UNRAVELING BRAIN SIGNALS THROUGH EXPERT COLLABORATION: MECHANISM-GUIDED ROUTING FOR HETEROGENEOUS EEG LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

EEG decoding across heterogeneous paradigms requires models to capture distinct yet complementary signal characteristics. Motor imagery, error-related potentials, and steady-state visual evoked potentials place different demands on temporal, spectral, and spatial representations, making it challenging for a single representation to capture the cues relevant to each paradigm. We propose MGNet, a mechanism-guided expert-routing framework that enables adaptive collaboration among specialized experts while preserving their complementary representations. MGNet combines a gated causal temporal model with three experts: a temporal module, a spectral dynamic graph module, and a Riemannian covariance module. A lightweight router and a Manhattan-distance-based fusion mechanism adaptively integrate expert representations to accommodate paradigm-dependent EEG characteristics. Experiments on three public benchmarks—BCI Competition IV-2a, BCI Challenge, and MAMEM—show that MGNet consistently outperforms the evaluated state-of-the-art baselines. Ablation studies further reveal paradigm-dependent expert contributions consistent with the signal characteristics motivating their design. These results support mechanism-guided specialization and adaptive expert collaboration as an effective approach to multi-paradigm EEG decoding.

1 INTRODUCTION

Many real-world learning problems are inherently *heterogeneous*: a single model must serve multiple tasks with distinct discriminative information in different representational forms. A common practice is to learn a shared encoder for all tasks; however, when the underlying mechanisms differ, shared representations often entangle task-specific cues and induce parameter interference. Brain–Computer Interface (BCI) decoding from Electroencephalography (EEG) is a representative example of this challenge (Wolpaw, 2013; Gu et al., 2021). Although EEG offers high temporal resolution and non-invasive acquisition, it suffers from low signal-to-noise ratio, non-stationary artifacts, and substantial variability across subjects and paradigms.

These challenges become more pronounced when a unified model must handle heterogeneous BCI paradigms, including motor imagery (MI), error-related potentials (ErrP), and steady-state visual evoked potentials (SSVEP), which rely on fundamentally different discriminative mechanisms: temporal morphology, spatial covariance geometry, and spectral-channel dynamics, respectively (Chavarriaga & Millán, 2010; Barachant et al., 2011; Oikonomou et al., 2016; Nikolopoulos, 2021). Compressing these heterogeneous cues into a shared representation can therefore produce task-mismatched features (Fig. 1). Moreover, online BCI systems require causal inference and robustness to EOG/EMG artifacts, motivating a mechanism-specific, task-adaptive architecture rather than a universal encoder.

Existing EEG decoders often rely on shared encoders or fixed fusion strategies, causing heterogeneous EEG mechanisms to be entangled across paradigms (Lawhern et al., 2018; Schirrmeister et al., 2017; Song et al., 2022; Altaheri et al., 2022; Song et al., 2018; Congedo et al., 2017). We propose **MGNet**, a mechanism-specific expert framework with adaptive routing that preserves temporal,

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.

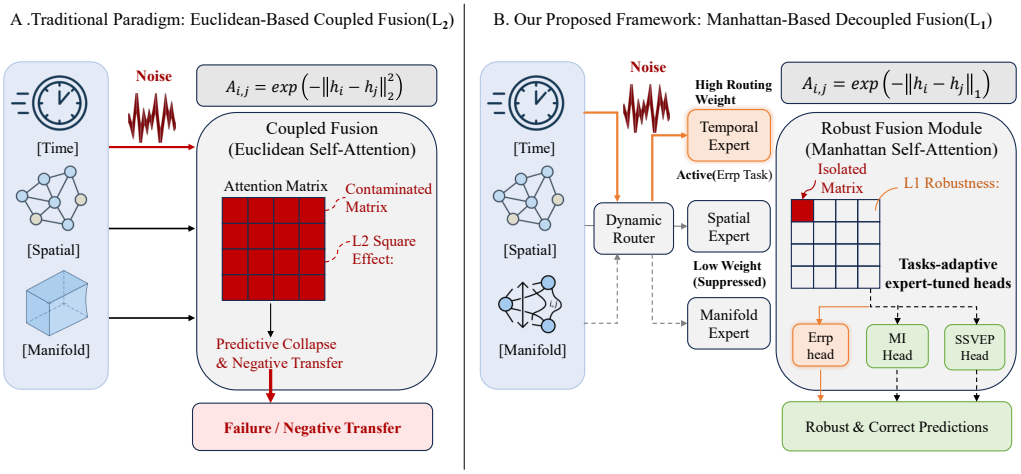


Figure 1: Motivation of MGNet. Shared representations may entangle heterogeneous EEG mechanisms, whereas mechanism-specific experts with adaptive routing preserve task-dependent representations.

spectral-dynamic graph, and Riemannian representations through dedicated experts and coordinates their contributions according to task-dependent EEG characteristics.

The main contributions are summarized as follows:

- **Mechanism-specific expert routing.** We propose a mechanism-specific expert framework with adaptive routing that preserves heterogeneous EEG representations and dynamically coordinates expert contributions, mitigating interference caused by a shared encoder.
- **Distance-aware expert fusion.** We introduce a Manhattan-distance-based fusion module that adaptively models cross-expert relationships and enables robust integration of heterogeneous representations.
- **Cross-paradigm validation.** Experiments on three representative BCI paradigms demonstrate consistent improvements and reveal interpretable task-dependent routing preferences.

2 Related Work

2.1 Deep Learning for EEG Decoding

Deep learning has largely replaced handcrafted pipelines such as FBCSP with end-to-end representation learning for EEG (Altaheri et al., 2023). Convolutional decoders (DeepConvNet, EEGNet) learn spectral-temporal patterns directly from signals (Schirrmeister et al., 2017; Lawhern et al., 2018), and subsequent work adds multi-scale, temporal-convolutional, and attention/Transformer modeling (EEG-Inception, EEG-TCNet, ATCNet, EEG Conformer) (Zhang et al., 2021; Ingolfsson et al., 2020; Altaheri et al., 2022; Song et al., 2022). Most of these models, however, still operate within a single shared feature extractor. Because EEG is highly non-stationary and shifts across subjects and paradigms (Gu et al., 2021), a shared encoder can suffer parameter interference when different tasks rely on different discriminative mechanisms. This motivates mechanism-specific representation learning rather than forcing heterogeneous paradigms into a universal feature space.

2.2 Graph and Riemannian Representations for EEG

Beyond temporal waveforms, EEG carries structural information along two complementary lines. Graph models such as DGCNN treat electrodes as nodes and learn inter-channel dependencies (Song et al., 2018), while convolutional-recurrent models jointly encode spectral, spatial, and temporal

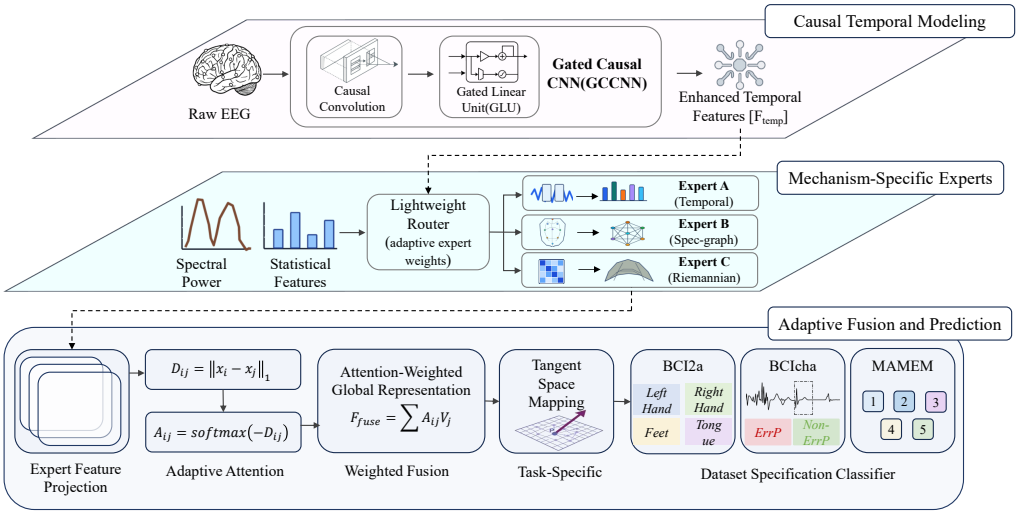


Figure 2: Overview of the proposed MGNet framework for online heterogeneous EEG decoding. MGNet first performs causal temporal modeling to obtain online-compatible temporal representations, and then extracts mechanism-specific features through temporal, spectral-dynamic graph, and Riemannian experts. A lightweight router and task-adaptive expert fusion module coordinate the expert outputs for task-specific prediction across different BCI paradigms.

cues for multichannel EEG decoding (Shen et al., 2020). Riemannian methods exploit the SPD covariance geometry of multi-channel EEG: SPD networks preserve manifold structure (Huang & Van Gool, 2017), and Riemannian representations have shown strong effectiveness for EEG classification (Congedo et al., 2017; Yger et al., 2016). However, these representations are typically developed as separate pipelines or combined through fixed fusion, which limits adaptation across heterogeneous paradigms. Recent multi-scale EEG networks, such as MSCFormer (Zhao et al., 2025), further combine multi-branch convolution with Transformer modeling to capture local and global EEG features. Nevertheless, they mainly enhance feature diversity through multi-scale feature extraction rather than explicitly routing mechanism-specific representations according to task characteristics. This motivates mechanism-specific expert pathways whose contributions can be adaptively routed across tasks.

2.3 Expert Routing and Adaptive Fusion for EEG

Mixture-of-Experts (MoE) offers a natural way to coordinate heterogeneous mechanisms: a gating network combines specialized experts (Jacobs et al., 1991), later scaled through sparse conditional computation and stochastic routing (Shazeer et al., 2017; Bengio et al., 2013). Recent EEG studies have also explored expert-based modeling, such as EvoMoE for SSVEP decoding, demonstrating the potential of adaptive expert specialization for EEG variability (Yang et al., 2025). However, existing MoE approaches typically specialize experts through parameter diversity, rather than explicitly preserving distinct signal generation mechanisms. For fusion, scaled dot-product attention (Vaswani et al., 2017) is widely used but sensitive to high-amplitude outliers, whereas distance-aware formulations improve stability under noise (Han et al., 2023). In heterogeneous EEG decoding, fusion should therefore not be a fixed universal operation but should adaptively coordinate mechanism-specific experts according to task-dependent preferences and representation complementarity.

3 Methodology

We cast EEG decoding as an *online heterogeneous representation learning* problem. Because MI, ErrP, and SSVEP rely on distinct temporal, spectral-channel, and covariance-geometric structures (Tangermann et al., 2012; Chavarriaga & Millán, 2010; Oikonomou et al., 2016; Barachant

et al., 2011), while online deployment demands causality, low latency, and artifact robustness, we propose MGNet, an online mechanism-specific expert framework with adaptive routing (Fig. 2). MGNet has four stages: (i) causal temporal modeling for online-compatible features, (ii) mechanism-specific parallel experts that separately model temporal dynamics, spectral-dynamic graph interactions, and Riemannian covariance geometry, (iii) adaptive routing and task-adaptive cross-expert fusion, and (iv) Riemannian tangent-space projection with task-specific prediction. The key idea is to *preserve* heterogeneous mechanisms in specialized experts and *adaptively* integrate them, rather than forcing them into a single encoder.

3.1 Causal Temporal Modeling

Let $X \in \mathbb{R}^{T\times C}$ be an EEG segment with T timesteps and C channels. Since online BCI cannot access future samples, we use causal convolution (Oord et al., 2016; Bai et al., 2018) as the temporal operation:

$$Y_t = \sum_{\tau=0}^{k-1} W_\tau X_{t-\tau}, \tag{1}$$

where Y_t depends only on current and past observations, preventing future-information leakage. A Gated Linear Unit (GLU) (Dauphin et al., 2017) then modulates local responses,

$$\mathrm{GLU}(Z) = Z_1 \odot \sigma(Z_2), \quad Z = \mathrm{Conv}(X), \tag{2}$$

which learns to suppress sporadic high-amplitude artifacts under the downstream loss without explicit EOG/EMG removal. Unlike RNNs (sequential, $O(T)$) or masked self-attention (non-causal-free but $O(T^2)$) (Vaswani et al., 2017), gated causal convolution is causal, parallelizable, and linear $O(kT)$, which suits high-rate EEG streams (see the Supplementary Material for a detailed comparison of temporal encoding alternatives). Its output F_{temp} is shared by all experts.

3.2 Mechanism-Specific Parallel Expert Processing

Most EEG decoders act as unified feature extractors (Song et al., 2022; Lawhern et al., 2018; Schirrmeister et al., 2017), which can entangle heterogeneous cues when paradigms differ. Since different paradigms emphasize different mechanisms, including time-locked waveforms for ErrP, spatial covariance for MI, and spectral-channel responses for SSVEP (Nakanishi et al., 2017; Liu et al., 2023), and multi-branch designs improve robustness by separating feature types (Ding et al., 2022; Mane et al., 2021; Zhao et al., 2025; Deng et al., 2024), MGNet preserves these mechanisms through parallel experts. Inspired by the mixture-of-experts principle (Jacobs et al., 1991; Shazeer et al., 2017; Ma et al., 2018), the layer couples a lightweight router with three experts organized by *representation mechanism* rather than by dataset (see the Supplementary Material for architectural and implementation details).

The router maps global summaries of F_{temp} (channel-wise variance, band-power, and spectral descriptors) to sample-level soft weights:

$$\mathbf{r} = [r_A, r_B, r_C] = \mathrm{softmax}(g(F_{\mathrm{temp}})), \tag{3}$$

trained only through the prediction objective, so no manual expert assignment is needed. Soft (rather than hard) routing retains complementary cues from non-dominant experts, and the learned weights offer an interpretable indicator of task-dependent expert preference, which we analyze in the ablation study. The three experts process complementary structures:

- **Expert A (Temporal).** Dilated causal convolutions capture time-locked waveforms and long-range temporal dependencies, suited to event-related patterns such as ErrP.
- **Expert B (Spectral-Dynamic Graph).** A dynamic GCN (Kipf & Welling, 2016) models spectral-channel interactions with a normalized sample-conditioned topology derived from spectral descriptors, capturing connectivity relevant to SSVEP.
- **Expert C (Riemannian Covariance).** Regularized covariance features are embedded on the SPD manifold (Barachant et al., 2011; Congedo et al., 2017; Yger et al., 2016; Kim et al., 2023) using training-set reference statistics, capturing spatial covariance geometry useful for MI.

3.3 TASK-ADAPTIVE DISTANCE-AWARE EXPERT FUSION

Expert outputs are heterogeneous in semantics and geometry, making direct concatenation inadequate (Ma et al., 2018). Let $\mathcal{E} = \{A, B, C\}$ denote the three experts. We first route each expert output by $\hat{F}_i = r_i F_i$ and project it into a common d-dimensional space:

$$H_i = \phi_i(\hat{F}_i), \qquad i \in \mathcal{E}, \tag{4}$$

where $\phi_i(\cdot)$ is an expert-specific projection head. This projection aligns the temporal, spectral-dynamic graph, and Riemannian expert outputs before cross-expert interaction.

We then compute pairwise expert consistency using the Manhattan distance:

$$D_{ij} = \|H_i - H_j\|_1, \qquad i, j \in \mathcal{E}. \tag{5}$$

The distance-aware fusion weights are obtained by normalizing the negative distances:

$$\alpha_{ij} = \frac{\exp(-D_{ij})}{\sum_{k\in\mathcal{E}} \exp(-D_{ik})}, \qquad \tilde{H}_i = \sum_{j\in\mathcal{E}} \alpha_{ij} H_j. \tag{6}$$

The global fused feature is then computed as

$$F_{\text{fuse}} = \frac{1}{|\mathcal{E}|} \sum_{i\in\mathcal{E}} \tilde{H}_i. \tag{7}$$

The fusion is task-adaptive because the router weights expert features for each sample before cross-expert interaction. The L_1 distance provides a robustness-motivated consistency bias through linear penalties on coordinate-wise deviations, which may arise from residual artifacts in specific EEG channels or frequency bands. Unlike dot-product attention (Vaswani et al., 2017), it measures feature discrepancies rather than inner-product similarity. Its empirical contribution is evaluated through ablation studies.

3.4 RIEMANNIAN PROJECTION AND TASK PREDICTION

The Riemannian expert yields SPD covariance features (Brooks et al., 2019); since fusion and prediction are Euclidean, we apply tangent-space projection to preserve geometry. For an SPD matrix C and Riemannian reference mean P,

$$C_{\tan} = \log_P(C) = P^{1/2} \log\Big(P^{-1/2} C P^{-1/2}\Big) P^{1/2}, \tag{8}$$

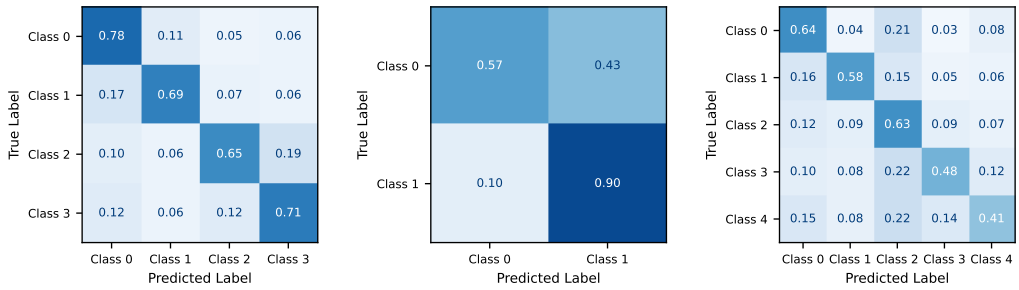
and the vectorized $C_{\tan}$ is used as the expert output. After fusion, F_{fuse} passes through a shared layer followed by task-specific heads (four-class MI for BCI IV-2a, frequency-class for MAMEM, binary ErrP for BCIcha). Only the head matching a sample's task is activated:

$$\mathcal{L} = \sum_{q\in\mathcal{Q}} \mathbb{I}(q = q_n)\, \mathcal{L}_q(y_n, \hat{y}_n), \tag{9}$$

using cross-entropy for MI/SSVEP and binary cross-entropy for ErrP. This shares upstream modules while avoiding incompatible output spaces.

4 EXPERIMENTAL EVALUATION

We evaluate MGNet in terms of **RQ1** decoding performance against neural and classical EEG baselines, **RQ2** inference efficiency, **RQ3** robustness to temporal-window changes, additive noise, and missing channels, and **RQ4** component contributions and routing behavior. Supplementary Material provides detailed dataset profiles, implementation settings, scalability curves, and extended expert-combination analyses.



(a) BCI2a (b) BCIcha (c) MAMEM

Figure 3: Normalized confusion matrices of MGNet on BCI2a, BCIcha, and MAMEM across all subjects.

Table 1: Global end-to-end task performance comparison.

Method	BCI2a	BCIcha	MAMEM
EEGNet	65.00%	75.93%	51.36%
DeepConvNet	35.60%	76.74%	47.45%
ShallowConvNet	62.69%	76.98%	57.00%
EEGConformer	58.41%	69.45%	47.18%
MAtt	62.96%	68.20%	54.54%
MSCFormer	63.04%	66.71%	57.73%
MGNet(Ours)	**69.87%**	**79.77%**	**59.18%**

4.1 Experimental Setup

4.1.1 Datasets and Task Characteristics.

We evaluate MGNet on three public BCI datasets covering heterogeneous EEG paradigms: BCI Competition IV-2a (BCI2a, motor imagery) (Tangermann et al., 2012), BCI Challenge (BCIcha, transient ErrP detection) (Chavarriaga & Millán, 2010), and MAMEM-SSVEP-II (MAMEM, frequency-driven SSVEP) (Oikonomou et al., 2016; Nikolopoulos, 2021). BCI2a contains 9 subjects, 22 channels, and 4 motor-imagery classes; BCIcha contains 16 subjects, 56 channels, and class-imbalanced ErrP trials; and MAMEM contains 11 subjects and 5 SSVEP classes. For MAMEM, we follow the preprocessed setting in MAtt and use eight occipital channels (PO7, PO3, POz, PO4, PO8, O1, Oz, O2) related to visual cortex activity. Accuracy is reported for BCI2a and MAMEM, and AUC is reported for the class-imbalanced BCIcha.

4.1.2 Baselines and Evaluation Metrics.

We compare MGNet with representative EEG decoders spanning convolutional decoders (DeepConvNet and ShallowConvNet (Schirrmeister et al., 2017)), a lightweight spatiotemporal network (EEGNet (Lawhern et al., 2018)), a convolutional Transformer (EEG Conformer (Song et al., 2022)), a recent multi-scale convolutional Transformer network (MSCFormer (Zhao et al., 2025)), and a representative manifold-aware attention model (MAtt (Congedo et al., 2017; Huang & Van Gool, 2017)). All baselines are re-implemented under a unified within-subject protocol with identical datasets, preprocessing, subject settings, and train/validation/test splits.

4.2 End-to-End Classification Performance (RQ1)

To answer **RQ1**, we first conduct a macro-level evaluation of end-to-end classification across the three paradigms. The results are summarized in Table 1.

Table 1 shows that MGNet outperforms all evaluated neural baselines under the same protocol, achieving 69.87% accuracy on BCI2a, 79.77% AUC on BCIcha, and 59.18% accuracy on MAMEM. Its consistent performance contrasts with paradigm-dependent baseline weaknesses, such as Deep-

Table 2: Comparison with classical EEG baselines. Values in parentheses denote standard deviations.

Data	Method	Base.	Ours	Δ	p
BCI2a	CSP+LDA	62.35 (17.43)	69.87 (16.56)	+7.52	.0742
BCI2a	FBCSP+LDA	60.65 (15.73)	69.87 (16.56)	+9.22	.0078
BCI2a	LE-MDM	54.28 (17.79)	69.87 (16.56)	+15.59	.0039
BCIcha	XDAWN+LDA	72.90 (12.05)	79.77 (12.32)	+6.87	.0034
MAMEM	CCA	51.82 (23.16)	59.18 (25.88)	+7.36	.0898
MAMEM	FBCCA	39.73 (14.83)	59.18 (25.88)	+19.45	.0078
MAMEM	TRCA	22.55 (5.79)	59.18 (25.88)	+36.64	.0049

Table 3: MGNet inference efficiency under the default input setting. MGNet has 1.764M trainable parameters. Thr. denotes throughput in trials/s and Lat. denotes latency in ms.

Dataset	**Ch.**	T	**Thr.**	**Lat.**
BCI2a	22	438	5964 ± 2201	0.90 ± 0.31
BCIcha	56	160	4482 ± 987	2.32 ± 1.35
MAMEM	8	125	4730 ± 57	0.41 ± 0.01

ConvNet on BCI2a and MAtt on BCIcha, supporting the value of mechanism-specific representations across heterogeneous EEG tasks.

Comparison with classical EEG pipelines. In addition to end-to-end neural EEG models, we compare MGNet with classical paradigm-specific pipelines that remain widely used in EEG decoding: CSP/FBCSP/LE-MDM for motor imagery, XDAWN for ErrP detection, and CCA/FBCCA/TRCA for SSVEP decoding. This comparison is important because these methods encode strong prior assumptions about the corresponding EEG paradigms and provide a complementary test of whether MGNet's gains are limited to deep neural baselines. Scores are reported as mean $\pm$ standard deviation across subjects; BCI2a and MAMEM use accuracy, BCIcha uses AUC, and p values are paired Wilcoxon tests.

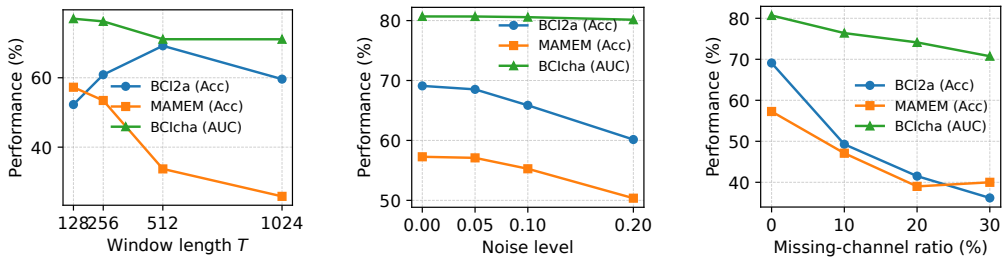
As shown in Table 2, MGNet improves over FBCSP+LDA on BCI2a and XDAWN+LDA on BCIcha, with paired Wilcoxon tests showing $p = 0.0078$ and $p = 0.0034$, respectively. On MAMEM, MGNet improves over the nearest CCA baseline by 7.36 points, although this comparison should be interpreted as a non-significant positive trend ($p = 0.0898$). These results indicate that MGNet remains competitive not only against deep neural architectures but also against established EEG-specific classical pipelines.

The confusion matrices in Fig. 3 further support these results. BCI2a shows a clear diagonal structure, with the main confusion occurring between Class 2 and Class 3 (19%), suggesting similar sensorimotor patterns. BCIcha exhibits lower recall for Class 0 than Class 1, consistent with class imbalance in ErrP detection. MAMEM shows more dispersed errors across SSVEP classes, indicating that frequency-dominant multi-class decoding remains more challenging.

4.3 Inference Efficiency Analysis (RQ2)

For **RQ2**, Table 3 reports batched throughput and single-trial latency under the default input settings. With 1.764M trainable parameters, MGNet achieves mean throughput of 4482–5964 trials/s and mean latency of 0.41–2.32 ms across datasets, supporting practical online inference under the tested conditions.

We avoid cross-implementation speed comparisons because hardware, framework, batch size, and kernel optimization affect timing. Supplementary Material reports temporal scalability on MAMEM's eight occipital channels: latency remains nearly constant and memory grows approximately linearly over the tested window lengths. The GCC layer has temporal complexity $O(kT)$ for kernel size k.



(a) Temporal window length T (b) Additive Gaussian noise (c) Missing-channel ratio

Figure 4: Robustness of MGNet under varying temporal windows, additive Gaussian noise, and missing-channel masking.

Table 4: Component and fusion ablation results.

Variant	BCI2a	MAMEM	BCIcha
Full	69.87	59.18	79.77
dot-product fusion	69.37	51.14	79.49
no_router	66.74	62.64	81.33
no_temporal	68.02	56.91	78.66
no_A	69.33	61.45	77.34
no_B	67.52	58.82	80.37
no_C	28.09	64.27	79.66
no_fusion	56.52	22.64	81.04

4.4 Robustness and Sensitivity Analysis (RQ3)

For **RQ3**, Fig. 4 evaluates temporal-window sensitivity, additive Gaussian noise, and random channel masking at inference. We report accuracy for BCI2a/MAMEM and AUC for BCIcha.

Temporal windows. BCI2a performs best at $T = 512$, whereas MAMEM and BCIcha favor shorter windows and degrade as T increases, suggesting that longer contexts can introduce task-irrelevant information.

Additive noise. BCI2a and MAMEM show moderate performance drops as Gaussian noise increases, while BCIcha remains nearly unchanged over the tested range.

Missing channels. Random channel masking causes the strongest degradation, particularly on BCI2a and MAMEM. BCIcha declines more gradually, suggesting less concentrated dependence on individual channels.

Overall, MGNet is more sensitive to window selection and channel loss than to moderate additive noise, motivating paradigm-specific temporal windows and channel-quality monitoring in deployment.

4.5 Ablation Study: Component and Expert Analysis (RQ4)

To answer **RQ4**, we analyze component removal, expert-combination diagnostics, and router behavior to examine module necessity, paradigm-dependent expert mechanisms, and routing allocation regimes.

4.5.1 Component Removal Ablation.

We disable the dynamic router, temporal stage, individual experts, Manhattan-distance bias, and fusion module in turn. Table 4 reports the results.

The contribution of each component is clearly task-dependent. On BCI2a, removing Expert C causes the largest degradation, indicating that the Riemannian covariance expert is essential for motor-imagery decoding, where covariance geometry provides highly discriminative spatial information.

Table 5: Average router weights across datasets.
A/B/C denote the three experts.

Dataset	A	B	C	Regime
BCI2a	0.333	0.333	0.333	ensemble-like
BCIcha	0.999	0.0006	0.0004	selective
MAMEM	0.331	0.340	0.330	ensemble-like

Table 6: Router collapse in pooled joint learning. Comparison with balance regularization.

Metric	Full soft	Balance reg.
Macro mean	60.11	57.05
Entropy	0.0227	1.0974
Max weight	0.9970	0.3481
Status	collapsed	balanced

Removing the fusion module also substantially reduces BCI2a performance. On MAMEM, the fusion module is the most critical component, suggesting that this task relies on coordinated expert integration. On BCIcha, removing Expert A yields the largest AUC drop, consistent with the temporal morphology of ErrP signals. Notably, the no-router variant outperforms the full model on MAMEM and BCIcha, suggesting that sample-level routing may introduce extra estimation burden under limited within-subject data.

Compared with conventional dot-product fusion, the proposed Manhattan-distance fusion achieves comparable performance on BCI2a and BCIcha while providing a clear advantage on MAMEM. This suggests that distance-based representation consistency is useful for aligning heterogeneous expert features under diverse signal characteristics. Overall, the ablation results indicate that MGNet benefits from mechanism-specific experts with task-dependent integration, while the router should be interpreted as a coordination and diagnostic module rather than an unconditional accuracy booster.

Extended single-expert and expert-combination results are provided in the Supplementary Material, further supporting the paradigm-dependent role of the specialized experts.

4.5.2 Router weight analysis.

Table 5 shows strongly selective average routing toward temporal Expert A on BCIcha, consistent with ErrP morphology and the component ablation. BCI2a and MAMEM instead have near-uniform average weights. These dataset-level patterns complement the ablations but do not imply that each paradigm selects one theoretically matched expert, or that near-uniform averages establish balanced routing for individual samples.

4.5.3 Router collapse and balance regularization.

We diagnose router collapse under pooled heterogeneous joint learning, separately from the main performance setting. In Table 6, the full soft router nearly collapses to one expert, whereas balance regularization increases entropy and reduces the maximum weight. However, the macro mean falls from 60.11 to 57.05. We therefore omit balance regularization from the main configuration and use this diagnostic to expose the trade-off between routing diversity and predictive performance.

5 Conclusion

In this paper, we proposed MGNet, a mechanism-specific expert framework for heterogeneous EEG decoding. Instead of using a single shared representation, MGNet preserves temporal, spectral-dynamic graph, and Riemannian covariance cues with dedicated experts, and coordinates them through adaptive routing and Manhattan-distance fusion. Experiments on three public datasets show strong and balanced performance, while ablations link MI to covariance geometry, ErrP to temporal modeling, and SSVEP to coordinated fusion. Router analysis further reveals selective and ensemble-like regimes, indicating that routing serves as a coordination and diagnostic module rather than an unconditional accuracy booster. These results support mechanism-specific representation learning for heterogeneous EEG decoding.

Ethics statement

This study uses existing public EEG datasets (BCI Competition IV-2a, BCI Challenge, and MAMEM-SSVEP-II) and involves no new participant recruitment or data collection. We did not attempt to re-identify participants. MGNet is evaluated as a research method for EEG decoding; the reported results do not establish clinical safety or effectiveness. Any future deployment should address neural-data privacy, informed consent, and performance variability across individuals.

Reproducibility statement

The MGNet architecture and learning objectives are described in Section 3. Dataset profiles, implementation settings, and data partitioning are documented in Appendix B, with additional experiments in Appendix C.We provide the implementation code in the supplementary archive `MGNet.zip` to facilitate reproduction of the reported results.

AI Use Statement

We used AI tools to assist with literature retrieval and discovery, LaTeX formatting, figure and table layout adjustments, and manuscript language editing and condensation. The authors independently verified the references and reviewed all AI-assisted revisions. The authors take full responsibility for the final content.

References

Hamdi Altaheri, Ghulam Muhammad, and Mansour Alsulaiman. Physics-informed attention temporal convolutional network for eeg-based motor imagery classification. *IEEE transactions on industrial informatics*, 19(2):2249–2258, 2022.

Hamdi Altaheri, Ghulam Muhammad, Mansour Alsulaiman, Syed Umar Amin, Ghadir Ali Altuwaijri, Wadood Abdul, Mohamed A Bencherif, and Mohammed Faisal. Deep learning techniques for classification of electroencephalogram (eeg) motor imagery (mi) signals: A review. *Neural computing and applications*, 35(20):14681–14722, 2023.

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

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, 2011.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Daniel Brooks, Olivier Schwander, Frédéric Barbaresco, Jean-Yves Schneider, and Matthieu Cord. Riemannian batch normalization for spd neural networks. *Advances in neural information processing systems*, 32, 2019.

Ricardo Chavarriaga and José del R Millán. Learning from eeg error-related potentials in noninvasive brain-computer interfaces. *IEEE transactions on neural systems and rehabilitation engineering*, 18(4):381–388, 2010.

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.

Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International conference on machine learning*, pp. 933–941. PMLR, 2017.

Haodong Deng, Mengfan Li, Jundi Li, Miaomiao Guo, and Guizhi Xu. A robust multi-branch multi-attention-mechanism eegnet for motor imagery bci decoding. *Journal of Neuroscience Methods*, 405:110108, 2024.

Yi Ding, Neethu Robinson, Su Zhang, Qiuhao Zeng, and Cuntai Guan. Tsception: Capturing temporal dynamics and spatial asymmetry from eeg for emotion recognition. *IEEE Transactions on affective computing*, 14(3):2238–2250, 2022.

Xiaotong Gu, Zehong Cao, Alireza Jolfaei, Peng Xu, Dongrui Wu, Tzyy-Ping Jung, and Chin-Teng Lin. Eeg-based brain-computer interfaces (bcis): A survey of recent studies on signal sensing technologies and computational intelligence approaches and their applications. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 18(5):1645–1666, 2021. doi: 10.1109/TCBB.2021.3052811.

Xing Han, Tongzheng Ren, Tan Nguyen, Khai Nguyen, Joydeep Ghosh, and Nhat Ho. Designing robust transformers using robust kernel density estimation. *Advances in Neural Information Processing Systems*, 36:53362–53384, 2023.

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.

Thorir Mar Ingolfsson, Michael Hersche, Xiaying Wang, Nobuaki Kobayashi, Lukas Cavigelli, and Luca Benini. Eeg-tcnet: An accurate temporal convolutional network for embedded motor-imagery brain–machine interfaces. In *2020 IEEE international conference on systems, man, and cybernetics (SMC)*, pp. 2958–2965. IEEE, 2020.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

Byung Hyung Kim, Jin Woo Choi, Honggu Lee, and Sungho Jo. A discriminative spd feature learning approach on riemannian manifolds for eeg classification. *Pattern Recognition*, 143:109751, 2023.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

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.

Ke Liu, Mingzhao Yang, Xin Xing, Zhuliang Yu, and Wei Wu. Sincmsnet: a sinc filter convolutional neural network for eeg motor imagery classification. *Journal of Neural Engineering*, 20(5): 056024, 2023.

Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H Chi. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1930–1939, 2018.

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.

Masaki Nakanishi, Yijun Wang, Xiaogang Chen, Yu-Te Wang, Xiaorong Gao, and Tzyy-Ping Jung. Enhancing detection of ssveps for a high-speed brain speller using task-related component analysis. *IEEE Transactions on Biomedical Engineering*, 65(1):104–112, 2017.

Spiros Nikolopoulos. Mamem eeg ssvep dataset ii (256 channels, 11 subjects, 5 frequencies presented simultaneously). *Centre Res. Technol. Hellas, Inf. Technol. Inst., Thessaloniki, Greece, Tech. Rep*, pp. 5, 2021.

Vangelis P Oikonomou, Georgios Liaros, Kostantinos Georgiadis, Elisavet Chatzilari, Katerina Adam, Spiros Nikolopoulos, and Ioannis Kompatsiaris. Comparative evaluation of state-of-the-art algorithms for ssvep-based bcis. *arXiv preprint arXiv:1602.00904*, 2016.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

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.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Fangyao Shen, Guojun Dai, Guang Lin, Jianhai Zhang, Wanzeng Kong, and Hong Zeng. Eeg-based emotion recognition using 4d convolutional recurrent neural network. *Cognitive Neurodynamics*, 14(6):815–828, 2020.

Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui. Eeg emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, 11(3): 532–541, 2018.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. Eeg conformer: Convolutional transformer for eeg decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2022.

Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J Miller, Gernot R Müller-Putz, et al. Review of the bci competition iv. *Frontiers in neuroscience*, 6:55, 2012.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Jonathan R. Wolpaw. Chapter 6 - brain–computer interfaces. In Michael P. Barnes and David C. Good (eds.), *Neurological Rehabilitation*, volume 110 of *Handbook of Clinical Neurology*, pp. 67–74. Elsevier, 2013. doi: https://doi.org/10.1016/B978-0-444-52901-5.00006-X. URL https://www.sciencedirect.com/science/article/pii/B978044452901500006X.

Xiaoli Yang, Yurui Li, Jianyu Zhang, Huiyuan Tian, Shijian Li, and Gang Pan. Evomoe: Evolutionary mixture-of-experts for ssvep-eeg classification with user-independent training. *IEEE Journal of Biomedical and Health Informatics*, 2025.

Florian Yger, Maxime Berar, and Fabien Lotte. Riemannian approaches in brain-computer interfaces: a review. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(10): 1753–1762, 2016.

Ce Zhang, Young-Keun Kim, and Azim Eskandarian. Eeg-inception: an accurate and robust end-to-end neural network for eeg-based motor imagery classification. *Journal of Neural Engineering*, 18(4):046014, 2021.

Wei Zhao, Baocan Zhang, Haifeng Zhou, Dezhi Wei, Chenxi Huang, and Quan Lan. Multi-scale convolutional transformer network for motor imagery brain-computer interface. *Scientific Reports*, 15(1):12935, 2025.

Table 7: Comparison of temporal encoding alternatives. T denotes sequence length and k denotes convolution kernel size.

Approach	Causal	Parallel	Complexity
Non-causal convolution	×	High	$O(kT)$
RNN/LSTM	✓	Low	$O(T)$ seq.
Masked Transformer	✓	High	$O(T^2)$
Gated causal conv.	✓	High	$O(kT)$

Table 8: Qualitative comparison of representative heterogeneous EEG architectures. Interference-risk entries express design expectations rather than measured quantities.

Criterion	Shared	Multi-task	Fixed	MGNet
Mechanism separation	No	Partial	Yes	Yes
Adaptation level	None	Task	None	Sample
Fusion policy	Shared	Fixed	Fixed	Adaptive
Interference risk	High	Medium	Low	Low

SUPPLEMENTARY MATERIAL

This supplementary material provides design-choice analysis, experimental details, and additional scalability and expert-specialization results for MGNet.

A DESIGN-CHOICE ANALYSIS

This section compares the major architectural choices in MGNet with representative alternatives. The comparisons clarify the design rationale and complement the empirical ablations in the main paper.

A.1 TEMPORAL ENCODING ALTERNATIVES

Table 7 compares commonly used temporal encoders. Gated causal convolution preserves temporal causality, supports parallel execution, and has linear complexity in sequence length, making it suitable for online EEG decoding (Oord et al., 2016; Bai et al., 2018). By contrast, recurrent models require sequential computation, whereas masked global self-attention has quadratic complexity in temporal length (Vaswani et al., 2017).

A.2 ARCHITECTURES FOR HETEROGENEOUS EEG PROCESSING

Table 8 qualitatively contrasts representative architectural paradigms. A shared encoder provides little mechanism separation, while a conventional multi-task model generally adapts at the task level. A fixed expert ensemble separates representations but lacks sample-dependent coordination. MGNet combines mechanism-specific experts with sample-level adaptive routing, following the general mixture-of-experts principle (Jacobs et al., 1991; Shazeer et al., 2017). These comparisons describe the representative configurations considered here rather than all possible implementations of each paradigm.

A.3 FUSION ALTERNATIVES

Expert outputs differ in semantic content and feature geometry. Concatenation alone and fixed averaging do not explicitly model adaptive cross-expert interactions. Table 9 summarizes representative fusion alternatives. Dot-product attention can be influenced by feature magnitudes, whereas Manhattan-distance attention uses linear coordinate-wise penalties as a robustness-motivated consistency bias (Vaswani et al., 2017; Han et al., 2023). The qualitative sensitivity rankings below express design intuition; actual robustness depends on feature scaling, normalization, and the complete model, and is evaluated empirically in the main paper.

Table 9: Qualitative design comparison of expert-fusion alternatives. Outlier-sensitivity entries are heuristic expectations, not universal guarantees or measured rankings.

Fusion Method	Adaptive	Outlier Sensitivity
Concatenation	No	High
Dot-product attention	Yes	Medium
Euclidean-distance attention	Yes	High
Manhattan-distance attention	Yes	Lower

Table 10: Profiles of the benchmark datasets. C denotes the number of input channels and f_s the sampling rate used in the stated configuration. For MAMEM, eight occipital channels are selected from the original 256-channel recordings.

Dataset	C	f_s	Subj.	Cls.	Dominant Cue
BCI2a	22	256 Hz	9	4	Covariance
BCIcha	56	128 Hz	16	2 (imb.)	Transient
MAMEM	8	250 Hz	11	5	Frequency

B Experimental Details

B.1 Dataset Profiles

The experiments cover three EEG paradigms with different dominant discriminative cues: motor imagery (MI), error-related potentials (ErrP), and steady-state visual evoked potentials (SSVEP). Table 10 summarizes their principal characteristics.

BCI2a (Motor Imagery). BCI Competition IV-2a contains recordings from nine subjects over two sessions (Tangermann et al., 2012). Each subject performs four MI tasks: left hand, right hand, feet, and tongue. The input configuration reported here uses 22 channels at 256 Hz. The task primarily depends on sensorimotor-rhythm modulation and spatial covariance structure.

BCIcha (Error-Related Potentials). The BCI Challenge dataset contains ErrP recordings collected during feedback in a P300-speller paradigm (Chavarriaga & Millán, 2010). It includes 16 subjects and 56 electrodes. Each trial lasts 1.25 s and is downsampled to 128 Hz. Because the two classes are imbalanced, performance is evaluated using AUC.

MAMEM (SSVEP). MAMEM-SSVEP-II was recorded with a 256-channel Electrical Geodesics Inc. system (Oikonomou et al., 2016; Nikolopoulos, 2021). Following the adopted preprocessing protocol, we retain eight occipital electrodes (PO7, PO3, POz, PO4, PO8, O1, Oz, and O2). The dataset contains 11 subjects and five stimulation frequencies: 6.66, 7.50, 8.57, 10.00, and 12.00 Hz.

B.2 Implementation Details

Hardware and optimization. All experiments are implemented in PyTorch and conducted on a Linux server with an NVIDIA A40 GPU (45 GB available memory), a 104-core CPU, and 251.5 GB RAM. Models are optimized using Adam with an initial learning rate of 10^{-3} and a batch size of 64. Early stopping uses a patience of 15 epochs based on validation loss, and the checkpoint with the lowest validation loss is retained for evaluation.

Input construction. For BCI2a, trials are cropped from sample indices 124 to 562, producing 438 temporal samples per trial. The dataset-specific temporal lengths used in the main experiments are reported in Table 3.

Baseline implementation. All neural baselines are re-implemented or adapted under the same preprocessing, subject split, and training protocol as MGNet. Dataset-specific input shapes, evaluation metrics, and train/validation/test partitions are kept identical across methods. For MSCFormer,

we adapt the output head and input interface to each dataset while keeping the same evaluation protocol. Hyperparameters are selected on the validation partition, and test labels are used only for final evaluation.

Dynamic graph construction. For the spectral-dynamic graph expert, the graph topology is constructed separately for each EEG segment. We first extract spectral descriptors from the temporal features, including channel-wise band-power statistics, and use them to compute sample-conditioned channel affinities. The resulting affinity matrix is symmetrized, augmented with self-loops, and normalized with a row-wise softmax. The normalized adjacency is then used by the graph convolution layer for spectral-channel message passing, with edge weights reflecting trial-dependent relationships.

Riemannian covariance estimation. For the Riemannian covariance expert, channel covariance matrices are estimated from each EEG segment and regularized before SPD embedding. Specifically, we use shrinkage and diagonal regularization:

$$\Sigma_\epsilon = (1-\lambda)\Sigma + \lambda \frac{\mathrm{tr}(\Sigma)}{C} I + \epsilon I, \tag{10}$$

where C is the number of channels and ϵ ensures positive definiteness. The Riemannian reference mean P used for tangent-space projection is estimated only from the training partition of each subject and dataset. Validation samples are used only for model selection, and test samples are used only for final evaluation. This protocol keeps the train/validation/test partitions separated and avoids information leakage.

Supplementary code. The implementation of MGNet is provided in the supplementary archive `MGNet.zip` to facilitate reproduction of the reported results.

B.3 Within-Subject Data Partitioning

The main decoding experiments follow a within-subject protocol: training, validation, and testing are conducted independently for each subject, and no trials from the test partition are used for optimization or model selection. The pooled heterogeneous joint-learning diagnostic in Table 6 is a separate experimental setting.

BCI2a. We use the official session-based protocol. The evaluation session is used as the test set. A class-balanced validation subset is constructed from the training session through stratified sampling, and the remaining trials form the training set.

BCIcha and MAMEM. Trials are split sequentially while preserving their original temporal order. The validation partition is used only for model selection and does not contribute gradient updates.

C Additional Experimental Results

C.1 Temporal Scalability

We analyze the computational behavior of MGNet as temporal window length T increases. The sweep is conducted on MAMEM using the eight selected occipital channels. For fixed kernel size k, gated causal convolution has complexity $O(kT)$, motivating an examination of temporal scaling without the quadratic attention matrices used by global self-attention.

Figure 5 reports latency and memory usage as T increases from 128 to 1024. Latency varies only slightly around 7.1 ms in this controlled sweep, while memory increases from 16.51 MB to 34.47 MB. Because absolute latency depends on batch size and the benchmarking protocol, this experiment evaluates scaling trends rather than directly comparing absolute values with the single-trial latency reported in Table 3. Over the evaluated range, MGNet exhibits stable latency and approximately linear memory growth.

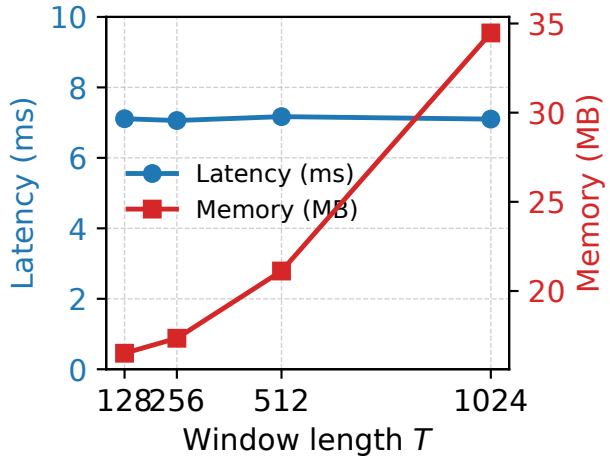


Figure 5: Temporal scalability of MGNet on MAMEM. Latency remains approximately stable over the evaluated window lengths, while measured memory usage grows approximately linearly with T.

Table 11: Expert-combination results. A, B, and C denote the temporal, spectral-dynamic graph, and Riemannian covariance experts, respectively. Values are accuracy (%) for BCI2a and MAMEM, and AUC (%) for BCIcha.

Data	A	B	C	A+B	A+C	B+C	A+B+C
BCI2a	44.37	49.15	**66.63**	43.13	41.47	55.25	43.21
MAMEM	62.55	**65.18**	58.18	61.64	59.73	56.45	57.18
BCIcha	74.16	76.26	67.21	76.99	73.11	75.43	**78.48**

C.2 Expert Specialization and Combinations

To examine whether the experts encode distinct task-dependent information, we train single-expert and expert-combination variants with the router and distance-aware fusion disabled. Each selected expert output is projected to a 128-dimensional representation, and multiple selected outputs are concatenated before the task-specific classifier.

The results reveal paradigm-dependent preferences. BCI2a obtains its strongest single-expert result with the Riemannian covariance expert (C, 66.63%), consistent with the importance of spatial covariance in MI. MAMEM favors the spectral-dynamic graph expert (B, 65.18%), consistent with its frequency-selective visual responses. BCIcha obtains the strongest result from the concatenated three-expert representation (A+B+C, 78.48%), suggesting that ErrP decoding benefits from complementary information.

These variants are independently retrained after bypassing the router and distance-aware fusion. They therefore measure the representational sufficiency of selected branches under a different classifier, rather than constituting nested removals from the trained full model. Their raw scores should be interpreted together with, rather than as replacements for, the component ablations in the main paper.

Routing-regime interpretation. Router weights characterize how the full model coordinates available experts. Table 5 reports strongly concentrated average weights on temporal Expert A for BCIcha and near-uniform average weights for BCI2a and MAMEM. These aggregate patterns complement the component and expert-combination analyses, but near-uniform dataset averages do not establish high-entropy routing for individual samples. The separate pooled-learning diagnostic in Table 6 directly reports router entropy and maximum weight to characterize collapse and its performance trade-off.