SPD-METAFORMER IS WHAT YOU NEED FOR SMALL-DATA BRAIN DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Brain signal decoding is challenging because neural recordings are noisy and vary across individuals, while labeled data are often limited. Recent attention-based models on the symmetric positive definite (SPD) manifold have nevertheless achieved strong performance using covariance and connectivity representations, yet the contribution of learned token weighting remains unclear. We examine two representative architectures, MAtt (based on log-Euclidean geometry) and GBWAtt (based on generalized Bures–Wasserstein geometry), and find that their learned attention weights remain close to uniform after training. We relate this behavior to bounded similarity parameterizations that, under the original softmax scaling, limit attention-weight contrast. Moreover, replacing learned weights with uniform weights, throughout training and evaluation, has little effect on mean predictive performance while preserving each model's original aggregation geometry. Motivated by these findings, we introduce SPD-MetaFormer, an attention-free architecture built on uniformly weighted Fréchet aggregation under log-Euclidean geometry. Its backbone uses a geodesic residual to update a summary token and a shared spectral feedforward map to transform all tokens, followed by a learned weighted readout. Token states remain SPD-valued until tangent-space classification. Across three EEG benchmarks, SPD-MetaFormer achieves competitive results relative to published Euclidean and manifold baselines. Separate matched reproductions test learned versus uniform weighting within MAtt and GBWAtt. These results suggest that, in the short-sequence and limited-data regimes studied, carefully designed SPD architectures can provide a simpler and effective alternative to adaptive manifold attention.

1 INTRODUCTION

Understanding the human brain is a long-standing challenge in science and medicine. Brain decoding connects neural recordings to cognitive and clinical states, supporting brain–computer interfaces, neurorehabilitation, and the study of neurological and psychiatric conditions such as ADHD, Parkinson's disease, and depression (GBD 2021 Nervous System Disorders Collaborators, 2024; GBD 2019 Mental Disorders Collaborators, 2022; Wolpaw et al., 2002; Chaudhary et al., 2016; Woo et al., 2017). The learning problem is demanding because recordings are high-dimensional, noisy, and heterogeneous, while labeled data are often limited and costly to collect (Lotte et al., 2018; Varoquaux, 2018). Reliable decoding must therefore recover stable structure without overfitting to subject- and session-specific variation.

Existing approaches include convolutional decoders for electroencephalography (EEG) signals (Lawhern et al., 2018; Mane et al., 2021) and methods that explicitly model second-order structure. In EEG, covariance matrices capture power and coordinated channel variation relevant to motor imagery (MI), steady-state visual evoked potentials (SSVEPs), and error-related negativity (ERN), while functional magnetic resonance imaging (fMRI) connectivity describes covariation among regional time series (Barachant et al., 2012; Congedo et al., 2017; Dai et al., 2020; You & Park, 2021). Although these quantities have different neuroscientific meanings, appropriately regularized estimates from both are symmetric positive definite (SPD) matrices. Once a metric is specified, SPD geometry supplies meaningful distances and averages without learning these operations from scarce labels, supporting classical Riemannian classifiers and deep SPD networks (Pennec et al., 2006; Arsigny et al., 2006; Huang & Van Gool, 2017; Brooks et al., 2019).

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.

Building on these representations, recent methods use Transformer-style attention to model interactions among manifold-valued tokens. Manifold attention assigns these tokens input-dependent weights before geometry-aware aggregation. MAtt (Pan et al., 2022) and GBWAtt (Jin et al., 2026) are representative examples, using Log–Euclidean and generalized Bures–Wasserstein geometries, respectively. Other methods explore deep SPD features, correlation matrices, and gyrovector spaces (Wang et al., 2023; Hu et al., 2025; Wang et al., 2025). Related Transformers likewise preserve SPD structure throughout their layers (Seraphim et al., 2024). Although these models achieve competitive performance, it remains unclear how much the learned attention weights contribute to the improvement.

We investigate this question in MAtt and GBWAtt by comparing learned attention weights with uniform weights while keeping the rest of each architecture unchanged. We find that both models learn nearly uniform attention weights. Replacing these weights with uniform weights during both training and evaluation has little effect on mean predictive performance, while preserving each model's original aggregation geometry (see Section 4.4; Appendix A.5 and Appendix B). Our analysis shows that bounded scores limit the contrast between attention weights at unit temperature, while temperature controls reveal no consistent benefit from sharper attention. Together, these findings suggest that attention in these two architectures acts primarily as geometry-aware averaging in the studied regime. This motivates using uniform geometric aggregation as the token mixer and designing effective transformations around it.

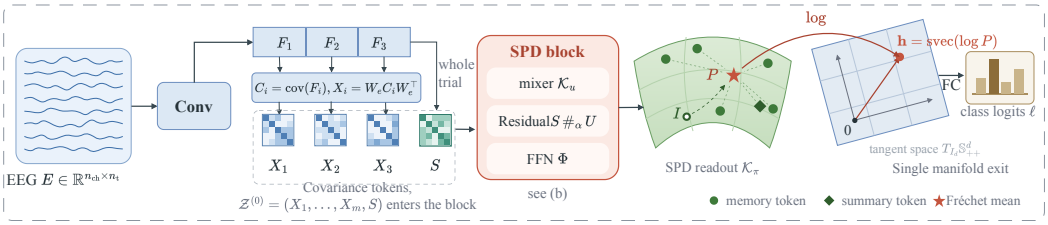
MetaFormer, developed for computer vision tasks (Yu et al., 2022), and related work on token mixing (Tay et al., 2021) mirror these observations by demonstrating that simplified token mixers suffice for effective representations. Adapting these ideas to small-data brain decoding requires operations that preserve SPD structure while combining information across covariance groups and keeping local evidence available for classification. We therefore introduce SPD-MetaFormer, which retains local covariance tokens and refines a separate summary through uniform geometric aggregation of transformed tokens. A shared spectral transformation improves matrix conditioning, and a geometric readout combines the local tokens and summary using learned weights shared across inputs. The resulting representation enters a tangent space only for final classification.

In summary, we make the following contributions:

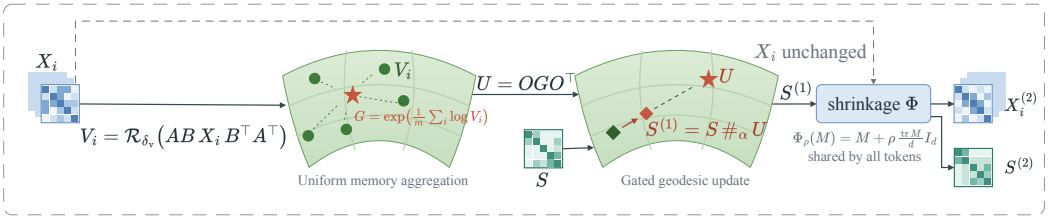
- **Controlled diagnosis of manifold attention.** Across three EEG benchmarks, MAtt and GBWAtt learn nearly uniform attention weights, and matched uniform-weight interventions change mean scores by at most 0.36 points. Score analysis bounds attention-weight contrast at unit temperature, while temperature controls reveal no consistent benefit from sharper attention in the studied settings (Section 4.4; Appendix B).
- **SPD-MetaFormer.** We introduce an attention-free SPD backbone that separates local covariance memories from a global summary. Uniform Log–Euclidean aggregation of transformed memories drives a gated geodesic update of the summary, while keeping local tokens available for readout. A shared, single-parameter spectral transformation preserves positive definiteness and cannot increase the spectral condition number of any token. A geometric readout then combines local and global information using learned weights shared across inputs (Sections 3.2–3.5).
- **Empirical evaluation.** SPD-MetaFormer achieves mean accuracies of 75.44% and 70.79% on MI and SSVEP and a mean AUC of 82.53% on ERN, exceeding the highest baseline means in the published comparison compiled by Jin et al. (2026). On ABIDE fMRI, it achieves 79.14% AUC, the highest mean among the compared baselines under shared subject splits.

2 Preliminaries

Let $\mathbb{S}^d$ and $\mathbb{S}^d_{++}$ denote the real symmetric and symmetric positive-definite (SPD) $d \times d$ matrices, respectively. For an eigendecomposition $A = U \operatorname{diag}(\lambda_1, \ldots, \lambda_d) U^\top$, matrix functions act on the eigenvalues: $f(A) = U \operatorname{diag}(f(\lambda_1), \ldots, f(\lambda_d)) U^\top$. In particular, $\log$ maps SPD matrices to $\mathbb{S}^d$, and $\exp$ maps symmetric matrices back to $\mathbb{S}^d_{++}$. A complementary operation is the congruence map $X \mapsto W X W^\top$, which preserves positive definiteness for full-row-rank $W \in \mathbb{R}^{r \times d}$ (Huang & Van Gool, 2017).



(a) SPD-MetaFormer



(b) SPD block

Figure 1: Overview of SPD-MetaFormer. **(a)** Convolutional features yield local covariance memories (blue) and an observed summary (green), embedded in a common SPD space. After the SPD block, a geometric readout with learned weights shared across trials (dotted spokes) pools all tokens for tangent-space classification. **(b)** The attention-free block updates only the summary through uniform geometric memory aggregation and a gated geodesic residual, then applies shared spectral shrinkage to every token. Circles, diamonds, and stars denote memories, the summary, and Fréchet means, respectively.

For $X, Y \in \mathbb{S}^d_{++}$, the Log-Euclidean (LE) metric measures distance in the matrix-logarithm coordinates (Arsigny et al., 2006):

$$d_{\mathrm{LE}}(X, Y) = \|\log X - \log Y\|_{\mathrm{F}}\,. \tag{1}$$

For $X_{1:m} = (X_1, \ldots, X_m) \in (\mathbb{S}^d_{++})^m$ and weights $\boldsymbol{w} = (w_1, \ldots, w_m)$ with $w_j \geq 0$ and $\sum_j w_j = 1$, the weighted Fréchet mean uniquely minimizes $\sum_{j=1}^m w_j d^2_{\mathrm{LE}}(P, X_j)$ over $P \in \mathbb{S}^d_{++}$ and has the closed form (Arsigny et al., 2006; 2007)

$$\mathrm{FM}^{\mathrm{LE}}_{\boldsymbol{w}}(X_{1:m}) = \exp\left(\sum_{j=1}^m w_j \log X_j\right). \tag{2}$$

Similarly, the LE geodesic interpolates between $X, Y \in \mathbb{S}^d_{++}$ while remaining SPD:

$$X \#_t Y = \exp((1-t)\log X + t \log Y)\,, \qquad t \in [0, 1]. \tag{3}$$

Section 3 uses congruence maps for covariance embeddings and mixer projections, spectral maps for token transformations, and log for the final tangent representation. The mean $\mathrm{FM}^{\mathrm{LE}}_{\boldsymbol{w}}$ supplies uniform mixing and weighted readout, while $\#_t$ defines the gated summary residual.

3 Proposed Method

3.1 Architecture Overview and Design Rationale

We introduce SPD-MetaFormer, a covariance-based decoder that uses uniform geometric mixing directly on SPD matrices. The central design problem is to integrate information across covariance groups and preserve the evidence contributed by each group. SPD-MetaFormer addresses this problem with local covariance memory tokens that retain group-specific second-order statistics and a data-dependent summary that accumulates global information. This role separation keeps the local sources accessible throughout summary refinement and final readout.

Following the MetaFormer perspective (Yu et al., 2022), Figure 1 separates geometric token mixing from shared spectral shrinkage, followed by geometric readout and tangent-space classification.

3.2 Covariance Tokens Separate Local Evidence from a Global State

The left side of Figure 1(a) constructs the role-aware input from the same trial at local and global resolutions.

Local covariance memory. The memory path represents an EEG trial through a collection of local second-order statistics. Given $E \in \mathbb{R}^{n_{\mathrm{ch}} \times n_{\mathrm{t}}}$ with n_{ch} channels and n_{t} time samples, spatial and temporal convolutions extract a feature map $F = f_{\mathrm{conv}}(E)$. Following the spatial–temporal decomposition used in EEG convolutional models and manifold attention (Wei et al., 2019; Pan et al., 2022; Jin et al., 2026), we divide F into m feature groups $F_i \in \mathbb{R}^{c_{\mathrm{f}} \times L_i}$ for $i = 1, \ldots, m$, each with c_{f} feature channels and L_i time samples, and convert each group into a regularized covariance,

$$C_i = \frac{1}{L_i - 1} \bar{F}_i \bar{F}_i^{\top} + \varepsilon I_{c_{\mathrm{f}}}, \qquad \bar{F}_i = F_i - \frac{1}{L_i} F_i \mathbf{1}_{L_i} \mathbf{1}_{L_i}^{\top}. \tag{4}$$

Here $I_{c_{\mathrm{f}}}$ is the $c_{\mathrm{f}} \times c_{\mathrm{f}}$ identity matrix and $\mathbf{1}_{L_i}$ is the L_i-dimensional all-ones vector. We set the diagonal floor to $\varepsilon = 10^{-5}$ to ensure positive definiteness and stabilize subsequent matrix operations. We preserve the original covariance scale, so each C_i records both the second-order structure and the energy of its feature group.

Observed global state. The summary construction follows the feature-grouping scheme. For temporal segments, C_{s} is computed from the complete feature sequence. For channel groups that each span the full trial, we aggregate their covariances as $C_{\mathrm{s}} = m^{-1} \sum_{i=1}^{m} C_i$. The resulting C_{s} initializes the trial-specific state updated by the memory mixer. The construction used for each task is specified with the corresponding data preprocessing.

Common SPD space. Geometric interaction requires placing local and global states in a common space. We therefore apply a shared congruence map (Huang & Van Gool, 2017),

$$X_i = W_{\mathrm{e}} C_i W_{\mathrm{e}}^{\top}, \qquad S = W_{\mathrm{e}} C_{\mathrm{s}} W_{\mathrm{e}}^{\top}, \qquad W_{\mathrm{e}} \in \mathbb{R}^{d \times c_{\mathrm{f}}}, \quad W_{\mathrm{e}} W_{\mathrm{e}}^{\top} = I_d. \tag{5}$$

Row orthonormality keeps the map full rank and hence preserves positive definiteness. When $c_{\mathrm{f}} = d$, we omit this map and use the identity. The result is the role-aware sequence $\mathcal{Z}^{(0)} = (X_1, \ldots, X_m, S)$ shown at the entrance to the SPD block in Figure 1. Across datasets, the model retains this token structure and adapts the number of memory tokens and their SPD dimension to the input feature structure. Appendix A specifies the exact configurations, feature shapes, and numerical eigenvalue floors. With both roles expressed in the same geometry, they form a valid input to the geometric mixer.

3.3 Uniform Geometric Mixing Refines the Summary

Uniform geometric mixing is motivated by the near-uniform allocations learned by MAtt and GBWAtt. This behavior suggests that their effective operation is geometric averaging rather than selective routing. We therefore retain learned SPD transformations and Fréchet aggregation while replacing input-dependent weights with uniform mixing. Figure 1(b) illustrates this design. Its empirical basis is examined in Section 4.4, with further analyses in Appendix A.5 and Appendix B.

Uniform memory aggregation. Let $B \in \mathbb{R}^{d \times d}$ be a learned orthogonal base map, and let $A, O \in \mathbb{R}^{d \times d}$ be learned value and output maps, respectively. Using eigenvalue rectification $\mathcal{R}_{\delta}$ as in SPDNet (Huang & Van Gool, 2017), the mixer computes

$$V_i = \mathcal{R}_{\delta_v}\left(A B X_i B^{\top} A^{\top}\right). \tag{6}$$

The transformed memory tokens have the uniform LE mean

$$G = \mathrm{FM}^{\mathrm{LE}}_{\frac{1}{m}\mathbf{1}_m}(\{V_i\}_{i=1}^{m}) = \exp\left(\frac{1}{m} \sum_{i=1}^{m} \log V_i\right). \tag{7}$$

The output map gives $U = OGO^{\top}$. We set $\delta_v = 10^{-4}$ to floor the eigenvalues of V_i, preserving positive definiteness and stabilizing the matrix logarithms. The intermediate G is the uniform LE Fréchet mean from Equation 2 (Arsigny et al., 2006; 2007), and U is the transformed memory mean passed to the gated summary update.

Gated summary update. The transformed memory mean U supplies cross-group information to the summary. We interpolate S and U along an LE geodesic,

$$S^{(1)} = S \#_\alpha U = \exp((1-\alpha)\log S + \alpha \log U), \qquad \alpha = \mathrm{sigmoid}(a), \tag{8}$$

where a is learned. The gate controls the update strength, and LE interpolation preserves positive definiteness. In parallel, the memory tokens follow the identity path $X_i^{(1)} = X_i$, yielding $\mathcal{Z}^{(1)} = (X_1^{(1)}, \ldots, X_m^{(1)}, S^{(1)})$. This separation parallels the distinct feature and aggregation roles in class-attention architectures (Touvron et al., 2021). The shared spectral shrinkage is then applied to both paths.

3.4 Shared Spectral Shrinkage Conditions Both Token Paths

The mixer deliberately gives the memory and summary different update histories. Before readout, the feedforward part of the SPD block applies the same spectral shrinkage to both, improving their conditioning and preserving their roles. For $M \in \mathbb{S}_{++}^d$, define

$$\Phi_\rho(M) = M + \rho \frac{\mathrm{tr}(M)}{d} I_d, \qquad \rho = 0.1 \exp(\theta), \tag{9}$$

where the scalar θ is learned, initialized at zero, and shared across tokens. Applying this map to the two paths gives

$$X_i^{(2)} = \Phi_\rho(X_i^{(1)}), \qquad S^{(2)} = \Phi_\rho(S^{(1)}). \tag{10}$$

The shared parameter ρ yields a token-adaptive correction whose magnitude scales with each token's mean eigenvalue.

The action of Φ_ρ on the eigenvalues is explicit. If M has eigenvalues λ_k and mean eigenvalue $\bar{\lambda} = d^{-1}\sum_k \lambda_k$, then $\Phi_\rho(M)$ retains the same eigenvectors and has eigenvalues $\lambda_k + \rho\bar{\lambda}$. Positive definiteness is therefore preserved, and the spectral condition number cannot increase. Up to the retained scale factor $1+\rho$, this operation is a convex combination of M and its isotropic component $\mathrm{tr}(M)I_d/d$, connecting it to covariance shrinkage (Ledoit & Wolf, 2004). The decoding objective directly optimizes ρ, which distinguishes it from a covariance-estimation coefficient. This token-wise operation complements the preceding cross-token mixer.

3.5 Trial-Independent Geometric Readout Retains Local and Global Evidence

After summary refinement and spectral shrinkage, the block returns $\mathcal{Z}^{(2)} = (X_1^{(2)}, \ldots, X_m^{(2)}, S^{(2)})$. The readout in Figure 1(a) fuses every token on the SPD manifold before the architecture makes its single transition to a Euclidean decision space.

Trial-independent geometric readout. The readout assigns one learned score b_i to each memory position and one score β to the summary. These scores are shared across trials and give the normalized role weights

$$\boldsymbol{\pi} = \frac{\big(\exp(b_1), \ldots, \exp(b_m), \exp(\beta)\big)}{\sum_{j=1}^m \exp(b_j) + \exp(\beta)}. \tag{11}$$

The readout $\mathcal{K}_\pi$ then forms their weighted LE Fréchet mean,

$$P = \exp\left(\sum_{i=1}^m \pi_i \log X_i^{(2)} + \pi_{m+1} \log S^{(2)}\right). \tag{12}$$

The mixer uses uniform weights for the gated summary update. The readout learns role weights for the final fusion of the memories and summary.

Single manifold exit. After geometric readout, the single SPD representation P is mapped to the tangent space at the identity and classified with one affine layer,

$$\mathbf{h} = \mathrm{svec}(\log P), \qquad \widehat{\mathbf{p}}(\cdot \mid E) = \mathrm{softmax}(W_h \mathbf{h} + b_h). \tag{13}$$

Here $\mathbf{h} \in \mathbb{R}^{d(d+1)/2}$, and $\widehat{p}(y \mid E) = \widehat{\mathbf{p}}_y$. The operator svec stacks the upper triangle and multiplies off-diagonal entries by $\sqrt{2}$, preserving the Frobenius inner product on symmetric matrices. This tangent head follows established SPD networks (Huang & Van Gool, 2017; Pan et al., 2022) and is applied once, after the manifold backbone has completed summary refinement and sequence readout. All modules are trained jointly with cross-entropy, with class weighting for the imbalanced ERN task.

Table 1: Comparison with published EEG results. MI and SSVEP report accuracy (%); ERN reports AUC (%). Baseline entries are reproduced from Table 3 of Jin et al. (2026).

Model	MI	SSVEP	ERN
Published results: mean ± SD			
EEGNet	61.84 ± 6.39	53.72 ± 7.23	74.28 ± 2.47
ShallowCNet	57.43 ± 6.25	56.93 ± 6.97	71.86 ± 2.64
SCCNet	71.95 ± 5.05	62.11 ± 7.70	70.93 ± 2.31
FBCNet	56.52 ± 3.07	53.09 ± 5.67	60.47 ± 3.06
TCNet-Fusion	71.45 ± 4.45	45.00 ± 6.45	70.46 ± 2.94
MBEEGSE	64.58 ± 6.07	56.45 ± 7.27	75.46 ± 2.34
SPDNet	72.93 ± 4.33	62.30 ± 3.12	72.05 ± 4.43
SPDNetBN	73.02 ± 3.67	62.76 ± 3.01	72.34 ± 3.46
MAtt	74.71 ± 5.01	65.19 ± 3.14	75.68 ± 2.23
GDLNet	69.32 ± 2.89	65.52 ± 2.86	78.23 ± 2.52
GBWAtt ($\theta = 1$)	74.54 ± 2.04	67.12 ± 2.94	80.42 ± 2.12
GBWAtt ($\theta = 1.5$)	74.95 ± 1.83	67.27 ± 2.62	80.90 ± 1.97
Our experiments: mean ± SD			
SPD-MetaFormer	$\mathbf{75.44 \pm 1.69}$	$\mathbf{70.79 \pm 2.17}$	$\mathbf{82.53 \pm 1.85}$

4 Experiments

We evaluate SPD-MetaFormer on three public EEG benchmarks spanning motor imagery (MI), steady-state visual evoked potentials (SSVEP), and error-related negativity (ERN). An additional study on the ABIDE fMRI cohort examines whether the same SPD processing architecture transfers beyond EEG, with the complete protocol and results reported in Appendix C. The evaluation first establishes its predictive performance against existing convolutional and geometry-aware EEG models. Controlled comparisons then examine the contribution of the complete SPD processing path and its static geometric readout. Finally, we revisit MAtt and GBWAtt through locked-uniform interventions and allocation-entropy measurements to determine whether their content-dependent weights contribute to predictive performance.

Datasets and protocol. We use BCIC-IV-2a for four-class MI decoding with nine subjects (Brunner et al., 2008), MAMEM-SSVEP-II for five-class SSVEP decoding with eleven subjects (Nikolopoulos, 2016), and BCI-ERN for binary ERN detection with sixteen subjects (Perrin et al., 2012), following prior manifold-attention studies (Pan et al., 2022; Jin et al., 2026). Each model is trained independently for each subject using seeds 0–9. For MI, we form a class-stratified validation subset from the training session and use the separate evaluation session for testing. For SSVEP and ERN, the first four sessions support training and validation (one held out), and the final session is used for testing, following MAtt (Pan et al., 2022). All paired ablations share the data splits and select the checkpoint with the lowest validation loss. Full training settings appear in Appendix A.3.

Metrics and uncertainty. We report accuracy for MI/SSVEP and ROC-AUC for imbalanced ERN, on a 0–100 scale. ERN AUC follows the evaluation convention of MAtt and GBWAtt (Pan et al., 2022; Jin et al., 2026). Our scores average seeds within subjects, then subjects equally; $\pm$ values and error bars report the mean within-subject sample SD across runs, also used for entropy dispersion. Paired bootstrap intervals resample subjects after seed averaging and are interpreted descriptively without multiplicity correction (Appendix A.3). Published baselines retain their source uncertainty definitions.

4.1 Comparison with Existing Methods

We compare SPD-MetaFormer with convolutional EEG models, including EEGNet (Lawhern et al., 2018), ShallowCNet (Schirrmeister et al., 2017), SCCNet (Wei et al., 2019), FBCNet (Mane et al., 2021), TCNet-Fusion (Musallam et al., 2021), and MBEEGSE (Altuwaijri et al., 2022). The geometry-aware baselines are SPDNet (Huang & Van Gool, 2017), SPDNetBN (Brooks et al., 2019), MAtt

Table 2: Whole-path controls under the same subject/seed grid and training recipe. Values are mean ± mean within-subject run SD (sample SD over ten runs per subject, averaged over subjects). MI/SSVEP: accuracy (%); ERN: AUC (%).

Configuration	MI	SSVEP	ERN
Convolutional features only	26.25 ± 2.03	24.35 ± 3.15	72.29 ± 3.71
SPD tangent readout	70.27 ± 2.26	59.35 ± 3.10	80.87 ± 2.49
Euclidean MetaFormer	59.32 ± 3.94	67.69 ± 3.06	77.65 ± 3.29
Memory-only geometric readout	74.41 ± 1.90	67.93 ± 3.15	80.83 ± 2.50
Full SPD-MetaFormer	$\mathbf{75.44 \pm 1.69}$	$\mathbf{70.79 \pm 2.17}$	$\mathbf{82.53 \pm 1.85}$

(Pan et al., 2022), GDLNet (Wang et al., 2024), and GBWAtt (Jin et al., 2026). Their results are taken from the published comparison of Jin et al. (2026) rather than from new reproductions.

SPD-MetaFormer achieves 75.44% accuracy on MI, 70.79% on SSVEP, and 82.53% AUC on ERN. These results exceed the highest published mean in each column by 0.49, 3.52, and 1.63 points, respectively. The largest improvement occurs on SSVEP, whereas the MI result remains close to the strongest reported baseline. The consistent performance across the three paradigms shows that an attention-free architecture can remain competitive when exact uniform LE mixing is combined with the shared spectral transformation and static geometric readout.

To examine whether the proposed SPD processing path transfers beyond EEG, we further evaluate SPD-MetaFormer on the multi-site ABIDE fMRI cohort. It achieves an AUC of 79.14% and exceeds the strongest compared baseline under matched subject splits. This result provides complementary evidence that the same geometry-aware architecture transfers to covariance representations from a distinct neuroimaging modality. The complete protocol and results are reported in Appendix C.

4.2 Whole-Path Comparisons

The benchmark results do not determine whether performance arises from the convolutional frontend, covariance tokenization alone, or the complete SPD block and static geometric readout. We therefore construct four controls that share the frontend initialization, subject–seed grid, and optimization recipe of SPD-MetaFormer. *Convolutional features only* flattens the feature map $F = f_{\mathrm{conv}}(E)$ and applies a linear classifier, removing covariance tokenization and all subsequent SPD operations. *SPD tangent readout* retains the covariance tokens, data-dependent summary, and shared congruence map W_{e}, but replaces the SPD block and static geometric readout with a linear classifier applied to the concatenated log-vectorizations of $\mathcal{Z}^{(0)} = (X_1, \ldots, X_m, S)$. *Euclidean MetaFormer* preserves the same covariance tokenization and shared congruence map, then processes the log-vectorized tokens using a pre-LayerNorm uniform mixer, additive residuals, and a two-layer GELU MLP with an expansion ratio of four, followed by mean readout. *Memory-only geometric readout* removes the summary and the SPD block, applying the static geometric readout (equation 12) to $X_1, \ldots, X_m$ alone.

SPD-MetaFormer outperforms all four controls on every dataset (Table 2). Relative to direct SPD tangent readout, it improves MI and SSVEP accuracy by 5.17 and 11.44 points and ERN AUC by 1.66 points. The Euclidean MetaFormer also performs worse despite having more parameters (Appendix A.4). Retaining the geometric readout but removing the summary and SPD block still lowers MI, SSVEP, and ERN by 1.03, 2.86, and 1.70 points. The paired subject-bootstrap intervals exclude zero in all twelve comparisons. The control hierarchy argues against four simpler explanations for the observed performance. The convolutional frontend alone is insufficient, constructing SPD tokens followed by direct tangent classification is insufficient, geometric readout of the local memories alone is insufficient, and generic uniform mixing in a Euclidean backbone does not recover the full model. These results support the complete SPD processing path rather than uniform mixing alone. Since the controls jointly alter transformations, capacity, and readout, the next experiment more narrowly isolates the static geometric readout.

Table 3: Readout ablations with the complete SPD block retained. All three controls change the final representation used for classification. Values are mean ± mean within-subject run SD (sample SD over ten runs per subject, averaged over subjects). MI/SSVEP: accuracy (%); ERN: AUC (%).

Configuration	MI	SSVEP	ERN
Concatenation readout	70.31 ± 2.45	61.73 ± 3.31	81.83 ± 1.93
Arithmetic pooling	75.31 ± 1.86	66.65 ± 2.35	81.90 ± 2.00
Summary-only readout	74.06 ± 1.85	67.03 ± 2.94	82.06 ± 2.04
Full SPD-MetaFormer	$\mathbf{75.44 \pm 1.69}$	$\mathbf{70.79 \pm 2.17}$	$\mathbf{82.53 \pm 1.85}$

Table 4: Training with learned versus locked-uniform weights in existing MAtt and GBWAtt reproductions. Values are mean ± mean within-subject run SD (sample SD over ten runs per subject, averaged over subjects); Δ is uniform minus learned, with a paired subject-bootstrap 95% interval. All scores are multiplied by 100. GBWAtt comparisons match $\theta = 1.5$ on both sides. These experiments use their own matched reproduction protocols (Appendix A.5).

Model / dataset	Learned	Uniform	Δ [95% interval]
MAtt / MI	74.84 ± 1.69	74.83 ± 1.73	$-0.01\ [-0.21, +0.17]$
MAtt / SSVEP	65.14 ± 3.33	64.91 ± 3.04	$-0.23\ [-1.22, +0.87]$
MAtt / ERN	74.87 ± 4.41	74.91 ± 4.60	$+0.04\ [-1.15, +1.23]$
GBWAtt ($\theta = 1.5$) / MI	75.21 ± 1.57	74.93 ± 1.94	$-0.28\ [-0.99, +0.31]$
GBWAtt ($\theta = 1.5$) / SSVEP	64.43 ± 3.42	64.06 ± 2.87	$-0.36\ [-1.06, +0.35]$
GBWAtt ($\theta = 1.5$) / ERN	80.97 ± 2.35	80.90 ± 2.47	$-0.07\ [-0.42, +0.26]$

4.3 Internal Readout Ablations

We retain the complete SPD block and modify only its static geometric readout. *Concatenation readout* stacks $\mathrm{svec}(\log X_i^{(2)})$ for all memory tokens together with $\mathrm{svec}(\log S^{(2)})$ and learns a new linear classifier. *Arithmetic pooling* replaces the weighted LE Fréchet mean in equation 12 with $P_{\mathrm{arith}} = \sum_{i=1}^{m} \pi_i X_i^{(2)} + \pi_{m+1} S^{(2)}$ while retaining the input-independent role weights $\boldsymbol{\pi}$. *Summary-only readout* discards the memory tokens and applies the original affine classifier to $\mathrm{svec}(\log S^{(2)})$.

The ablations reveal complementary roles for token retention and pooling geometry. Concatenation retains every token but removes the structured geometric pooling operation, producing the largest losses on MI and SSVEP at 5.13 and 9.06 points, respectively. Summary-only readout lowers the mean on all three tasks, indicating that the memory tokens retain information that complements the refined summary at classification. The effect of pooling geometry is more task-dependent. Arithmetic pooling reduces SSVEP accuracy by 4.15 points but changes MI accuracy by only 0.13 points. Overall, the static geometric readout provides the most consistent performance.

4.4 Locked-Uniform Attention and Allocation Entropy

We revisit ten-seed experiments using MAtt and the released implementation of GBWAtt. Within each architecture, the locked-uniform control fixes every attention row to $\mathbf{u} = m^{-1}\mathbf{1}_m$ during both training and evaluation while preserving the frontend, value transformation, aggregation geometry, and classifier. MAtt retains its LE aggregation, whereas GBWAtt retains the weighted square-root aggregation in its released implementation. We denote the GBWAtt parameter by θ_{GBW} and set it to 1.5 for both variants in each pair. The evaluation contains 720 matched subject–seed pairs across the two architectures and three datasets, and Appendix A.5 details the reproduction protocols.

Locking the attention weights changes mean performance by at most 0.36 points across all six architecture–dataset comparisons (Table 4), and every paired interval includes zero. Under these configurations, the residual content-dependent variation in the learned weights therefore provides no measurable improvement over uniform allocation. The intervals are not equivalence tests, so the result establishes practical insensitivity at the measured precision rather than an exact zero effect.

Table 5: Attention entropy at the validation-selected checkpoint, averaged over test queries/trials, then seeds and subjects. $\overline{\mathrm{SD}}(H)$ averages the within-subject sample SD over ten runs. Natural logarithms give nats. Uniform weights have $H = \log m$ and $H/\log m = 1$ exactly. GBWAtt uses $\theta = 1.5$.

Model / dataset	m	H_{learned}	$\overline{\mathrm{SD}}(H)$	H_{uniform}	$H_{\text{learned}}/\log m$
MAtt / MI	3	1.098605	8.12×10^{-7}	1.098612	0.999994
MAtt / SSVEP	7	1.945906	8.97×10^{-7}	1.945910	0.999998
MAtt / ERN	3	1.098578	1.98×10^{-5}	1.098612	0.999969
GBWAtt / MI	3	1.098410	3.44×10^{-5}	1.098612	0.999815
GBWAtt / SSVEP	7	1.945243	1.33×10^{-4}	1.945910	0.999657
GBWAtt / ERN	3	1.097353	6.41×10^{-4}	1.098612	0.998854

We next determine whether this performance insensitivity coincides with nearly uniform learned allocations. For a row-stochastic attention matrix $\Gamma(Y) \in \mathbb{R}^{m\times m}$, we measure its entropy on the test set at the selected checkpoint.

$$H(\Gamma) = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{m}\Gamma_{ij}\log\Gamma_{ij}, \qquad \overline{H}(\Gamma) = \frac{H(\Gamma)}{\log m}. \tag{14}$$

Uniform allocation gives $H(\Gamma) = \log m$ and $\overline{H}(\Gamma) = 1$. Learned MAtt attains a mean normalized entropy above 0.99996 on every dataset, while GBWAtt remains above 0.99885 (Table 5). Both architectures therefore allocate nearly equal mass across their short token sequences. The entropy measurement characterizes the allocations learned by the original models, while the locked-uniform intervention shows that removing their residual input dependence has little effect on mean predictive performance.

The inverse-log score used by both implementations further constrains the attainable allocation contrast (Pan et al., 2022; Jin et al., 2026). For a nonnegative distance d_{ij}, the score $s_{ij} = 1/(1+\log(1+d_{ij}))$ lies in $(0, 1]$. At unit softmax temperature, any two weights in the same row satisfy $\Gamma_{ij}/\Gamma_{ik} = \exp(s_{ij} - s_{ik}) \le e$. The score parameterization therefore bounds the pairwise allocation ratio independently of the input. This structural constraint alone does not imply the near-maximal empirical entropies. Temperature controls further show that sharper attention does not consistently improve performance over uniform mixing (Appendix B). Together, the intervention, entropy measurements, and score analysis show that MAtt and GBWAtt operate in a nearly uniform regime in which content-dependent weighting has little effect on predictive performance. Combined with the whole-path and readout controls, this diagnosis supports the design emphasis of SPD-MetaFormer on uniform geometric mixing, shared spectral transformation, and static geometric readout.

5 Conclusion

In this work, we revisited the role of adaptive manifold attention in small-data brain decoding. Matched experiments on MAtt and GBWAtt show that their learned allocations remain close to uniform and that replacing them with locked-uniform weights has little effect on predictive performance in the settings studied. Motivated by this finding, we introduced SPD-MetaFormer, an attention-free architecture built around uniform geometric mixing, geodesic summary refinement, shared spectral transformation, and geometric readout. Across three EEG benchmarks, SPD-MetaFormer achieves competitive performance relative to existing Euclidean and manifold-based methods, while additional results on ABIDE support its applicability to fMRI connectivity. Overall, our results suggest that, for short sequences and limited data, the effectiveness of SPD models can arise from geometry-aware representation, transformation, and aggregation without requiring strongly selective input-dependent weighting. An important direction for future work is to understand why optimization nevertheless drives existing manifold-attention models toward nearly uniform allocations. Although bounded similarity scores constrain attainable attention contrast, they do not fully explain the near-maximal entropies observed in practice. A theoretical characterization of how score parameterization, optimization dynamics, data scale, and SPD geometry interact may help determine when adaptive manifold attention is necessary and when uniform geometric aggregation is sufficient.

AI USE STATEMENT

Generative AI tools were used to assist with brainstorming and refining research ideas, generating portions of the experimental code, and improving the clarity and readability of the manuscript. All AI-assisted ideas, code, and text were reviewed, verified, and, where necessary, modified by the authors. The authors take full responsibility for the final methodology, implementation, results, and content of the paper.

REPRODUCIBILITY STATEMENT

Appendix A describes the feature extraction, covariance construction, parameterization, data splits, and training protocols. All datasets used in this work are publicly available. We will release our code upon acceptance.

REFERENCES

Ghadir Ali Altuwaijri, Ghulam Muhammad, Hamdi Altaheri, and Mansour Alsulaiman. A multi-branch convolutional neural network with squeeze-and-excitation attention blocks for EEG-based motor imagery signals classification. *Diagnostics*, 12(4):995, 2022. doi: 10.3390/diagnostics12040995. URL https://doi.org/10.3390/diagnostics12040995.

Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Log-Euclidean metrics for fast and simple calculus on diffusion tensors. *Magnetic Resonance in Medicine*, 56(2):411–421, 2006. doi: 10.1002/mrm.20965. URL https://doi.org/10.1002/mrm.20965.

Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications*, 29(1):328–347, 2007. doi: 10.1137/050637996. URL https://doi.org/10.1137/050637996.

Alexandre Barachant, Stéphane Bonnet, Marco Congedo, and Christian Jutten. Multiclass brain–computer interface classification by Riemannian geometry. *IEEE Transactions on Biomedical Engineering*, 59(4):920–928, 2012. doi: 10.1109/TBME.2011.2172210. URL https://doi.org/10.1109/TBME.2011.2172210.

Daniel Brooks, Olivier Schwander, Frédéric Barbaresco, Jean-Yves Schneider, and Matthieu Cord. Riemannian batch normalization for SPD neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/6e69ebbfad976d4637bb4b39de261bf7-Abstract.html.

Clemens Brunner, Robert Leeb, Gernot Müller-Putz, Alois Schlögl, and Gert Pfurtscheller. BCI Competition 2008—Graz data set A. Technical report, Graz University of Technology, 2008. URL https://www.bbci.de/competition/iv/desc_2a.pdf.

Ujwal Chaudhary, Niels Birbaumer, and Ander Ramos-Murguialday. Brain–computer interfaces for communication and rehabilitation. *Nature Reviews Neurology*, 12(9):513–525, 2016. doi: 10.1038/nrneurol.2016.113. URL https://doi.org/10.1038/nrneurol.2016.113.

Marco Congedo, Alexandre Barachant, and Rajendra Bhatia. Riemannian geometry for EEG-based brain–computer interfaces; a primer and a review. *Brain-Computer Interfaces*, 4(3):155–174, 2017. doi: 10.1080/2326263X.2017.1297192. URL https://doi.org/10.1080/2326263X.2017.1297192.

Cameron Craddock, Yassine Benhajali, Carlton Chu, Francois Chouinard, Alan Evans, András Jakab, Budhachandra Singh Khundrakpam, John David Lewis, Qingyang Li, Michael Milham, Chaogan Yan, and Pierre Bellec. The Neuro Bureau preprocessing initiative: Open sharing of preprocessed neuroimaging data and derivatives. *Frontiers in Neuroinformatics*, 7, 2013. doi: 10.3389/conf.fninf.2013.09.00041. URL https://doi.org/10.3389/conf.fninf.2013.09.00041. Conference Abstract: Neuroinformatics 2013.

R. Cameron Craddock, G. Andrew James, Paul E. Holtzheimer, Xiaoping P. Hu, and Helen S. Mayberg. A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Human Brain Mapping*, 33(8):1914–1928, 2012. doi: 10.1002/hbm.21333. URL https://doi.org/10.1002/hbm.21333.

Mengyu Dai, Zhengwu Zhang, and Anuj Srivastava. Analyzing dynamical brain functional connectivity as trajectories on space of covariance matrices. *IEEE Transactions on Medical Imaging*, 39(3):611–620, 2020. doi: 10.1109/TMI.2019.2931708. URL https://doi.org/10.1109/TMI.2019.2931708.

Adriana Di Martino, Chao-Gan Yan, Qingyang Li, Erin Denio, Francisco X. Castellanos, Kaat Alaerts, Jeffrey S. Anderson, Michal Assaf, Susan Y. Bookheimer, Mirella Dapretto, Ben Deen, Sonja Delmonte, Ilan Dinstein, Birgit Ertl-Wagner, Damien A. Fair, Louise Gallagher, Daniel P. Kennedy, Christopher L. Keown, Christian Keysers, Janet E. Lainhart, Catherine Lord, Beatriz Luna, Vinod Menon, Nancy J. Minshew, Christopher S. Monk, Sophia Mueller, Ralph-Axel Müller, Mary Beth Nebel, Joel T. Nigg, Kirsten O'Hearn, Kevin A. Pelphrey, Scott J. Peltier,

Jeffrey D. Rudie, Stefan Sunaert, Marc Thioux, J. Michael Tyszka, Lucina Q. Uddin, Judith S. Verhoeven, Nicole Wenderoth, Jillian L. Wiggins, Stewart H. Mostofsky, and Michael P. Milham. The Autism Brain Imaging Data Exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry*, 19(6):659–667, 2014. doi: 10.1038/mp.2013.78. URL https://doi.org/10.1038/mp.2013.78.

GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019. *The Lancet Psychiatry*, 9(2):137–150, 2022. doi: 10.1016/S2215-0366(21)00395-3. URL https://doi.org/10.1016/S2215-0366(21)00395-3.

GBD 2021 Nervous System Disorders Collaborators. Global, regional, and national burden of disorders affecting the nervous system, 1990–2021: A systematic analysis for the Global Burden of Disease Study 2021. *The Lancet Neurology*, 23(4):344–381, 2024. doi: 10.1016/S1474-4422(24)00038-3. URL https://doi.org/10.1016/S1474-4422(24)00038-3.

Chen Hu, Rui Wang, Xiaoning Song, Tao Zhou, Xiao-Jun Wu, Nicu Sebe, and Ziheng Chen. A correlation manifold self-attention network for EEG decoding. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 5372–5380, 2025. doi: 10.24963/ijcai.2025/598. URL https://doi.org/10.24963/ijcai.2025/598.

Zhiwu Huang and Luc Van Gool. A Riemannian network for SPD matrix learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017. doi: 10.1609/aaai.v31i1.10866. URL https://doi.org/10.1609/aaai.v31i1.10866.

Shaocheng Jin, Tao Zhou, Rui Wang, Ziheng Chen, Xiaoqing Luo, Xiao-Jun Wu, and Josef Kittler. Towards robust EEG decoding based on Riemannian self-attention. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 11225–11235, 2026. doi: 10.1145/3770855.3818867. URL https://doi.org/10.1145/3770855.3818867.

Jeremy Kawahara, Colin J. Brown, Steven P. Miller, Brian G. Booth, Vann Chau, Ruth E. Grunau, Jill G. Zwicker, and Ghassan Hamarneh. BrainNetCNN: Convolutional neural networks for brain networks; towards predicting neurodevelopment. *NeuroImage*, 146:1038–1049, 2017. doi: 10.1016/j.neuroimage.2016.09.046. URL https://doi.org/10.1016/j.neuroimage.2016.09.046.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SJU4ayYgl.

Vernon J. Lawhern, Amelia J. Solon, Nicholas R. Waytowich, Stephen M. Gordon, Chou P. Hung, and Brent J. Lance. EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018. doi: 10.1088/1741-2552/aace8c. URL https://doi.org/10.1088/1741-2552/aace8c.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004. doi: 10.1016/S0047-259X(03)00096-4. URL https://doi.org/10.1016/S0047-259X(03)00096-4.

Fabien Lotte, Laurent Bougrain, Andrzej Cichocki, Maureen Clerc, Marco Congedo, Alain Rakotomamonjy, and Florian Yger. A review of classification algorithms for EEG-based brain–computer interfaces: A 10 year update. *Journal of Neural Engineering*, 15(3):031005, 2018. doi: 10.1088/1741-2552/aab2f2. URL https://doi.org/10.1088/1741-2552/aab2f2.

Ravikiran Mane, Effie Chew, Karen Chua, Kai Keng Ang, Neethu Robinson, A. Prasad Vinod, Seong-Whan Lee, and Cuntai Guan. FBCNet: A multi-view convolutional neural network for brain-computer interface. *arXiv preprint arXiv:2104.01233*, 2021. URL https://arxiv.org/abs/2104.01233.

Yazeed K. Musallam, Nasser I. AlFassam, Ghulam Muhammad, Syed Umar Amin, Mansour Alsulaiman, Wadood Abdul, Hamdi Altaheri, Mohamed A. Bencherif, and Mohammed Algabri. Electroencephalography-based motor imagery classification using temporal convolutional network fusion. *Biomedical Signal Processing and Control*, 69:102826, 2021. doi: 10.1016/j.bspc.2021.102826. URL https://doi.org/10.1016/j.bspc.2021.102826.

Spiros Nikolopoulos. MAMEM EEG SSVEP dataset II (256 channels, 11 subjects, 5 frequencies presented simultaneously). Figshare dataset, version 2, 2016. URL https://doi.org/10.6084/m9.figshare.3153409.v2.

Yue-Ting Pan, Jing-Lun Chou, and Chun-Shu Wei. MAtt: A manifold attention network for EEG decoding. In *Advances in Neural Information Processing Systems*, volume 35, pp. 31116–31129, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/c981fd12b1d5703f19bd8289da9fc996-Abstract-Conference.html.

Xavier Pennec, Pierre Fillard, and Nicholas Ayache. A Riemannian framework for tensor computing. *International Journal of Computer Vision*, 66(1):41–66, 2006. doi: 10.1007/s11263-005-3222-z. URL https://doi.org/10.1007/s11263-005-3222-z.

Margaux Perrin, Emmanuel Maby, Sébastien Daligault, Olivier Bertrand, and Jérémie Mattout. Objective and subjective evaluation of online error correction during P300-based spelling. *Advances in Human-Computer Interaction*, 2012:578295, 2012. doi: 10.1155/2012/578295. URL https://doi.org/10.1155/2012/578295.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, 38(11):5391–5420, 2017. doi: 10.1002/hbm.23730. URL https://doi.org/10.1002/hbm.23730.

Mathieu Seraphim, Alexis Lechervy, Florian Yger, Luc Brun, and Olivier Etard. Structure-preserving transformers for sequences of SPD matrices. In *32nd European Signal Processing Conference*, pp. 1451–1455, 2024. doi: 10.23919/EUSIPCO63174.2024.10715089. URL https://doi.org/10.23919/EUSIPCO63174.2024.10715089.

Yi Tay, Dara Bahri, Donald Metzler, Da-Cheng Juan, Zhe Zhao, and Che Zheng. Synthesizer: Rethinking self-attention for transformer models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10183–10192, 2021. URL https://proceedings.mlr.press/v139/tay21a.html.

Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 32–42, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Touvron_Going_Deeper_With_Image_Transformers_ICCV_2021_paper.html.

Gaël Varoquaux. Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage*, 180:68–77, 2018. doi: 10.1016/j.neuroimage.2017.06.061. URL https://doi.org/10.1016/j.neuroimage.2017.06.061.

Rui Wang, Xiao-Jun Wu, Hui Li, and Josef Kittler. Riemannian self-attention mechanism for SPD networks. *arXiv preprint arXiv:2311.16738*, 2023. doi: 10.48550/arXiv.2311.16738. URL https://arxiv.org/abs/2311.16738.

Rui Wang, Chen Hu, Ziheng Chen, Xiao-Jun Wu, and Xiaoning Song. A Grassmannian manifold self-attention network for signal classification. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 5099–5107, 2024. doi: 10.24963/ijcai.2024/564. URL https://www.ijcai.org/proceedings/2024/564.

Rui Wang, Chen Hu, Xiaoning Song, Xiaojun Wu, Nicu Sebe, and Ziheng Chen. Towards a general attention framework on gyrovector spaces for matrix manifolds. In *Advances in Neural Information Processing Systems*, volume 38, pp. 112051–112091, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/a2e3b4132ab2e0b7a21e6e75da7f91a9-Abstract-Conference.html.

Chun-Shu Wei, Toshiaki Koike-Akino, and Ye Wang. Spatial component-wise convolutional network (SCCNet) for motor-imagery EEG classification. In *International IEEE EMBS Conference on Neural Engineering*, pp. 328–331, 2019. doi: 10.1109/NER.2019.8716937. URL https://www.merl.com/publications/TR2018-210.

Jonathan R. Wolpaw, Niels Birbaumer, Dennis J. McFarland, Gert Pfurtscheller, and Theresa M. Vaughan. Brain–computer interfaces for communication and control. *Clinical Neurophysiology*, 113(6):767–791, 2002. doi: 10.1016/S1388-2457(02)00057-3. URL https://doi.org/10.1016/S1388-2457(02)00057-3.

Choong-Wan Woo, Luke J. Chang, Martin A. Lindquist, and Tor D. Wager. Building better biomarkers: Brain models in translational neuroimaging. *Nature Neuroscience*, 20(3):365–377, 2017. doi: 10.1038/nn.4478. URL https://doi.org/10.1038/nn.4478.

Kisung You and Hae-Jeong Park. Re-visiting Riemannian geometry of symmetric positive definite matrices for the analysis of functional connectivity. *NeuroImage*, 225:117464, 2021. doi: 10.1016/j.neuroimage.2020.117464. URL https://doi.org/10.1016/j.neuroimage.2020.117464.

Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. MetaFormer is actually what you need for vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10819–10829, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Yu_MetaFormer_Is_Actually_What_You_Need_for_Vision_CVPR_2022_paper.html.

A Implementation and Evaluation Details

A.1 Feature Extraction and Covariance Construction

Table 6 specifies the frontend used for the reported SPD-MetaFormer results. Every convolution is followed by ordinary batch normalization. There is no additional activation between these two Conv–BN stages. Temporal padding with an even kernel gives one additional feature sample. Covariances are centered, divided by $L_i - 1$, and regularized by $10^{-5}I$. Neither trace normalization nor an additional dimension-dependent multiplier is applied.

Table 6: Dataset-specific SPD-MetaFormer frontend. $C_{\text{in}} \to C_{\text{out}}$ denotes convolutional feature channels, with kernel sizes shown in parentheses.

	MI	SSVEP	ERN
Input dimensions	22×438	8×125	56×160
Spatial convolution	$1 \to 22$ $(22, 1)$	$1 \to 125$ $(8, 1)$	$3 \to 69$ $(56, 1)$
Temporal convolution	$22 \to 20$ $(1, 12)$	$125 \to 15$ $(1, 36)$	$69 \to 48$ $(1, 64)$
Temporal padding	6	18	32
Feature matrix	20×439	15×126	48×161
Memory grouping	time	time	channel banks
Memory group sizes	$147, 146, 146$	7×18	3×16 channels
Memory covariance dimension	20	15	16
Summary	full-trial cov.	full-trial cov.	bank-cov. mean
Entry map	$20 \to 15$	identity	$16 \to 8$
Memory + summary tokens	$3 + 1$	$7 + 1$	$3 + 1$

For ERN, the input is repeated into three streams and both convolutions use three groups, yielding three separately learned banks of sixteen channels. Each bank covariance uses all 161 output samples. The reported ERN model uses the arithmetic summary.

A.2 Parameterization and Numerical Operations

Entry maps have orthonormal rows and the square base map is orthogonal. They are parameterized through the thin QR decomposition of the transpose of an unconstrained matrix, with diagonal-sign correction. The value map is initialized as $A = I + 0.1E/\sqrt{d}$ and the output map as $O = I + 0.02E'$, with independent standard-normal entries in E, E'. These two square matrices are unconstrained during optimization. Initialization near the identity is not a hard full-rank constraint.

The implementation symmetrizes matrix outputs. It applies an eigenvalue floor of 10^{-6} before the entry map and 10^{-4} after the value map. In float32, every matrix logarithm evaluates $\log \max(\lambda, 10^{-6})$ spectrally; this also protects the logarithm of the output congruence in Equation 8. Thus, the exact SPD formulas in the main text have explicit numerical regularization at their logarithms. Matrix exponentials and spectral functions use eigendecompositions; backpropagation uses divided differences with their repeated-eigenvalue limit. All reported runs use float32 with two CPU threads per process.

The mixer changes only the summary before the spectral map, but Equation 9 then changes *all* tokens. For completeness, the conditioning and shrinkage identities are

$$\kappa_2(\Phi_\rho(Z)) = \frac{\lambda_{\max} + \rho\mu}{\lambda_{\min} + \rho\mu} \le \frac{\lambda_{\max}}{\lambda_{\min}}, \qquad \mu = \frac{\operatorname{tr}(Z)}{d}. \tag{15}$$

$$\frac{\Phi_\rho(Z)}{1+\rho} = \frac{1}{1+\rho} Z + \frac{\rho}{1+\rho}\mu I_d. \tag{16}$$

The normalization in the second identity explains the relation to isotropic shrinkage; it is not an extra operation performed by the network. Initial values are $\alpha = 0.2$, $\rho = 0.1$, $b_i = 0$, and $\beta = 2.2$.

A.3 Training and Data Splits

Regularization applies to all trainable parameters, including the scalar gate, spectral coefficient parameter, and readout parameters; these are not excluded from weight decay. The SSVEP recipe

Table 7: Fixed training recipes for SPD-MetaFormer and its paired ablations.

	MI	SSVEP	ERN
Optimizer	AdamW	Adam	AdamW
Learning rate	2×10^{-3}	5×10^{-3}	5×10^{-3}
Weight decay	10^{-4}	10^{-1}	10^{-4}
Batch size	128	64	64
Epochs	350	180	130
Learning-rate schedule	cosine	constant	cosine
Gradient norm limit	5	none	5
Cross-entropy	unweighted	unweighted	class-weighted
Selection	min. val. loss	min. val. loss	min. val. loss

therefore includes Adam's coupled L_2 penalty of 0.1. No hyperparameters are retuned separately for the ablation configurations.

For MI, the preprocessed training and test sessions contain 288 trials each. Signals are cropped to indices $[124, 562)$, giving 438 samples. Sorting training trials by class and reserving the last nine trials within each class gives 252 training and 36 validation trials; the separate evaluation session supplies 288 test trials. For SSVEP and ERN, we follow the session-wise MAtt protocol (Pan et al., 2022). The first four sessions are used for model development. One of them is reserved for validation, and the remaining session is used for testing. ERN subjects are 2, 6, 7, 11, 12, 13, 14, 16, 17, 18, 20, 21, 22, 23, 24, and 26. Labels in the held-out test portion are not used for checkpoint selection.

Writing q_{sr} for the test metric of subject s and run r, the EEG scores report $\bar{q} \pm \overline{\mathrm{SD}}_{\mathrm{run}}$:

$$\bar{q}_s = \frac{1}{R}\sum_{r=1}^{R} q_{sr}, \qquad \bar{q} = \frac{1}{N}\sum_{s=1}^{N} \bar{q}_s. \tag{17}$$

$$\overline{\mathrm{SD}}_{\mathrm{run}} = \frac{1}{N}\sum_{s=1}^{N} \sqrt{\frac{1}{R-1}\sum_{r=1}^{R}(q_{sr} - \bar{q}_s)^2}. \tag{18}$$

Here $R = 10$, except that the temperature study uses $R = 3$. This summarizes training-run variability within subjects and defines both score error bars and the entropy SD in Table 5. For paired comparisons, subjects remain the sampling unit: the same averaging is applied to within-subject differences before subjects are resampled, using 10,000 bootstrap draws. Cross-paper comparisons in Table 1 have no paired test because the requisite baseline subject-level predictions are unavailable in the supplied table.

A.4 Ablation Definitions and Parameter Counts

The Euclidean control uses $D = d(d+1)/2$ log-vector features per token. Its mixer averages all pre-normalized tokens, including the summary, and adds this common vector to every token; a shared pre-normalized $D \to 4D \to D$ GELU MLP with an additive residual then transforms all tokens. The readout averages the resulting token vectors. New classifiers and the Euclidean MLP are initialized inside isolated RNG contexts, preserving the shared initialization and the external RNG stream of the paired full model. Concatenation changes classifier width, so it is a readout design comparison rather than a parameter-matched control.

The main-text comparisons describe selected structural and readout questions; they are exploratory ablations, not a preregistered confirmatory test of all architectural choices.

A.5 Matched Uniform Controls for MAtt and GBWAtt

The attention intervention uses existing ten-seed runs, with 90, 110, and 160 runs per model/condition on MI, SSVEP, and ERN. Each learned model is compared with its own uniform counterpart, with identical subject/seed keys, splits, learning rate, regularization, and epoch budget. The GBWAtt comparisons use $\theta = 1.5$ on both sides; the available $\theta = 1$ learned runs are not paired against a $\theta = 1.5$ uniform control.

Table 8: Trainable parameter counts for the configurations reported in the main ablation tables.

Configuration	MI	SSVEP	ERN
Convolutional features only	41,014	78,375	90,329
SPD tangent readout	8,114	73,725	75,289
Euclidean MetaFormer	122,954	185,805	85,765
Memory-only geometric readout	6,677	69,532	75,076
Concatenation readout	8,791	74,402	75,483
Arithmetic pooling	7,356	70,211	75,272
Summary-only readout	7,351	70,202	75,267
Full SPD-MetaFormer	7,356	70,211	75,272

These reproductions follow their recorded settings rather than the SPD-MetaFormer recipe. MAtt uses its mixed Adam/Stiefel optimizer with learning rates 5×10^{-4}, 5×10^{-3}, and 10^{-2} for MI, SSVEP, and ERN. GBWAtt uses RiemannianAdam with learning rates 2.5×10^{-3}, 10^{-3}, and 5×10^{-4}. Both use weight decay 0.1, batch sizes 128/64/32, and epoch budgets 350/180/130 in the same dataset order. The historical MAtt JSON rows do not store batch size; the listed values are the runner defaults, rather than independently verified per-run metadata.

MI uses the training-session holdout and separate test session described above. For SSVEP, both architectures use the split of the released MAtt and GBWAtt data loaders: sessions 1–3 (trials 1–300) for training, session 4 (trials 301–400) for validation, and session 5 (trials 401–500) for testing. For ERN, both architectures use the session-wise split adopted by MAtt, with one of the first four sessions reserved for validation. Within each architecture, the learned and uniform variants use identical splits. These runs support the *within-model* attention intervention. Their absolute scores are not inserted into the main benchmark as if all training protocols were identical.

The released-structure GBWAtt implementation fixes its geometric matrix parameter to the identity and aggregates transformed values as

$$R_i = \left(\sum_{j=1}^{m} A_{ij} V_j^{1/2} \right)^2 , \tag{19}$$

with the same power transformations in the learned and uniform arms. This is the implemented square-root-coordinate mean, not an assertion that it equals the exact Bures–Wasserstein barycenter for arbitrary noncommuting matrices. MAtt retains its LE mean. Thus, the intervention changes the allocation weights without replacing either model's aggregation rule.

Entropy is measured on the full test set after restoring the minimum-validation-loss checkpoint. It is averaged first over query rows and test trials, then over the ten seeds and subjects. Recorded float32 uniform entropies agree with the analytical $\log m$ to within 2×10^{-6}; Table 5 shows the analytical reference. The reported entropy comparisons concern these checkpoints and short token sequences, not the entire training trajectory or all possible attention architectures.

B TEMPERATURE SCALING IN MANIFOLD ATTENTION

We examine whether adjusting the softmax temperature can recover a predictive benefit from learned attention. Keeping each model's score s_{ij} and aggregation geometry unchanged, we use $A_{ij}^{(\tau)} = \mathrm{softmax}_j(s_{ij}/\tau)$. For both MAtt and GBWAtt, we compare a retrained uniform control, seven fixed temperatures $\tau \in \{1, 0.3, 0.1, 0.03, 0.01, 0.003, 0.001\}$, and a learnable $\log \tau$ initialized at $\tau = 1$ or 0.01. Each condition is trained from scratch on all 9/11/16 MI/SSVEP/ERN subjects with seeds 0–2, yielding 2,160 runs with matched initialization and minibatch order.

Training recipes follow Appendix A.5, with $\theta_{\mathrm{GBW}} = 1.5$. For learnable temperature, the learning rate is 0.01, weight decay is zero, and τ is constrained to $[10^{-4}, 10]$. All splits are disjoint: MI retains the split in Appendix A.3, while SSVEP and ERN use consecutive training/validation/test blocks of 300/100/100 and 180/60/100 trials, respectively. Validation cross-entropy selects the checkpoint and, within each subject/seed, the fixed temperature or learnable-temperature initialization. Results average seeds within subjects before averaging across subjects.

Table 9: Temperature controls with three seeds. Default uses $\tau = 1$; tuned and learnable settings are selected by validation loss. Scores are accuracy (MI/SSVEP) or AUC (ERN), multiplied by 100. Δ is tuned minus uniform, with a paired subject-bootstrap 95% interval.

Model / dataset	Default	Uniform	Tuned	Learnable	Δ [95% interval]
MAtt / MI	74.74	75.04	74.09	73.34	-0.95 $[-2.40, +0.31]$
MAtt / SSVEP	64.39	65.52	66.27	64.58	$+0.76$ $[-0.36, +1.88]$
MAtt / ERN	71.35	72.36	73.65	71.28	$+1.28$ $[-1.78, +4.53]$
GBWAtt / MI	75.37	75.18	74.81	73.33	-0.37 $[-1.45, +0.54]$
GBWAtt / SSVEP	64.24	63.55	64.00	63.97	$+0.45$ $[-0.52, +1.36]$
GBWAtt / ERN	75.94	75.21	75.66	74.52	$+0.45$ $[-1.60, +2.92]$

Lower temperatures substantially sharpen attention (Figure 2): at $\tau = 0.001$, normalized entropy ranges from 0.027 to 0.293 across the six settings, but mean performance drops by 2.61–23.73 points relative to uniform (Figure 3). After validation selection, fixed temperatures change performance by -0.95 to $+1.28$ points, with all paired intervals including zero (Table 9). Learnable temperature likewise provides no consistent improvement. Thus, despite gains over the default temperature in some settings, changing temperature alone does not reliably recover a predictive advantage over uniform geometric mixing in these experiments.

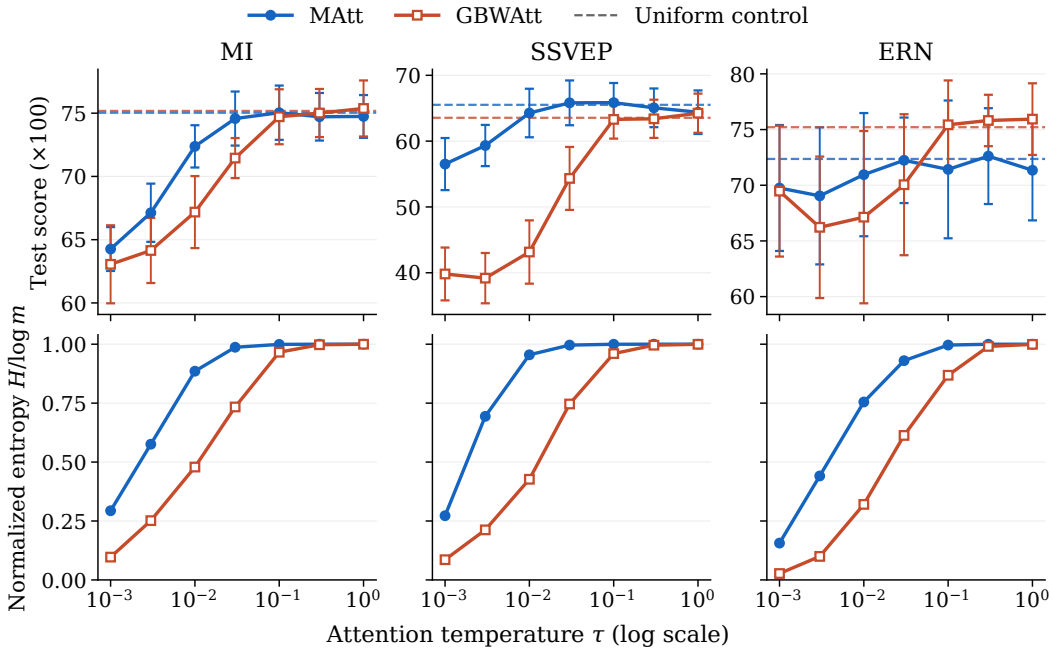


Figure 2: Temperature, predictive performance, and allocation entropy after full retraining. Top: mean test scores with the within-subject sample SD over three runs, averaged over subjects; dashed lines show each model's retrained uniform control. Scores are accuracy (MI/SSVEP) or AUC (ERN), multiplied by 100, with separate vertical scales across datasets. Bottom: normalized attention entropy, where one denotes uniform allocation. All points average three seeds within subjects.

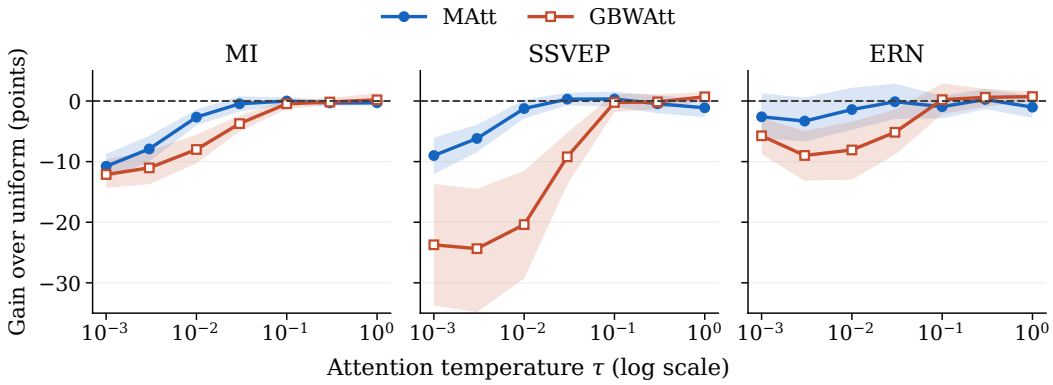


Figure 3: Performance differences from retrained uniform mixing at each fixed temperature. Positive values favor learned weighting. Shading shows pointwise paired subject-bootstrap 95% intervals after averaging seeds within subjects. The dashed line denotes zero difference, and the vertical scale is shared across datasets.

C FUNCTIONAL CONNECTIVITY: ABIDE-I

We evaluate SPD-MetaFormer on ABIDE-I autism classification to assess its applicability to fMRI connectivity beyond EEG, keeping the backbone unchanged.

Data and tokens. The CPAC-preprocessed ABIDE-I release (Di Martino et al., 2014) of the Preprocessed Connectomes Project (Craddock et al., 2013) contains 1,035 subjects (505 with autism, 530 controls) from 20 sites. We use its band-pass-filtered and global-signal-regressed CC200 time series (Craddock et al., 2012). Four evenly spaced windows of $\mathrm{round}(T/2)$ frames, spanning each T-frame scan from its first to last frame, provide four memory tokens; the full scan provides the summary token. Each token is a shrunk Pearson correlation matrix $0.3R + 0.7I_{200}$. The entry map (Equation 5) projects these tokens to $d = 32$, without a convolutional frontend. The resulting model has 10,538 parameters.

Protocol. All methods share ten 80/10/10 subject splits (828/103/104 subjects; seeds 0–9), stratified by site and diagnosis. SPD-MetaFormer uses a fixed AdamW recipe: learning rate 2×10^{-3}, weight decay 10^{-4}, batch size 32, class-weighted cross-entropy, gradient-norm clipping at 5, and cosine decay over at most 20 epochs. The best validation-AUC checkpoint is retained, with early stopping after five unimproved epochs. AUC uses positive-class probabilities; accuracy and balanced accuracy use the threshold in $\{0.20, 0.21, \ldots, 0.80\}$ maximizing validation accuracy. Each test split is evaluated once. We report mean $\pm$ population SD ($\mathrm{ddof} = 0$) across the ten splits.

Baselines. Euclidean baselines use the unshrunk full-scan correlation R: MLP and L_2 logistic regression on its upper triangle, BrainNetCNN (Kawahara et al., 2017), and a two-layer GCN on thresholded R (Kipf & Welling, 2017). SPDNet-BN (Brooks et al., 2019) uses the shrunk full-scan matrix with BiMap widths 100/50/30. MAtt (Pan et al., 2022) uses the same five tokens as SPD-MetaFormer and $d_{\mathrm{out}} = 32$, with divided-difference spectral derivatives and a validation-selected AdamW recipe. All model and threshold selection uses validation data only.

Table 10: ABIDE-I (CC200) test performance: mean $\pm$ population SD ($\times 100$) over ten shared subject splits (seeds 0–9). SPD-MetaFormer and MAtt use $d = 32$. Bold marks the largest mean in each column.

Method	AUC	Accuracy	Balanced acc.
Euclidean, full-scan correlation R			
MLP on $\mathrm{vec}(R)$	78.65 ± 2.67	69.23 ± 2.65	69.16 ± 2.48
L_2 logistic regression on $\mathrm{vec}(R)$	78.48 ± 3.37	71.25 ± 3.11	71.22 ± 3.32
BrainNetCNN (Kawahara et al., 2017)	76.66 ± 4.59	69.04 ± 4.07	69.87 ± 4.89
GCN on thresholded R (Kipf & Welling, 2017)	72.65 ± 3.65	63.85 ± 3.73	65.40 ± 3.04
Riemannian			
SPDNet-BN (Brooks et al., 2019)	78.90 ± 3.05	71.15 ± 3.30	70.85 ± 3.37
MAtt (Pan et al., 2022), $d_{\mathrm{out}} = 32$	77.24 ± 3.61	68.37 ± 4.00	68.18 ± 4.52
SPD-MetaFormer, $d = 32$	$\mathbf{79.14} \pm 3.53$	$\mathbf{71.83} \pm 4.67$	$\mathbf{71.79} \pm 4.66$

Results. SPD-MetaFormer achieves the highest mean AUC (79.14), accuracy (71.83), and balanced accuracy (71.79) among the methods in Table 10. At $d = 32$, its mean AUC exceeds MAtt by 1.90 points while using approximately 57% fewer parameters (10,538 versus 24,482). These results support the applicability of the same compact, attention-free SPD backbone to both EEG covariance and fMRI connectivity.

D SPECTRAL FFN IMPROVES ROBUSTNESS TO TOKEN GRANULARITY

Our spectral FFN Φ_ρ (Section 3.4) adds a trace-proportional isotropic correction to each covariance token using one learned scalar. We test its contribution by shortening the windows used to estimate memory tokens, making covariance estimation more challenging.

Setup. We vary the number of temporal windows K ($3K$ memory tokens for ERN's three channel banks) and train with or without the FFN, replacing it by the identity in the latter case. The summary, frontend, splits, training recipe, and checkpoint selection remain fixed. All 9/11/16 MI/SSVEP/ERN subjects and seeds 0–9 are included: 4,400 runs in total. The main-model settings are $K = 3/7/1$; the pre-specified comparison settings are $K = 18/4/6$, respectively. Statistics follow Appendix A.3, with unadjusted paired t-tests over subjects.

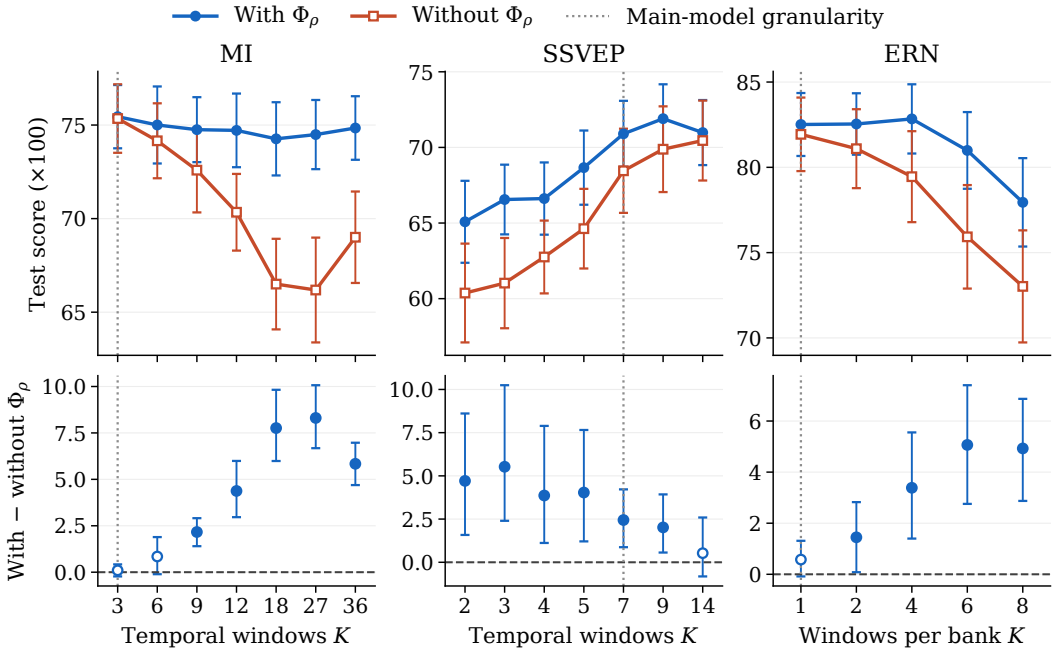


Figure 4: **Contribution of the spectral FFN across token granularities.** Top: accuracy (MI/SSVEP) or AUC (ERN), $\times 100$, with the within-subject sample SD over ten runs, averaged over subjects. Bottom: FFN gain with subject-bootstrap 95% intervals; filled markers indicate intervals excluding zero. Ten seeds are averaged within each subject. Dotted lines mark the main-model settings; these single-threaded reruns differ from Table 1 by at most 0.11 points.

Robustness from one parameter. The FFN substantially reduces the performance loss from finer MI and ERN tokens (Figure 4). At the pre-specified settings, it improves MI accuracy by 7.76 points (95% interval $[5.99, 9.82]$) and ERN AUC by 5.06 points ($[2.76, 7.40]$), benefiting 9/9 and 15/16 subjects, respectively (both $p < 0.002$). Across all seven MI granularities, accuracy stays within 74.3–75.5%: a range of only 1.2 points, versus 9.2 without the FFN. SSVEP also benefits at its main setting ($+2.45$ points, $[0.87, 4.22]$, $p = 0.021$) and at the pre-specified $K = 4$ ($+3.86$, $[1.12, 7.89]$, $p = 0.073$). At the coarser main settings of MI and ERN, the gains are small ($+0.10$ and $+0.58$; both intervals include zero), locating the clearest benefit in the more demanding tokenizations.

Spectral conditioning. At the selected checkpoints, the geometric-mean condition number of memory tokens after the FFN stays below $10^{2.3}$ across all settings, whereas it reaches $10^{5.9}$ without the FFN. The learned correction remains mild ($0.04 \leq \rho \leq 0.07$). Together, these results support the FFN as a compact spectral conditioner that improves robustness to token granularity with a single learned parameter.