source scenarios, if the minimum Euclidean distance between source seeds falls below $10\,\mathrm{mm}$. Consequently, the 100k development and 10k independent test counts per split reflect strictly post-QC samples.

Each finalized sample stores the tuple $(\boldsymbol{Y}, \boldsymbol{S}, \boldsymbol{X}, \text{support})$ alongside metadata concerning all the settings and information needed to generate the sample. $\boldsymbol{Y}$ is stored post-referencing, while $\boldsymbol{S}$ is maintained at vertex resolution prior to being reduced to 994 parcels and the DCT basis during data loading.

Full reproducibility is ensured through the release of the frozen generator codebase, alongside all necessary configuration files, fixed anatomy/leadfield/connectivity/noise banks, and precise scenario seeds.

## B Montage Agnosticism

The term montage-agnostic refers to three fundamental aspects that distinguish this implementation that we collect as follows:

1. The prediction remains invariant to the ordering of EEG channels; permuting the EEG channels and their corresponding electrode coordinates does not alter the output in source space.
2. The same model parameters can process varying numbers of electrodes, without any layer whose dimensionality depends on the number of channels $C$.
3. Inference does not require a subject- or montage-specific derivation field.

The first two properties follow directly from the architecture, while the third depends on how accurately the step operator can be inferred from the electrode geometry.

We have articulated these points explicitly because the existing literature typically addresses each aspect in isolation. In contrast, our approach addresses all of them jointly, specifically in ESI context. Many current architectures, including those tied to $C$, such as the dense layers of DeepSIF Sun et al. (2022), require tailoring to a specific channel configuration. The study that most closely aligns with our approach is ConvDip Hecker et al. (2021) which interpolates sensor measurements onto a fixed two-dimensional scalp grid before applying a convolutional neural network. Although this procedure makes the network agnostic to $C$, it still requires montage-specific preprocessing and loses the electrodes' original three-dimensional geometry. Moving from ESI architectures to more general EEG foundation models, several studies have advanced this paradigm by representing electrodes as an unordered set of spatial tokens, which inspired our approach. However, these models focus exclusively on learning input representations rather than combining them with the propagation operator and biophysical constraints, which our approach directly integrates.

**Sensor-set representation.** Each available electrode is represented as an individual token. Its features contain $D = 64$ temporal DCT coefficients, their log-norm, and Fourier features of the electrode's three-dimensional position. We then apply self-attention only over valid sensor tokens using an attention mask $\boldsymbol{m}$. Consequently, no model parameter depends on sensor identity or sensor count. Electrode positions are represented continuously rather than through learned channel-name embeddings, allowing the same encoder to process coordinates that were not assigned a discrete identity during training.

Before tokenization, the EEG is normalized using a single mask-aware RMS scale per sample,

$$a = \sqrt{\frac{\sum_{c,t} m_c Y_{c,t}^2}{T \sum_c m_c}}, \qquad \widetilde{\mathbf{Y}} = \frac{\mathbf{Y}}{a}$$

The rationale for using a scale stems from the fact that independent normalization for each channel would eliminate the relative amplitudes between the electrodes, which is precisely the information we want to preserve.

## C Evaluation and Implementation Details

We specify the ground truth, scoring rules, and evaluation conditions behind the main tables, followed by implementation details needed to reproduce the results.

### C.1 GROUND TRUTH AND THE COMMON SOURCE SPACE

Synthetic reconstructions and learned Localize-MI predictions are evaluated in the same 994-parcel source space on the `fsaverage` cortical surface. The main tables use Euclidean distances between parcel centroids, consistent with DeepSIF and ConvDip, and Euclidean peak-to-site distance on Localize-MI (Sun et al., 2022; Hecker et al., 2021; Mikulan et al., 2020); Appendix C.3 examines cortical-surface geodesics. Centre-of-mass errors are Euclidean distances between coordinates in the same millimetre frame.

Synthetic ground truth is derived from the generator's recorded vertex activity before the reduction used to form training targets. This keeps the evaluation reference independent of the model's parcel-aggregation convention. We then define:

- **support** $S^\star$: parcels containing at least one active vertex;
- **peak** $p^\star$: the parcel containing the active vertex with the highest temporal RMS;
- **centre of mass**: the RMS-weighted mean of the active vertices' parcel coordinates.

The peak is determined from vertex RMS rather than from the temporally compressed parcel target, whose aggregation can change which parcel ranks first. All methods therefore share the same peak definition; samples with no active vertex energy have no peak-based score.

Because the generator rescales the sensor signal to a target SNR without rescaling the stored sources, the evaluation does not score absolute source amplitude.

### C.2 RECONSTRUCTION METRICS

Let $s_r$ be a method's score for parcel $r$, $d(\cdot,\cdot)$ the Euclidean distance between parcel centroids, and $\widehat{S}$ the thresholded support.

**Localisation.** Localisation error is the distance from the highest-scoring parcel to the strongest true source,

$$\mathrm{LE} = d\Big(\arg\max_r s_r,\ p^\star\Big), \tag{8}$$

as reported for synthetic data in Table 2. We use the strongest true source rather than the nearest active parcel: for the two strongest learned imagers, the predicted peak lies inside the true support in 94% of samples, so nearest-support error is often zero even when the strongest source is misplaced.

**Detection.** AUPRC is average precision over the 994 parcel scores within each observation, then averaged across observations. This measures how well a method ranks parcels for each EEG recording, irrespective of its score scale across recordings. Observations with an empty or complete true support have undefined AUPRC and are excluded from that metric.

Precision, recall and $F_1$ are computed from $\widehat{S}$ for each observation and then averaged; mean $F_1$ is not recomputed from mean precision and recall.

**Extent.** Spatial dispersion follows Sun et al. (2022),

$$\mathrm{SD} = \sqrt{\frac{\sum_r d_r^2\, w_r}{\sum_r w_r}}, \qquad d_r = \min_{j \in S^\star} d(r, j), \tag{9}$$

where $w_r$ is estimated source energy: squared amplitude for point reconstructions and draw-averaged normalised energy for sampled reconstructions. Distance to the true support penalises both misplaced and diffuse energy; a perfect reconstruction scores zero.

**Summarising and pooling.** LE and SD are summarised by the median across observations, whereas detection metrics are summarised by the mean. For each seed, we pool the SEREEGA and TVB records before computing these summaries; the tables then report mean and standard deviation across three seeds for learned methods and a single value for each classical method.

### C.3 EUCLIDEAN AND GEODESIC DISTANCE

The main tables use Euclidean parcel-centroid distances for comparability with prior work. Because a straight line can cross a sulcus or the midline, we repeat the distance-based evaluation along the cortical surface to test whether this convention changes the conclusions.

Within each hemisphere, geodesics are computed between the vertices nearest the parcel centroids using the MMP algorithm (Mitchell et al., 1987); edge-graph paths would be 9% longer at the median. Because the hemispheres are disconnected on the mesh, a cross-hemisphere distance is defined as the Euclidean distance plus the largest within-hemisphere geodesic (222 mm), a penalty rather than a physical path length. Checkpoints, examples and posterior draws are unchanged between evaluations. For Localize-MI, we measure the surface distance to the parcel containing the stimulation site; no corresponding geodesic can be computed for the benchmark's classical results, which provide no parcel maps.

Table 7: **Euclidean and cortical-surface distances.** The same 256-channel reconstructions as Table 2, rescored with Euclidean parcel-centroid distance (Eucl.) or cortical-surface geodesic distance (Geod.). LE and SD are medians in mm; wrong hemisphere is the percentage of synthetic peaks opposite the strongest true source. Localize-MI uses Euclidean distance to the stimulation site and geodesic distance to its parcel. Learned entries are mean ± standard deviation over three seeds; bold marks the best value in each ranked column.

| | Synthetic test sets | | | | | Localize-MI | |
|---|---|---|---|---|---|---|---|
| **Method** | LE Eucl. (mm)↓ | LE Geod. (mm)↓ | SD Eucl. (mm)↓ | SD Geod. (mm)↓ | Wrong hemisphere (%) | Eucl. (mm)↓ | Geod. (mm)↓ |
| MNE | 60.3 | 104.4 | 5.4 | 12.9 | 34.0 | 34.8 | — |
| sLORETA | 39.8 | 66.8 | 25.4 | 62.7 | 27.9 | 34.6 | — |
| dSPM | 35.7 | 79.9 | 45.7 | 97.5 | 29.0 | 28.4 | — |
| ConvDip | $22.5_{\pm 0.9}$ | $36.8_{\pm 2.0}$ | $35.3_{\pm 0.6}$ | $78.6_{\pm 0.6}$ | $23.5_{\pm 0.2}$ | $26.4_{\pm 0.3}$ | $40.6_{\pm 2.8}$ |
| DeepSIF | $11.2_{\pm 0.1}$ | $13.6_{\pm 0.0}$ | $5.0_{\pm 0.0}$ | $8.5_{\pm 0.1}$ | $21.1_{\pm 0.1}$ | $57.6_{\pm 5.0}$ | $100.4_{\pm 14.5}$ |
| DeepSIF$^{\mathrm{AR}}$ | $11.9_{\pm 0.3}$ | $14.4_{\pm 0.5}$ | $3.1_{\pm 0.1}$ | $4.3_{\pm 0.1}$ | $20.8_{\pm 0.6}$ | $62.6_{\pm 8.9}$ | $114.2_{\pm 9.9}$ |
| Ours$^{\mathrm{Det}}$ | $\mathbf{9.7_{\pm 0.0}}$ | $\mathbf{11.6_{\pm 0.1}}$ | $3.7_{\pm 0.1}$ | $5.0_{\pm 0.1}$ | $19.4_{\pm 0.2}$ | $22.9_{\pm 0.9}$ | $40.8_{\pm 6.1}$ |
| **Ours**$^{\mathrm{AR}}$ | $11.1_{\pm 0.5}$ | $13.7_{\pm 0.9}$ | $\mathbf{2.9_{\pm 0.0}}$ | $\mathbf{3.6_{\pm 0.1}}$ | $20.9_{\pm 0.2}$ | $\mathbf{22.3_{\pm 1.4}}$ | $\mathbf{39.8_{\pm 5.6}}$ |

Table 7 preserves the main synthetic conclusions: Ours$^{\mathrm{Det}}$ has the lowest LE and Ours$^{\mathrm{AR}}$ the lowest SD under both distance definitions. Geodesic LE is roughly 20% higher for our models and DeepSIF, but at least 60% higher for ConvDip and the classical methods, whose peaks also cross hemispheres more often; Euclidean distance thus understates their larger anatomical errors. On Localize-MI, our variants remain far closer to the stimulated parcel than DeepSIF, while ConvDip is similar under this alternative metric. Euclidean values remain in the main tables for comparability with published results.

### C.4 THE OPERATING POINT

Precision, recall and $F_1$ require an operating point, but the methods' scores have different scales. For example, fitted thresholds are 0.36 for DeepSIF, 0.44 for ours and 0.95 for ConvDip, so a common numerical threshold would not be a common decision rule.

For each method and synthetic test set, we select the threshold that maximises mean per-observation $F_1$ on a hash-selected calibration slice and apply it unchanged to the disjoint scoring slice. The calibration observations are excluded from precision, recall and $F_1$, while threshold-free metrics use the full test set. These test labels determine only the reported operating point, not checkpoint or architecture selection.

The search grid for deterministic outputs extends to $0.99$, avoiding truncation of ConvDip's high optimum; the sampled-output grid ends at $0.9$, above every fitted optimum in these runs. Classical inverse methods are instead thresholded by a fitted quantile of their amplitude scores, since those scores are not probabilities.

### C.5 DISTRIBUTIONAL METRICS

The uncertainty evaluation scores distributions of sampled supports, with the imager's thresholded support included as a point-mass energy-score baseline and its spatial dispersion used for error ranking.

**Energy score.** Let $z^{(1)}, \dots, z^{(S)} \in \{0,1\}^P$ be sampled support indicators and $z^\star$ the true support. We use the sample energy score (Gneiting & Raftery, 2007),

$$\widehat{\mathrm{ES}} = \frac{1}{S}\sum_{m=1}^{S} \left\| z^{(m)} - z^\star \right\|_2 - \frac{1}{2S(S-1)} \sum_{m \neq n} \left\| z^{(m)} - z^{(n)} \right\|_2, \tag{10}$$

where self-pairs are excluded from the second sum. For a deterministic support the spread term vanishes, leaving its distance to the truth; this permits the point-support baseline in Table 4 to be scored on the same binary target and observations as the sampled methods. Its support is thresholded at the operating point fitted in the same evaluation pass.

**Marginal-mass sets.** The $\alpha$ marginal-mass set in Figure 3 is the smallest set of parcels holding $\alpha$ of the normalised inclusion mass, taken in descending order. Its size measures spatial concentration, not the coverage of a credible region: inclusion probabilities sum to expected support cardinality, and normalising them does not define a distribution over peak locations or entire supports. Even if such a set intersects a true active parcel, that overlap has no nominal $\alpha$-coverage interpretation. Full-support containment, $\Pr(Z \subseteq C \mid H)$, is a different joint event that must be evaluated directly.

**Marginal calibration.** Parcel inclusion probabilities are the sample frequencies over draws. For each simulator, expected calibration error (ECE) pools all parcel-observation pairs into ten equal-width probability bins and computes

$$\mathrm{ECE} = \sum_{b:\,|B_b|>0} \frac{|B_b|}{N_{\mathrm{pool}}} \left| \overline{\pi}_b - \overline{z}_b \right|, \tag{11}$$

where $N_{\mathrm{pool}}$ is the number of parcel-observation pairs, $\overline{\pi}_b$ the mean predicted inclusion probability in bin $b$, and $\overline{z}_b$ its observed activation frequency. The table averages the two simulator-specific ECE values per seed. Because most parcels are inactive, low-probability pairs dominate this diagnostic; a small ECE alone does not establish conditional calibration or informative uncertainty.

**Excess AURC.** To test whether uncertainty identifies difficult observations, we rank them by confidence and measure mean peak-localisation error over the most confident fraction retained. The confidence score is the mean pairwise distance between nonempty draw centroids for sampled methods, and the imager's spatial dispersion for the point-support baseline. Excess AURC subtracts the area attained by an oracle ordering of the same errors (Geifman et al., 2019); lower values indicate better error ranking. It retains the units and distribution of peak error, so differences between models cannot be attributed to confidence ranking alone. We average simulator-specific areas per seed, using graph-geodesic peak error for every row and fixed jitter to break confidence ties.

### C.6 POSTERIOR SAMPLING

Each draw is a support sampled autoregressively at temperature 1.0, without repeated parcels, until a stop symbol or the 32-parcel cap. The main synthetic comparisons use $S = 50$ draws per observation and Localize-MI uses $S = 100$; Appendix D.1 examines sensitivity to $S$. Empty draws remain in binary-support scores and inclusion probabilities, but are excluded from centroid-based quantities because they have no centroid.

The draw-count ablation also reports mean and best-of-$S$ distances from draw-support centroids to the true centre of mass. The latter uses ground truth to select a draw and is therefore an oracle diagnostic, not a deployable prediction; neither quantity establishes calibration of the support distribution.

### C.7 Evaluation conditions

**Data.** Each simulator and montage has a separately generated 10k-scenario test set with distinct seeds and rebalanced source-count weights; no scenario identifier overlaps training. The synthetic rows of Tables 2 and 3 use the same 10k scenarios per simulator at each layout. Table 4 uses a common 1k-scenario subset per simulator from the 256-channel sets, with $S = 50$ draws for sampled methods. The 256-only models are trained on both simulators at that layout, while the montage-generalisation model is trained on both simulators at 32, 64, 90 and 128 channels.

**Real data.** Localize-MI provides 61 averaged stimulation runs. We take the stimulation target as the midpoint of the stimulated contacts, transform its native-surface coordinates to `fsaverage` with a subject-specific affine, and assign it to the nearest parcel. The EEG input is the final 60 ms of the stimulation-locked epoch average, resampled to the model input length and normalised by global RMS. Recording electrodes are aligned to the target montage's head frame, then mapped to its positions with spherical-spline interpolation (Perrin et al., 1989); this gives each fixed-montage model the geometry on which it was trained.

Because only the stimulation site is known, not the complete active support, median Euclidean peak-to-site distance (Mikulan et al., 2020) and mean per-run AUPRC for the stimulated parcel are the primary real-data metrics. The site-only F1 and SD entries in Table 3 treat the stimulated parcel as the target and use an Otsu threshold; unlike their synthetic counterparts, they do not measure full-support recovery. The classical peak-to-site values in Table 2 are the benchmark's published results with subject-specific head models; these provide no parcel scores for AUPRC.

**Undefined and excluded cases.** Undefined values are excluded per metric rather than by discarding an entire observation: peak-based errors require a nonzero true source, and parcel ranking requires a nonempty, noncomplete true support. Empty predicted supports receive zero precision and recall; when a sampled support is empty, peak localisation falls back to the highest-scoring parcel.

### C.8 Implementation and reproducibility

**Shared training protocol.** Ours, DeepSIF and ConvDip use the same data splits, optimiser and validation schedule for imager training. We train with AdamW at learning rate $1.2 \times 10^{-3}$ and weight decay $0.01$ on matrix weights, using batches of 32 and eight gradient-accumulation steps for an effective batch of 256. Each micro-batch contains one simulator-montage group; accumulation balances groups for our imager and DeepSIF, whereas ConvDip follows the default group order over the same training data. After 400 warmup updates, the learning rate is halved following four validations without improvement; validation uses 2048 held-out examples per group every 256 updates, and training stops after twelve unsuccessful validations. We clip gradients at norm 1.0, use dropout 0.1 and no input augmentation, and select the checkpoint with the lowest validation field loss.

The imager objective gives equal weight to focal binary cross-entropy on parcel support ($\alpha = 0.25$, $\gamma = 2$) and scale-invariant normalised MSE on the source field. The latter fits an optimal scale per observation before comparing fields; targets are RMS-normalised and projected onto 64 DCT components. All learned inputs are normalised by global EEG RMS over valid channels, with the same rule applied by baseline adapters.

**Our model.** Our imager has width 256, six blocks, eight attention heads and one refinement step, and predicts 64 coefficients and a support logit per parcel. Its physics prior uses sLORETA at regularisation levels $0.01$, $0.05$ and $0.25$, plus log leadfield norm and log resolution diagonal per parcel. The leadfield is reconstructed from electrode positions by a frozen predictor, and the resulting operators are cached per montage; their effects are measured in Appendices D.2 and D.3.

**The support decoder.** The second stage is a two-layer transformer over frozen parcel tokens, with pointer-style scores over 994 parcels and a stop symbol (Vinyals et al., 2015). Training targets are ordered by descending source energy and capped at 32 parcels, covering 98% of the synthetic test supports. We train with teacher forcing, learning rate $10^{-3}$, batch size 16 and a fixed 12k-update budget; validation scores 1024 examples every 500 updates and selects the best checkpoint.

Table 8: **Posterior draws.** The montage-generalisation model is evaluated on 1k held-out 256-channel scenarios per simulator, pooled across SEREEGA and TVB. Expected and best-of-$S$ errors average distances from sampled-support centroids to the true centre of mass; point-collapse centroid and peak errors are medians, and precision is a mean. Values are mean $\pm$ standard deviation over three seeds; best-of-$S$ is an oracle.

| draws $S$ | expected (flat by construction) | point-collapse centre of mass | point-collapse peak | best-of-$S$ (oracle) | precision (detection) |
|---|---|---|---|---|---|
| 1 | $18.5_{\pm 0.0}$ | $23.4_{\pm 0.1}$ | $57.8_{\pm 0.8}$ | $18.5_{\pm 0.0}$ | $0.43_{\pm 0.01}$ |
| 2 | $18.6_{\pm 0.3}$ | $17.7_{\pm 0.4}$ | $27.8_{\pm 1.2}$ | $13.7_{\pm 0.3}$ | $0.36_{\pm 0.01}$ |
| 5 | $18.7_{\pm 0.2}$ | $12.8_{\pm 0.2}$ | $17.3_{\pm 0.7}$ | $9.9_{\pm 0.2}$ | $0.53_{\pm 0.00}$ |
| 10 | $18.7_{\pm 0.2}$ | $12.1_{\pm 0.1}$ | $15.9_{\pm 0.4}$ | $7.8_{\pm 0.1}$ | $0.53_{\pm 0.00}$ |
| 20 | $18.6_{\pm 0.2}$ | $11.8_{\pm 0.1}$ | $14.5_{\pm 0.1}$ | $6.4_{\pm 0.1}$ | $0.56_{\pm 0.01}$ |
| 30 | $18.6_{\pm 0.2}$ | $11.7_{\pm 0.2}$ | $14.2_{\pm 0.3}$ | $5.6_{\pm 0.0}$ | $0.57_{\pm 0.01}$ |
| 50 | $18.7_{\pm 0.2}$ | $11.7_{\pm 0.1}$ | $14.3_{\pm 0.5}$ | $4.9_{\pm 0.0}$ | $0.57_{\pm 0.01}$ |
| 100 | $18.6_{\pm 0.2}$ | $11.6_{\pm 0.0}$ | $13.9_{\pm 0.2}$ | $4.1_{\pm 0.0}$ | $0.55_{\pm 0.01}$ |
| 200 | $18.6_{\pm 0.2}$ | $11.6_{\pm 0.2}$ | $14.0_{\pm 0.2}$ | $3.4_{\pm 0.0}$ | $0.56_{\pm 0.01}$ |

The independent ablation keeps the frozen imager, decoder trunk and pointer scoring, but predicts parcels in one step; the parameter-count difference is 8,705 out of 2.18M.

**Baselines.** MNE (H"am"al"ainen & Ilmoniemi, 1994), sLORETA (Pascual-Marqui et al., 1994) and dSPM (Dale et al., 2000) use the same parcel leadfield, double-precision inversion, no depth weighting and Tikhonov regularisation $0.05\,\mathrm{tr}(GG^\top)/C$. sLORETA normalises by the square root of the resolution-matrix diagonal and dSPM by the inverse-operator row norm; each method scores parcels by reconstructed temporal RMS. DeepSIF (Sun et al., 2022) and ConvDip (Hecker et al., 2021) follow their reference architectures, with three recurrent layers for DeepSIF (Reynaud et al., 2024) and a $9 \times 9$ scalp image for ConvDip. Both require a fixed sensor count and are trained separately by montage. To evaluate support ranking and apply our decoder to DeepSIF, we add a parcelwise detection head trained with the same focal objective as our imager (Section 3.2).

**Environment and provenance.** Runs use one NVIDIA A100 (40 GB) GPU and record software versions, split rules and group counts in `provenance.json`; the evaluated runs use Python 3.11 and PyTorch 2.13. Evaluations retain per-observation records and fitted operating points, from which the table generator computes reported values and the paper audit checks them.

# D Ablations

This appendix supports the main experiments with four checks: sensitivity to the number of posterior draws, error from leadfield reconstruction, the contribution of the physics prior, and the choice of training simulators. Metric definitions and implementation details are in Appendix C.

## D.1 Posterior draw count

Increasing the number of draws $S$ improves the Monte Carlo approximation to a fixed learned support distribution, not the distribution itself. Table 8 separates this sampling effect from predictive accuracy: expected draw error should remain stable, whereas point estimates formed from empirical marginals can change with $S$. Best-of-$S$ uses the ground truth to choose a draw and is therefore an oracle diagnostic rather than a deployable estimate.

The point-collapse peak is especially sensitive to very few draws, whereas the centre of mass stabilises sooner. At $S = 50$, the centre-of-mass and peak errors are within 0.08 and 0.34 mm, respectively, of their $S = 200$ values when averaged over seeds. This supports $S = 50$ for the main synthetic comparisons. Marginal-based diagnostics remain draw-count sensitive: ECE decreases from 0.00085 at 50 draws to 0.00056 at 200, while union recall rises from 0.68 at five draws to 0.94 at 200. We therefore keep $S$ fixed when comparing sampled methods.

Table 9: **Predicted versus true leadfield.** The same frozen multi-montage imager is evaluated with each leadfield on the same synthetic test sets. LE is the median Euclidean error in mm and AUPRC the mean across observations; both simulators are pooled before averaging over three seeds.

| | | LE (mm) ↓ | | AUPRC ↑ | |
|---|---|---|---|---|---|
| **Montage** | **Seen by predictor** | predicted | true | predicted | true |
| 21 | no | 16.5 | 16.5 | 0.521 | 0.521 |
| 32 | yes | 11.7 | 11.7 | 0.678 | 0.678 |
| 64 | yes | 10.5 | 10.5 | 0.738 | 0.739 |
| 76 | no | 10.4 | 10.4 | 0.747 | 0.748 |
| 90 | yes | 10.3 | 10.3 | 0.755 | 0.756 |
| 128 | yes | 10.1 | 10.1 | 0.769 | 0.770 |
| 256 | no | 13.0 | 9.9 | 0.598 | 0.775 |

Table 10: **The physics prior.** Paired deterministic imagers differ only in the prior; training regimes and distance/AUPRC definitions follow Table 3. The with-prior arm has three seeds, whereas the no-prior arm has two under joint training and one at fixed 256; cells show observed seed variation only. Bold marks the better point estimate in each pair.

| | LE (mm) ↓ | | AUPRC ↑ | |
|---|---|---|---|---|
| **Evaluation set** | with | without | with | without |
| *Jointly trained on 32, 64, 90, and 128 channels* | | | | |
| Seen layouts (pooled) | $10.6_{\pm 0.1}$ | $\mathbf{10.0_{\pm 0.1}}$ | $0.735_{\pm 0.005}$ | $\mathbf{0.761_{\pm 0.001}}$ |
| 76 channels (unseen) | $10.4_{\pm 0.1}$ | $\mathbf{9.8_{\pm 0.2}}$ | $0.747_{\pm 0.005}$ | $\mathbf{0.770_{\pm 0.002}}$ |
| 21 channels (unseen) | $\mathbf{16.5_{\pm 0.1}}$ | $18.8_{\pm 0.1}$ | $\mathbf{0.521_{\pm 0.005}}$ | $0.501_{\pm 0.004}$ |
| 256 channels (unseen) | $13.0_{\pm 0.2}$ | $\mathbf{9.4_{\pm 0.1}}$ | $0.598_{\pm 0.006}$ | $\mathbf{0.791_{\pm 0.002}}$ |
| Localize-MI | $\mathbf{22.4_{\pm 1.0}}$ | $28.9_{\pm 0.1}$ | $\mathbf{0.102_{\pm 0.004}}$ | $0.042_{\pm 0.011}$ |
| *Separately trained on 256 channels* | | | | |
| 256 channels (seen) | $9.7_{\pm 0.0}$ | **8.9** | $0.787_{\pm 0.005}$ | **0.818** |
| Localize-MI | $22.9_{\pm 0.9}$ | **22.3** | $\mathbf{0.093_{\pm 0.010}}$ | 0.080 |

### D.2 Leadfield reconstruction

Our imager constructs its physics prior from a leadfield predicted from electrode positions rather than requiring a subject-specific forward model. To isolate the cost of this approximation, we evaluate the same frozen imager with the predicted and true leadfields. The leadfield predictor was trained on 32-, 64-, 90-, and 128-channel layouts; 21, 76, and 256 channels were held out.

Table 9 shows negligible differences on six layouts, including the held-out 21- and 76-channel arrangements of familiar positions. At 256 channels, where the electrode positions are new, the true leadfield substantially improves both metrics and brings the frozen imager close to the montage-matched reference in Table 3. Leadfield prediction is thus a major source of the synthetic transfer gap at this layout.

### D.3 The physics prior

The imager's physics prior comprises sLORETA estimates and per-parcel leadfield statistics (Appendix C.8). We remove these features while holding the architecture, training data, recipe, and seed fixed, so Table 10 tests whether physical guidance adds value beyond learned geometry.

On synthetic data, removing the prior improves LE and AUPRC in most settings, including layouts where leadfield reconstruction is accurate. This indicates that the imager can recover useful spatial structure from simulated training data without explicit inverse features. At unseen 256 channels, the larger gain is also consistent with leadfield-prediction error (Appendix D.2); the sparse 21-channel layout is the exception, where the prior helps both metrics.

On Localize-MI, the multi-montage model instead benefits from the prior: it improves both peak–site distance and AUPRC, and ranks the stimulated parcel higher in 81 of 122 runs after averaging the two completed paired seeds. It also concentrates the ranking: the seed-averaged median size of the 90% marginal-mass set is 565 parcels with the prior and 814 without it. For the fixed 256-channel model, the ranking gain is smaller and currently based on one no-prior seed. The contrast suggests that physical guidance helps transfer to real EEG and sparse sensing, even when it limits performance on simulated test sets.

### D.4 Training simulator diversity

The main tables use models trained on both simulators. To test that choice, we train our imager and DeepSIF on TVB, SEREEGA, or both, then evaluate each on the two synthetic test sets separately and on Localize-MI. Separate synthetic columns show the direction of transfer between simulators.

Table 11: **Training simulators and transfer.** Deterministic imagers at 256 channels, evaluated on separate 10k-scenario test sets per simulator and on 61 Localize-MI runs. LE, SD, and peak–site are Euclidean medians in mm; AUPRC is a mean. Values are mean ± standard deviation over three seeds.

| | | Tested on TVB | | | Tested on SEREEGA | | | Localize-MI | |
|---|---|---|---|---|---|---|---|---|---|
| **Model** | **Trained on** | LE ↓ | AUPRC ↑ | SD ↓ | LE ↓ | AUPRC ↑ | SD ↓ | Peak–site ↓ | AUPRC ↑ |
| $\text{Ours}^{\text{Det}}$ | TVB | $8.6_{\pm 0.1}$ | $0.830_{\pm 0.002}$ | $3.4_{\pm 0.1}$ | $44.3_{\pm 0.4}$ | $0.171_{\pm 0.006}$ | $19.8_{\pm 0.2}$ | $38.1_{\pm 3.7}$ | $0.046_{\pm 0.009}$ |
| | SEREEGA | $16.4_{\pm 0.5}$ | $0.608_{\pm 0.016}$ | $6.2_{\pm 0.1}$ | $14.5_{\pm 0.2}$ | $0.674_{\pm 0.005}$ | $4.5_{\pm 0.1}$ | $22.7_{\pm 0.3}$ | $0.075_{\pm 0.015}$ |
| | both | $7.6_{\pm 0.2}$ | $0.848_{\pm 0.004}$ | $3.5_{\pm 0.0}$ | $13.4_{\pm 0.1}$ | $0.727_{\pm 0.006}$ | $4.0_{\pm 0.1}$ | $22.9_{\pm 0.9}$ | $0.093_{\pm 0.010}$ |
| DeepSIF | TVB | $10.6_{\pm 0.2}$ | $0.717_{\pm 0.006}$ | $6.2_{\pm 0.1}$ | $34.8_{\pm 1.6}$ | $0.402_{\pm 0.025}$ | $19.3_{\pm 0.5}$ | $68.4_{\pm 4.8}$ | $0.025_{\pm 0.010}$ |
| | SEREEGA | $13.3_{\pm 0.0}$ | $0.616_{\pm 0.001}$ | $6.0_{\pm 0.1}$ | $11.8_{\pm 0.0}$ | $0.769_{\pm 0.001}$ | $4.9_{\pm 0.0}$ | $28.9_{\pm 2.2}$ | $0.067_{\pm 0.002}$ |
| | both | $10.8_{\pm 0.2}$ | $0.711_{\pm 0.001}$ | $5.4_{\pm 0.0}$ | $11.7_{\pm 0.1}$ | $0.787_{\pm 0.000}$ | $4.8_{\pm 0.0}$ | $57.6_{\pm 5.0}$ | $0.055_{\pm 0.014}$ |

Table 11 shows asymmetric transfer: training only on TVB performs poorly on SEREEGA, whereas the reverse shift is milder. Our imager trained on both improves LE and AUPRC on each synthetic test set even over the specialist trained on that simulator. DeepSIF gains on SEREEGA under pooled training, although its TVB specialist retains a small LE and AUPRC advantage on that simulator.

Localize-MI provides a transfer test generated by neither simulator. Pooled training gives our imager the highest AUPRC (0.0926 versus 0.0752 and 0.0464), while preserving the peak–site accuracy of SEREEGA-only training (22.9 versus 22.7 mm). DeepSIF transfers best when trained on SEREEGA alone, but both our pooled and SEREEGA-only imagers place the peak closer and rank the stimulated parcel higher. These real-data results support pooling the simulators for our model, rather than choosing one solely from its own synthetic test set.

## E Spatio-temporal compression of the source field

The model estimates a parcel coefficient field $\boldsymbol{S}_D \in \mathbb{R}^{P \times D}$ with $P = 994$ and $D = 64$ rather than the source field $\boldsymbol{S} \in \mathbb{R}^{V \times T}$ with $V = 20484$ and $T = 500$, reducing the unknowns from $10^7$ to $6.4 \times 10^4$. Both axes are reduced, but only the temporal one because the data are compressible along it, and this appendix reports the measurements that establish the difference.

We estimated the covariance of the source field separately along each axis on 6k held-out scenarios (generated using SEREEGA) after parcel reduction, taking the waveform of every active parcel for the temporal spectrum and the parcel map at a set of time points for the spatial one. The temporal covariance is close to low rank, with twenty-two of $500$ directions accounting for $95\%$ of the variance and three accounting for half, as the smoothness of the waveforms would suggest, whereas the spatial covariance needs $685$ of $994$ directions for the same fraction. Because a sparse signal has a flat spectrum even without spatial organisation, we added a control drawing supports independently at random with the mean cardinality observed in the data, $12.6$ active parcels. That control requires $920$ directions, so the parcel maps carry more structure than sparsity alone would produce, though not enough to admit a low-dimensional spatial basis.

Table 12: **Principal directions required to reach a given fraction of the variance**, on 6,000 held-out scenarios. Temporal directions are counted out of $T = 500$ and spatial directions out of $P = 994$.

| Axis | Dimension | 50% | 90% | 95% | 99% |
|---|---|---|---|---|---|
| Temporal (waveforms) | 500 | 3 | 16 | 22 | 37 |
| Spatial (parcel maps) | 994 | 121 | 522 | 685 | 900 |
| Spatial, sparse null | 994 | 422 | 855 | 920 | 976 |

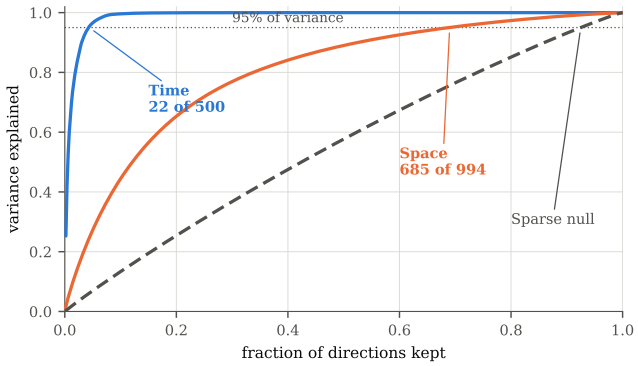



Figure 5: **Cumulative variance along each axis**, plotted against the fraction of directions retained since the two axes have different dimensions. The temporal curve saturates within a few percent of its directions, whereas the spatial curve remains close to the diagonal an incompressible signal would follow.

Since explained variance is not the quantity on which source estimates are judged, we also measured what spatial truncation would cost on the quantities the main text reports. Projecting the parcel maps onto their first $k$ principal directions and retaining 64 of them, the budget adopted along time, leaves 81% of the field unexplained and moves the median peak by 24.4 mm; even at 512 directions the peak still moves 9.6 mm, the entire localisation error of Ours$^{\mathrm{Det}}$ at 256 channels (Table 13). No spatial budget comparable to the temporal one preserves the estimate, which is why the spatial reduction is not a truncation of a learned basis. The operator $\mathbf{\Pi}$ is the fixed 994-parcel atlas on which the simulator defines its patches, so it removes structure the labels never contained, and these numbers describe what the variance-based alternative would have cost instead.

Table 13: **Effect of spatial truncation.** Medians over the parcel maps of the same 6,000 scenarios.

| Directions $k$ | Relative error | Peak displacement (mm) | Peak preserved |
|---|---|---|---|
| 16 | 0.941 | 52.2 | 4.6% |
| 64 | 0.811 | 24.4 | 14.3% |
| 256 | 0.507 | 13.9 | 33.5% |
| 512 | 0.311 | 9.6 | 44.0% |

It remained to choose how many temporal components to keep and whether the basis should be fixed or estimated, so we compared an orthonormal DCT-II basis against a truncated SVD basis fitted to 50k training waveforms and evaluated both on 6k held-out ones. The fitted basis is more accurate at every smaller budget and the two meet near $D \approx 61$, with the DCT reaching a median relative error of 0.031 at $D = 64$ against 0.035 for the fitted basis (Table 14).

Table 14: **Rate–distortion of the temporal basis.** Median relative $\ell_2$ error and median Pearson correlation against the component budget $D$; compression is relative to 500 float32 samples.

| | | DCT | | SVD (fitted) | |
|---|---|---|---|---|---|
| $D$ | **Compression** | **Rel. error** | $r$ | **Rel. error** | $r$ |
| 32 | 15.6× | 0.189 | 0.982 | **0.138** | 0.990 |
| 48 | 10.4× | 0.062 | 0.998 | **0.048** | 0.999 |
| 64 | 7.8× | **0.031** | **0.9995** | 0.035 | 0.9994 |

We adopt the fixed basis for reasons beyond the small difference in accuracy. An SVD basis estimated from SEREEGA and TVB waveforms encodes the temporal structure those engines happen to produce, and since part of our evaluation is on real EEG, an analytic basis has no fitted parameters that could fail to transfer. Orthonormality of $\mathbf{\Phi}$ also makes the mean-squared error between coefficients equal to that between waveforms, so compressing the target leaves the training objective unchanged, while the same projection can be applied to the sensors (recall $\boldsymbol{Y}_D$ from Section 2). The budget suits the forward operator as well, since a rank-64 truncation of the parcel gain retains $99.975\%$ of its energy on `standard_1005_128`.

Two caveats bound these conclusions. The spectra were measured on 1-second windows at 500 Hz, so $D = 64$ is calibrated to that regime and sharper transients would require more components. The spatial analysis characterises the parcel maps rather than the vertex field, and therefore supports only the narrower statement that no further spatial reduction is available once the parcellation is fixed, the parcellation itself following from the resolution argument of Section 3.

## F Discussion and limitations

**Transfer from simulation to real EEG.** The Localize-MI results provide evidence that the learned inverse mapping transfers beyond the simulators used for training. Under the shared two-simulator training protocol, our full model achieves 22.3 mm median peak-to-site distance and 0.090 stimulated-parcel AUPRC, compared with 57.6 mm and 0.055 for DeepSIF (Table 2). The corresponding 61.3% distance reduction is specific to this comparison. DeepSIF trained only on SEREEGA reaches 28.9 mm, showing that the size of the transfer gap depends on the training distribution. Our deterministic imager still achieves a lower distance under both pooled and SEREEGA-only training. The evidence therefore supports improved transfer under the evaluated protocols, while also showing that stronger synthetic performance does not necessarily imply stronger real-data performance.

**The role of physical guidance.** The physics-prior ablation reveals a trade-off between synthetic reconstruction and real-data transfer. Removing the prior improves localization and parcel ranking in most synthetic settings, yet worsens the multi-montage model's Localize-MI distance from 22.4 to 28.9 mm and AUPRC from 0.102 to 0.042. This suggests that physical guidance can help under distribution shift even when it constrains performance on simulated data. The effect is not uniform: for the separately trained 256-channel model, the prior improves real-data AUPRC but does not improve peak-to-site distance.

**Scope of montage generalization.** Montage-agnostic processing means that the same learned parameters accept different electrode counts and coordinates; it does not imply equal accuracy across layouts. Performance is maintained on the unseen 76-channel arrangement, while sparse coverage at 21 channels and new electrode positions at 256 channels produce larger errors (Table 3). Replacing the predicted leadfield with the true leadfield reduces the unseen 256-channel error from 13.0 to 9.9 mm, identifying leadfield approximation as a major contributor to that transfer gap. The anatomy, cortical parcellation, and coordinate system remain fixed. Consequently, removing the need for a supplied leadfield at inference does not remove dependence on the forward models used during training or establish generalization across anatomical variability.

**Interpretation of support uncertainty.** The decoder represents uncertainty in which parcels are active together and in support cardinality. It does not sample waveform shape or amplitude: ev-

ery draw selects rows from the same predicted coefficient field. Its lower energy score and excess AURC on held-out synthetic data support the predictive value of these sampled supports, but do not establish uniformly better reconstruction or isolate the contribution of dependencies alone (Table 4). The independent decoder also learns different marginal probabilities; isolating dependence effects would require a comparison with matched marginals. Excess AURC additionally depends on the underlying localization-error distribution, while pooled ECE is dominated by inactive parcels. Neither low ECE nor concentrated marginal-mass sets establishes calibrated coverage. Finally, training uses one energy-defined ordering per target support, and sampling is capped at 32 parcels, covering 98% of the synthetic test supports.

**Real-data and baseline evaluation boundaries.** Localize-MI is the only real-EEG benchmark evaluated here and comprises 61 stimulation-locked averaged runs. The known stimulation site supports localization and stimulated-parcel ranking, but does not identify the complete active support or validate reconstructed waveforms and support probabilities. The results therefore do not establish performance on arbitrary spontaneous EEG or single-trial recordings. SESAME is evaluated only on synthetic data and uses a dipolar source representation that produces substantially smaller supports than the parcel-level targets. Its scores consequently reflect both inference quality and source-model mismatch, and should not be interpreted as establishing superiority over Bayesian ESI in general. Broader validation across anatomies, recording conditions, and source representations remains necessary.