# DOES ATTENTION MATTER IN MOTOR IMAGERY EEG DECODERS? DIAGNOSING AND REPAIRING ATTENTION FAILURE

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

## ABSTRACT

Attention is widely used in hybrid motor-imagery (MI) EEG decoders to dynamically select and integrate evidence for each trial. However, performance gains of the complete decoder and its attention maps do not reveal whether this trial-specific weighting actually contributes to classification. At frozen checkpoints, we replace attention with cross-sample or training-set-mean weights while holding the current trial's Value tensors and all parameters fixed; uniform replacement provides an additional fixed-route control. Across eight decoders on BCIC-IV-2b and BCIC-IV-2a, no architecture shows jointly supported dependence on the two empirical replacements in both datasets. Probes and controlled perturbations trace this failure to token convergence, Value-space equivalence, and downstream bypass or compensation. We therefore develop SOAMNet, which preserves multi-scale temporal trajectories, separates class queries, and forces evidence through one sample-conditioned attention path. Across four MI datasets and three seeds, SOAMNet achieves the highest mean BA and Cohen's $\kappa$ among 12 baselines. Cross-sample and training-set-mean replacement consistently reduce its BA on every dataset; the weaker mean decrease ranges from 18.61 to 40.95 percentage points, and both participant-bootstrap intervals exclude zero throughout. Replacing attention with a matched trainable non-attention head lowers BA by 3.51–5.22 points. SOAMNet thus realizes functional reliance on sample-conditioned attention weights while achieving competitive classification performance.

## 1 INTRODUCTION

Attention has become a routine component of hybrid motor-imagery (MI) EEG decoders. Convolution extracts local spectral and spatial features, while attention is expected to select and integrate the sensorimotor and temporal evidence relevant to each trial. This division is physiologically plausible because MI responses evolve across channels, mu and beta rhythms, and event-related desynchronization and synchronization (Pfurtscheller & Lopes da Silva, 1999). Yet neither the performance of a complete decoder nor its attention maps establishes that classification depends on the sample-conditioned weights produced by its attention block.

ATCNet, EEG Conformer, CTNet, TCFormer, DBConformer, MVCNet, RatioWaveNet, and SSTAF exemplify this hybrid design (Altaheri et al., 2023; Song et al., 2023; Zhao et al., 2024; Altaheri et al., 2025; Wang et al., 2026; 2025; Siino et al., 2026; Muna et al., 2025). We ask a direct counterfactual question: does balanced accuracy deteriorate when a frozen model no longer uses the attention weights computed for the current trial? Our audit changes only those weights, retaining the current trial's Value tensors and every other computation. On BCIC-IV-2b, no audited decoder loses BA consistently under both cross-sample and training-set-mean replacement, and no architecture receives joint support on both BCIC-IV-2b and BCIC-IV-2a. Zeroing the attention output can nevertheless remain harmful, showing that a branch may contribute a useful representation while its sample-conditioned weighting is dispensable.

This distinction is not resolved by system-level accuracy, heatmaps, or block deletion: these respectively compare complete architectures, describe weight morphology, or remove weighting together with its Value transformations and projections. Related work has exposed limitations in attention



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.

faithfulness and identifiability (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019; Serrano & Smith, 2019; Abnar & Zuidema, 2020; Chefer et al., 2021; Brunner et al., 2020). We instead trace a route perturbation through attention weights, routed representations, and task performance.

The analysis locates weak functional reliance at three points. Row-stochastic attention preserves the token-common component of Value; the effective rank of centered Value tokens limits which weight changes remain visible; and residuals, pooling, or downstream temporal modules can expose or reconstruct evidence outside the changed route. Representation probes, equal-norm Value-subspace perturbations, and architectural factorials test these predictions on BCIC-IV-2b.

These mechanisms motivate three constraints for a route-dependent decoder: retain identifiable local evidence until routing, keep route changes visible after Value multiplication, and prevent sufficient class evidence from bypassing the audited route. SOAMNet—the *Single-Path Orthogonal-Class Attention Multi-Window Network*—aligns complete temporal trajectories from overlapping windows, uses class-conditioned routing with a uniform-null readout, and ends the decision path at attention pooling. We evaluate it on BCIC-IV-2b, BCIC-IV-2a, BNCI2015-001, and Cho2017.

Our contributions are:

1. a seven-intervention frozen-checkpoint audit that separates branch participation, route form, token alignment, and current-trial adaptivity;
2. an information-path account of attention failure through token consensus, Value-space equivalence, and decision bypass, supported by probes, controlled perturbations, and architectural factorials; and
3. SOAMNet, a constructive route-dependent decoder that ranks first in mean BA and $\kappa$ among twelve baselines across four datasets, while both empirical route replacements reliably reduce BA on every dataset.

## 2 Related work

**MI decoding.** Classical FBCSP combines subband spatial filtering with a linear classifier (Ramoser et al., 2000; Ang et al., 2008; Blankertz et al., 2008); compact neural baselines include EEGNet and FBCNet (Lawhern et al., 2018; Mane et al., 2021) within the wider deep EEG literature (Lotte et al., 2018; Schirrmeister et al., 2017; Roy et al., 2019). Hybrid decoders instead couple convolutional representations to attention at different points in the information path. ATCNet routes overlapping windows before a TCN, EEG Conformer places a Transformer after convolutional tokenization, CTNet and TCFormer fuse convolutional and Transformer features, and recent DBConformer, MVCNet, RatioWaveNet, SSTAF, and EEG-CSANet introduce parallel branches, specialized front ends, or finer attention gates (Altaheri et al., 2023; Song et al., 2023; Zhao et al., 2024; Altaheri et al., 2025; Wang et al., 2026; 2025; Siino et al., 2026; Muna et al., 2025; Cai et al., 2025). These choices determine whether unrouted features, residual tokens, pooling, or downstream modules can provide an alternative route to the classifier.

**Functional evaluation of attention.** Self-attention computes content-conditioned mixtures (Vaswani et al., 2017), but plausible weights may disagree with attribution measures, alternative weights may preserve predictions, and deletion can confound several computations (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019; Serrano & Smith, 2019). Head pruning and specialization likewise show that many learned heads can be removed with little performance change (Michel et al., 2019; Voita et al., 2019); attention flow and relevance propagation account explicitly for mixing and skip connections (Abnar & Zuidema, 2020; Chefer et al., 2021). Attention can also be non-identifiable when sequence length exceeds head dimension (Brunner et al., 2020). We evaluate the computation directly: with the current trial's Values fixed, does replacing its route change task performance? This complements attribution stress tests (Sundararajan et al., 2017; Shrikumar et al., 2017; Adebayo et al., 2018; Hooker et al., 2019) and distinguishes checkpoint dependence from compensation after retraining.

**Failure through information shortcuts.** Patch tokenization trades temporal resolution for representation and compute (Nie et al., 2023), while temporal convolutions provide a strong sequence-modeling alternative (Bai et al., 2018). Direct attention perturbations in forecasting Transformers

have linked weak effects to inadequate latent representations (Liang et al., 2025). MI hybrids add domain-specific encoders, Value geometry, and downstream temporal decoders. Shortcut learning predicts that optimization will exploit any sufficient path (Geirhos et al., 2020); residual connections make such access explicit (He et al., 2016). Our framework tests whether token consensus, low-rank Value, or downstream access makes the audited route dispensable.

# 3 FUNCTIONAL RELIANCE OF ATTENTION

## 3.1 ATTENTION, REPRESENTATIONS, AND DECISIONS

Let an encoder map EEG trial $X$ to $L$ Value tokens. For one attention head,

$$A(X) = \mathrm{softmax}\Big(Q(X)K(X)^\top/\sqrt{d_k}\Big), \qquad Z(X) = A(X)V(X), \tag{1}$$

and the downstream prediction is

$$f_A(X) = g(Z(X), R(X)), \tag{2}$$

where $R(X)$ denotes residuals, unrouted branches, raw-token pooling, or any other alternative access to the representation. Replacing $A(X)$ by $A_m(X)$ while holding $V(X)$, all parameters, and $R(X)$ fixed gives

$$D_A^{(m)} = \|A - A_m\|_F, \quad D_Z^{(m)} = \|(A - A_m)V\|_F, \quad \Delta_m = \mathrm{BA}(f_A) - \mathrm{BA}(f_{A_m}). \tag{3}$$

These quantities separate a changed map, a changed routed representation, and a changed task decision.

## 3.2 FROZEN-CHECKPOINT FUNCTIONAL AUDIT

Seven interventions probe distinct properties of a frozen attention path. Zero removes its routed output; uniform and canonical one-hot weights compare learned mixing with fixed routes; shuffle and reverse keys break weight–Value alignment; and cross-sample and training-set-mean weights test whether the current trial's route is replaceable by a label-free maximum-dissimilarity cyclic partner or a fixed empirical template. Supplementary Section B gives exact definitions and the synchronized-permutation implementation check.

We define *functional reliance* as dependence of task performance on the weights generated for the current trial. For participant $i$ and empirical replacement $k \in \{X, M\}$, let $\Delta_{ik} = \mathrm{BA}_{i,\mathrm{clean}} - \mathrm{BA}_{i,k}$ and let $\mu_k$ be its participant-level mean. The primary intersection–union test is

$$H_0 : (\mu_X \le 0) \cup (\mu_M \le 0), \qquad H_1 : (\mu_X > 0) \cap (\mu_M > 0). \tag{4}$$

Reliance is supported only when both participant-bootstrap 95% intervals have positive lower endpoints. We also report the continuous strength

$$R_{\mathrm{core}} = \min_{k \in \{X,M\}} \widehat{\mu}_k. \tag{5}$$

Uniform replacement is a route-form control; for an analytic uniform-null construction, it verifies the construction rather than learned route selection. Zeroing separately tests whether the complete routed branch participates.

## 3.3 AUDIT RESULTS AND CONCLUSION

We applied this protocol to eight hybrid decoders on BCIC-IV-2b. Before intervention, the corresponding checkpoints reach 71.27–78.54% clean BA (Supplementary Table S10). Figure 1 reports the mean BA and $\kappa$ decreases under all seven interventions, allowing branch removal, fixed routes, disrupted token alignment, and empirical route replacement to be compared within one view.

Across architectures, $R_{\mathrm{core}}$ ranges from $-0.51$ to 0.89 pp, and no model has positive lower confidence bounds for both empirical effects. Zeroing substantially degrades ATCNet, EEG Conformer, DBConformer, and RatioWaveNet, whereas replacing their nonzero routes has little aggregate effect. The routed representations are useful, but the current trial's particular mixtures are replaceable. To reduce the risk of a dataset-specific audit conclusion, we repeated the same protocol on BCIC-IV-2a and obtained the same cross-dataset conclusion: no architecture receives joint support on both datasets (Supplementary Fig. S1 and Table S2).

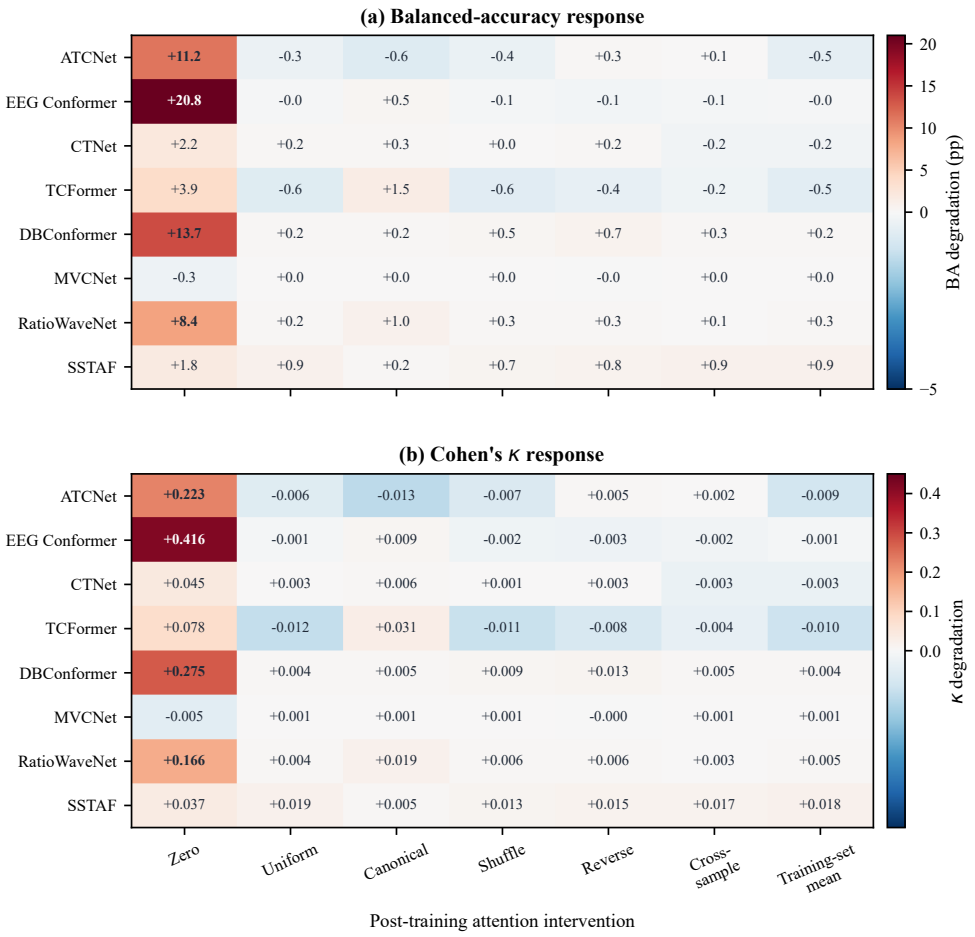



Figure 1: Unified frozen-checkpoint audit on BCIC-IV-2b. Each cell reports participant-mean degradation after replacing only attention weights while retaining the current trial's Values and all trained parameters. BA is in percentage points and $\kappa$ uses its native scale; positive values denote worse performance.

## 4 Attention Failure Mechanisms

A route affects classification only if discriminative evidence remains in token-varying Values, a weight change produces a non-equivalent change in $AV$, and the downstream network transmits that change to the logits. Failures of these three conditions yield token consensus, Value-space equivalence, and decision equivalence through bypass or compensation.

### 4.1 Attention-Invariant Token Consensus

Decompose Value into its token mean and centered disagreement:

$$\mu(V) = L^{-1}\mathbf{1}^\top V, \qquad V_\perp = V - \mathbf{1}\mu(V), \qquad \mathbf{1}^\top V_\perp = 0. \tag{6}$$

Because every row-stochastic $A$ satisfies $A\mathbf{1} = \mathbf{1}$,

$$AV = \mathbf{1}\mu(V) + AV_\perp. \tag{7}$$

The consensus term reaches every query unchanged. If it contains a sufficient class statistic, classification need not exploit token differences. Leakage-safe probes fitted to $\mu(V)$, $V_\perp$, and $AV$ distinguish readability from the classifier's actual route dependence.

### 4.2 VALUE-SPACE AND DECISION EQUIVALENCE

Two attention matrices are representation-equivalent whenever

$$(A - A')V_\perp = 0. \tag{8}$$

For one query, a legal zero-sum route perturbation has $L - 1$ degrees of freedom, while the visible map has rank at most $\mathrm{rank}(V_\perp)$; hence

$$\dim \ker T_V \geq (L-1) - \mathrm{rank}(V_\perp), \qquad T_V : \delta a \mapsto \delta a V_\perp. \tag{9}$$

Nominal width therefore does not guarantee identifiable routing when effective token disagreement collapses. Even when $AV$ changes, the local approximation

$$\Delta f \approx J_g \,\mathrm{vec}((A' - A)V_\perp) \tag{10}$$

shows how downstream insensitivity, residuals, parallel branches, pooling, or temporal modules can absorb the change. Equal-norm null and Value-sensitive perturbations isolate representation equivalence; residual-by-Value and route-by-TCN factorials test downstream access and compensation.

### 4.3 EMPIRICAL DIAGNOSIS

We applied a common diagnostic suite to 72 frozen BCIC-IV-2b checkpoints. It measures Value consensus energy and centered rank, split-safe linear probes, route geometry, and relative changes in attention, routed representations, and logits under cross-sample replacement. Figure 2 summarizes the resulting information-path evidence.

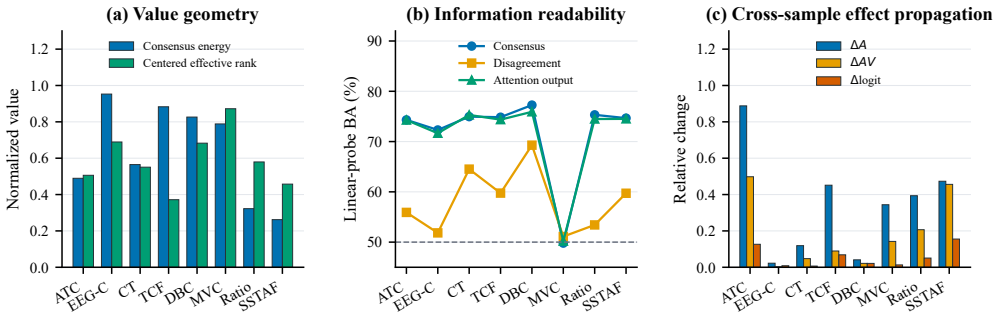



Figure 2: Information-path diagnostics on BCIC-IV-2b. (a) Value consensus energy and centered effective rank. (b) Linear-probe BA for consensus, token disagreement, and attention output; the dashed line marks chance. (c) Relative changes in attention, routed representations, and logits under cross-sample replacement. Each change is normalized by the corresponding clean tensor norm, so the three stages localize attenuation rather than form a conserved ratio.

Class-readable information is concentrated in the attention-invariant consensus in all architectures except MVCNet: consensus probes reach 72.31–77.26% BA, and attention-output probes differ by only $-1.34$ to $+0.34$ pp. High token correlation and low centered rank further reduce the visible route subspace; EEG Conformer and DBConformer additionally learn nearly uniform routes. Effect propagation localizes later attenuation: TCFormer's relative change contracts from $0.451$ in $A$ to $0.090$ in $AV$, while CTNet contracts from $0.119$ to $0.048$ and then $0.007$ at the logits. ATCNet and RatioWaveNet retain larger $AV$ changes but attenuate them downstream; MVCNet and SSTAF likewise show little task effect after visible internal changes. These patterns support the three mechanisms at successive points along the information path. Complete values are in Supplementary Tables S11 and S12.

## 5 SOAMNET: DESIGNING AGAINST ATTENTION FAILURE

### 5.1 FAILURE-TO-DESIGN MAPPING

Each diagnosed mechanism yields a path constraint: preserve local evidence identity until routing, make route changes visible after Value multiplication, and deny alternative access to sufficient class

evidence. SOAMNet, the *Single-Path Orthogonal-Class Attention Multi-Window Network*, maps these constraints to concrete components in Table 1; Figure 3 shows the resulting single-path architecture. Classification and intervention tests separately evaluate practical utility and realized route dependence.

Table 1: Mechanism-to-design mapping for SOAMNet.

| Failure | Target property | Realization |
|---|---|---|
| Token consensus | Preserve evidence identity | Align complete trajectories from overlapping windows |
| Value equivalence | Keep route changes visible | Class-wise temporal Values and uniform-centered readout |
| Prediction bypass | Require the audited route | Terminate the decision path at attention pooling |

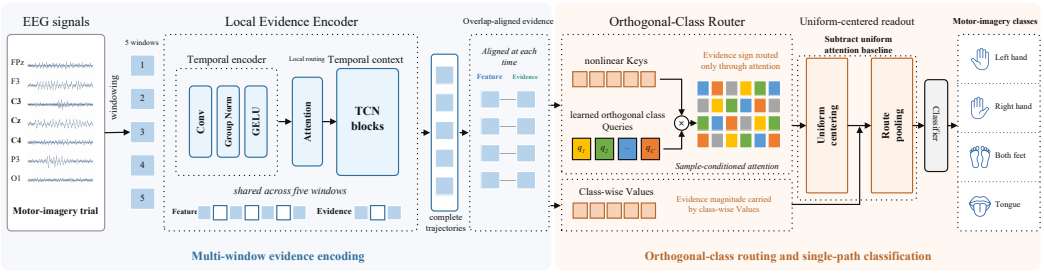



Figure 3: SOAMNet aligns complete feature and class-evidence trajectories from five shared CNN–attention–TCN windows. Orthogonal class queries, nonlinear keys, and global context define sample-conditioned attention; class-wise Values combine local evidence with routed context. Uniform-centered route pooling is the sole path to the classifier.

### 5.2 Aligned evidence and class-conditioned routing

The local encoder uses five overlapping windows and retains complete TCN feature and class-evidence trajectories rather than only terminal decisions. If $\mathcal{I}_t$ contains the windows covering global position $t$, aligned features and evidence are

$$H_t = |\mathcal{I}_t|^{-1} \sum_{w \in \mathcal{I}_t} Z_t^{(w)}, \qquad E_t = |\mathcal{I}_t|^{-1} \sum_{w \in \mathcal{I}_t} E_t^{(w)}. \tag{11}$$

where $Z_t^{(w)}$ and $E_t^{(w)}$ are respectively the feature vector and class-evidence vector produced at position $t$ by window $w$. The resulting sequences contain 15, 18, 22, and 13 tokens on BCIC-IV-2b, BCIC-IV-2a, BNCI2015-001, and Cho2017. The evidence extractor and routing path are optimized jointly as one end-to-end model.

Keys combine normalized $H_t$ with temporally standardized $E_t$. Let $r(E)$ aggregate the five terminal window scores into a class vector, let $c = g_{\mathrm{corr}}(E)$ be a learned correction, and define the global class context $m = r(E) + 0.25c$. These quantities condition the route without direct access to the logits:

$$k_t = f_K([\mathrm{LN}(H_t); \mathrm{Std}_t(E_t)]) \in \mathbb{R}^{64}, \tag{12}$$

$$s_{ct} = q_c^\top k_t / \sqrt{64} + p_{ct} + \gamma_c \tanh(c_c) h_{ct} + \gamma_m \tanh(m_c) h_{ct},$$

$$a_{ct} = \mathrm{softmax}_t(s_{ct}), \tag{13}$$

where orthogonally initialized $q_c$ are class queries, $p_{ct}$ is a learned class-specific positional bias, zero-mean $h_c$ are distinct temporal carriers, and $\gamma_c, \gamma_m$ are learned scales.

### 5.3 UNIFORM-NULL SINGLE-PATH READOUT

Class-wise Values combine local evidence with the signed correction $c_c$ and global-evidence magnitude $|m_c|$:

$$v_{ct} = \lambda_\ell(E_{tc} + f_V(H_t)_c) + \lambda_c c_c h_{ct} + \lambda_m |m_c| h_{ct}. \tag{14}$$

Here $f_V$ is a local Value adapter and $\lambda_\ell, \lambda_c, \lambda_m$ are learned scales. Let $d_{ct} = a_{ct} - T^{-1}$, $\rho_c = [\sum_t d_{ct}^2/((T-1)/T)]^{1/2}$, and $\bar{v}_c = T^{-1}\sum_t v_{ct}$. The sole classifier input is

$$z_c = \tau\left(\sum_t d_{ct} v_{ct} + \rho_c \bar{v}_c\right) + \bar{\rho}\,\beta\xi_c, \qquad \bar{\rho} = C^{-1}\sum_c \rho_c. \tag{15}$$

Here $\tau$ is a learned logit scale, $\xi_c$ is a fixed class axis evenly spaced from $-1$ to 1, and $\beta$ calibrates its route-gated operating point. Uniform attention makes $d_{ct} = \rho_c = \bar{\rho} = 0$ and therefore gives $z_c = 0$ for any values of $c_c$ and $m_c$. Neither context term has a path to the logits outside the audited readout: no encoder logit, correction output, aggregated evidence, or feature residual is added after Eq. (15). Throughout the audit, SOAMNet attention denotes this final class-conditioned route $a_{ct}$, which is the sole input to the classifier. A fixed nonuniform route can still yield nonzero logits, so cross-sample and training-set-mean replacement remain the empirical tests of current-trial dependence. Training uses cross-entropy with query-orthogonality regularization; full optimization and calibration details appear in Supplementary Section A.

## 6 EXPERIMENTAL DESIGN

The experiments ask whether trained hybrids require current-trial attention, where route influence disappears, and whether the diagnosed mechanisms can guide a route-dependent decoder with competitive classification. Frozen-checkpoint interventions, internal diagnostics, and the complete SOAMNet evaluation answer these questions in sequence.

**Data and splits.** All evaluations are subject specific and use disjoint train, validation, and test splits. BCIC-IV-2b has three channels, two classes, and nine participants; sessions 1–2, 3, and 4–5 supply the three splits (Tangermann et al., 2012). BCIC-IV-2a has 22 channels and four classes for nine participants (Tangermann et al., 2012). BNCI2015-001 provides 13-channel right-hand/feet imagery from 12 participants (Faller et al., 2012), and Cho2017 provides left/right-hand imagery from 48 participants (Cho et al., 2017). Together the frozen splits contain 24,240 clean trials. Dataset roles, windows, preprocessing, and integrity controls are reported in Supplementary Sections A and F.

**Audit and mechanism analysis.** The common seed-2026 audit covers ATCNet, EEG Conformer, CTNet, TCFormer, DBConformer, MVCNet, RatioWaveNet, and SSTAF with the same participant splits. Seven QK-softmax replacements are applied jointly while current-trial Values and parameters remain fixed; topology-specific vector gates receive a separate compatible audit. The same 72 BCIC-IV-2b checkpoints provide attention geometry, Value consensus and rank, leakage-safe probes, and propagation from $A$ through $AV$ to logits. Equal-norm perturbations and residual-by-Value and route-by-TCN factorials test equivalence, bypass, and compensation. Full targets and execution checks appear in Supplementary Tables S3–S6.

**Constructive validation.** Table 2 reports the four-dataset comparison of SOAMNet with twelve baselines. Neural methods use seeds 2026, 2036, and 2046, identical participant-wise splits, one checkpoint per participant–seed pair, validation-BA selection, and no ensembling. Epoch caps never exceed 300; SOAMNet training is capped at 120 epochs on BCIC-IV-2b and 100 elsewhere. Seven independently trained controls test a nonlinear non-attention head, local-only and global-only Values, temporal alignment, orthogonal queries, uniform-null readout, and direct logit bypass. All use the same splits, encoder architecture, seeds, budgets, and selection rules. Optimization settings and computational profiles are in Supplementary Tables S8 and S9; full control results appear in Tables S18 and S19.

Table 2: Within-subject classification across four MI datasets. Neural-model entries are mean (population standard deviation) over three seeds; FBCSP is deterministic. Bold entries denote the best result for each metric on each dataset.

| Method | BCIC-IV-2b | | BCIC-IV-2a | | BNCI2015-001 | | Cho2017 | |
|---|---|---|---|---|---|---|---|---|
| | BA (%) | $\kappa$ | BA (%) | $\kappa$ | BA (%) | $\kappa$ | BA (%) | $\kappa$ |
| FBCSP | 74.34 | 0.486 | 42.94 | 0.239 | 76.20 | 0.524 | 61.10 | 0.222 |
| EEGNet | 71.21(2.45) | .423(.049) | 57.28(1.70) | .430(.023) | 72.86(2.22) | .457(.044) | 57.37(.48) | .147(.009) |
| FBCNet | 73.29(.24) | .465(.005) | 55.90(.31) | .412(.004) | 76.48(.41) | .530(.008) | 67.38(.84) | .347(.017) |
| EEG Conformer | 74.18(2.07) | .483(.041) | 50.66(.82) | .342(.011) | 76.50(1.03) | .530(.021) | 68.12(.60) | .362(.012) |
| CTNet | 73.05(.19) | .460(.004) | 64.87(1.22) | .532(.016) | 74.38(.75) | .488(.015) | 60.97(.34) | .219(.007) |
| TCFormer | 74.99(1.69) | .498(.034) | 67.08(.16) | .561(.002) | 76.78(1.06) | .536(.021) | 67.18(.52) | .343(.011) |
| ATCNet | 76.96(.97) | .538(.020) | 67.93(.67) | .572(.009) | 81.11(1.18) | .622(.024) | 68.78(.87) | .375(.017) |
| DBConformer | 78.07(.33) | .560(.007) | 46.49(2.30) | .287(.031) | 79.18(.57) | .584(.011) | 68.13(.37) | .363(.008) |
| MVCNet | 76.68(.88) | .533(.018) | 51.07(1.11) | .348(.015) | 77.58(1.45) | .551(.029) | 60.67(.81) | .213(.016) |
| RatioWaveNet | 75.17(.11) | .502(.002) | 66.85(.53) | .558(.007) | 77.79(1.25) | .556(.025) | 65.39(.40) | .308(.008) |
| SSTAF | 72.92(1.75) | .457(.035) | 37.00(.44) | .160(.006) | 73.86(.46) | .477(.009) | 59.25(.28) | .185(.005) |
| EEG-CSANet | 75.76(.45) | .514(.009) | 49.24(1.94) | .323(.026) | 78.84(.26) | .577(.005) | 67.40(.94) | .348(.019) |
| SOAMNet (ours) | **80.10(.20)** | **.602(.004)** | **69.44(.78)** | **.593(.010)** | **85.53(.65)** | **.711(.013)** | **72.75(.79)** | **.455(.016)** |

**Metrics and inference.** We report BA and Cohen's $\kappa$ (Cohen, 1960). Participants are the independent units; three-seed results are first averaged within participant. Functional reliance uses 50,000 common participant-bootstrap resamples and requires both empirical 95% intervals to have positive lower endpoints. Classification comparisons use 100,000-resample participant-bootstrap intervals and Holm-adjusted paired $t$ and Wilcoxon tests across twelve comparators within each dataset and metric. Component controls use participant-paired $t$ intervals. Instrumented models must reproduce original logits within $10^{-6}$ and pass the synchronized-permutation identity.

# 7 Quantitative results

## 7.1 Functional audit and mechanism tests

The eight BCIC-IV-2b classifiers reach 71.27–78.54% clean BA, yet their $R_{\mathrm{core}}$ values remain between $-0.51$ and 0.89 pp, with no positive lower bounds for both empirical replacements. Zeroing four models by 8.39–20.81 pp shows that useful routed representations need not entail sample-conditioned mixing. BCIC-IV-2a gives the same joint conclusion; Supplementary Tables S2, S4, and S10 report the complete matrices and intervals.

Mechanism tests reproduce the predicted information-path interactions. Across 45 model–participant cases, consensus energy and token correlation correlate negatively with route-induced prediction flips ($\rho = -0.761$ and $-0.719$), whereas attention probe gain correlates positively ($\rho = 0.428$). Equal-norm Value-space controls further show that a null-space perturbation leaves $AV$ and BA unchanged, while a Value-sensitive perturbation changes $AV$ and reduces BA by 0.43 pp (Supplementary Table S13).

Architectural controls localize downstream access. Residual removal and Value centering each leave route effects below 1.8 pp, but together produce 24.1–27.8 pp degradation. A TCN erases an 18.6-pp current-route advantage with unrestricted access yet produces a 22.3-pp advantage after a routing bottleneck. Supplementary Tables S11–S15 and Fig. S3 give the complete protocols and values.

## 7.2 SOAMNet classification and functional reliance

SOAMNet combines frozen-checkpoint route dependence with competitive classification, ranking first in BA and $\kappa$ on all four datasets (Table 2).

Participant pairing confirms the ranking: all 48 SOAMNet-minus-comparator means are positive for each metric, and 46/48 bootstrap intervals exclude zero. Against ATCNet, paired BA gains are 3.15 [1.64, 4.75], 1.52 [0.69, 2.39], 4.42 [2.55, 6.81], and 3.97 [3.13, 4.81] pp across the four datasets (Supplementary Table S17).

Both empirical interventions have positive participant-bootstrap lower bounds on every dataset; $R_{\mathrm{core}}$ ranges from 18.61 to 40.95 pp, versus $-1.74$ to 4.67 pp among the audited hybrids (Table 3). Uniform replacement remains a separate analytic structural-null check.

Table 3: SOAMNet route audit: mean BA degradation (pp) [95% participant-bootstrap CI].

| Dataset | Uniform null | Cross-sample | Train mean | $R_{\mathrm{core}}$ |
|---|---|---|---|---|
| BCIC-IV-2b | 30.10 [21.01,38.38] | 31.37 [20.72,41.31] | 28.46 [19.61,36.88] | 28.46 [19.50,36.88] |
| BCIC-IV-2a | 44.44 [39.00,49.45] | 41.73 [36.36,46.82] | 40.95 [33.94,47.34] | 40.95 [33.94,46.28] |
| BNCI2015-001 | 35.53 [28.00,42.37] | 40.17 [30.84,48.59] | 33.81 [26.03,41.26] | 33.81 [26.03,41.26] |
| Cho2017 | 22.75 [19.27,26.29] | 39.04 [31.62,46.64] | 18.61 [15.56,21.75] | 18.61 [15.56,21.75] |

Table 4: BA change (pp) of trained controls relative to complete SOAMNet; negative values denote degradation.

| Variant | 2b | 2a | BNCI | Cho | Macro |
|---|---|---|---|---|---|
| Non-attention head | -3.51 | -5.22 | -4.45 | -3.98 | -4.29 |
| Local content only | -4.51 | -7.57 | -4.14 | -4.29 | -5.13 |
| Global carrier only | -2.38 | -0.95 | -2.29 | -2.45 | -2.02 |
| No temporal alignment | -2.42 | -1.12 | -3.31 | -2.46 | -2.33 |
| No orthogonal queries | -2.26 | -0.63 | -2.85 | -2.13 | -1.97 |
| Standard readout | -2.85 | -1.04 | -3.00 | -2.44 | -2.33 |
| Direct logit bypass | -2.65 | -0.68 | -2.97 | -2.63 | -2.23 |

### 7.3 Component controls

Every trained control reduces BA on every dataset, and all 28 participant-paired intervals exclude zero (Table 4). The non-attention head loses 3.51–5.22 pp; global-only and local-only routing lose 2.02 and 5.13 pp on macro average.

Frozen pathway diagnostics show that global class context establishes the principal sample-conditioned route, while local temporal content further improves classification (Supplementary Table S19).

## 8 Discussion

Across the audited hybrids, empirical route replacement usually changes BA only marginally even when zeroing is harmful. The diagnostics trace this gap to token-common Values, Value geometry that suppresses route changes, and residual, pooling, or temporal access outside the altered path. Convolution remains compatible with route dependence when it preserves token disagreement without creating a sufficient bypass.

SOAMNet shows that current-trial attention dependence can coexist with competitive BA and $\kappa$ across four datasets. Its controls distinguish complementary roles: global class context establishes the principal route, while temporally resolved content provides additional classification gains. The route is functionally used, although its weights are not direct localizations of neural sources.

Route-dependent decoders should preserve task-relevant token differences, expose route changes after Value projection, and require sufficient evidence to cross an auditable sample-conditioned path. Structured or physiologically motivated maps alone do not establish functional use (Wang et al., 2026; 2025; Siino et al., 2026; Muna et al., 2025), and predictive weights may reflect participant-specific correlates (Haufe et al., 2014). Our conclusions concern subject-specific MI decoding; transfer and distribution shift remain open.

## 9 Conclusion

Frozen-checkpoint audits show that hybrid MI decoders often lose or bypass attention-route information. SOAMNet establishes sample-conditioned route dependence while improving classification across four datasets, demonstrating how functional audits can guide route-dependent decoder design.

### Ethics statement

This study analyzes previously released, de-identified EEG datasets and involves no new participant recruitment, data collection, or intervention. Dataset provenance is documented in Section 6.

### Reproducibility statement

The data splits, preprocessing, model configurations, optimization budgets, checkpoint-selection rules, intervention definitions, statistical procedures, and computational profiles are documented in Sections 6 and G. The appendix also reports seed-level effects, intervention-integrity checks, and the complete quantitative matrices underlying the main figures and tables.

### AI use statement

Generative AI tools were used to aid or polish the writing of this paper. The authors reviewed all AI-assisted revisions and take full responsibility for the paper's content.

## References

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197, 2020. doi: 10.18653/v1/2020.acl-main.385.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, pp. 9505–9515, 2018.

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, 2023. doi: 10.1109/TII.2022.3197419.

Hamdi Altaheri, Fakhri Karray, and Amir-Hossein Karimi. Temporal convolutional transformer for EEG based motor imagery decoding. *Scientific Reports*, 15(1), 2025. doi: 10.1038/s41598-025-16219-7.

Kai Keng Ang, Zheng Yang Chin, Haihong Zhang, and Cuntai Guan. Filter bank common spatial pattern (FBCSP) in brain-computer interface. In *2008 IEEE International Joint Conference on Neural Networks*, pp. 2390–2397, 2008. doi: 10.1109/IJCNN.2008.4634130.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. In *International Conference on Learning Representations*, 2018.

Benjamin Blankertz, Ryota Tomioka, Steven Lemm, Motoaki Kawanabe, and Klaus-Robert Müller. Optimizing spatial filters for robust EEG single-trial analysis. *IEEE Signal Processing Magazine*, 25(1):41–56, 2008. doi: 10.1109/MSP.2008.4408441.

Gino Brunner, Yang Liu, Damián Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. On identifiability in transformers. In *International Conference on Learning Representations*, 2020.

Xiangrui Cai, Shaocheng Ma, Lei Cao, Jie Li, Tianyu Liu, and Yilin Dong. Fusion of multiscale features via centralized sparse-attention network for EEG decoding. *arXiv preprint arXiv:2512.18689*, 2025. doi: 10.48550/arXiv.2512.18689.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 782–791, 2021.

Hohyun Cho, Minkyu Ahn, Sangtae Ahn, Moonyoung Kwon, and Sung Chan Jun. EEG datasets for motor imagery brain–computer interface. *GigaScience*, 6(7), 2017. doi: 10.1093/gigascience/gix034.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104.

Josef Faller, Carmen Vidaurre, Teodoro Solis-Escalante, Christa Neuper, and Reinhold Scherer. Autocalibration and recurrent adaptation: Towards a plug and play online ERD-BCI. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 20(3):313–319, 2012. doi: 10.1109/TNSRE.2012.2189584.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Stefan Haufe, Frank Meinecke, Kai Görgen, Sven Dähne, John-Dylan Haynes, Benjamin Blankertz, and Felix Bießmann. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, 87:96–110, 2014. doi: 10.1016/j.neuroimage.2013.10.067.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, pp. 9737–9748, 2019.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3543–3556, 2019. doi: 10.18653/v1/N19-1357.

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.

Zida Liang, Jiayi Zhu, and Weiqiang Sun. Why attention fails: The degeneration of transformers into MLPs in time series forecasting. *arXiv preprint arXiv:2509.20942*, 2025.

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.

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

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, volume 32, pp. 14014–14024, 2019.

Ummay Maria Muna, Md. Mehedi Hasan Shawon, Md Jobayer, Sumaiya Akter, and Saifur Rahman Sabuj. SSTAF: Spatial-spectral-temporal attention fusion transformer for motor imagery classification. *IEEE Access*, 13:158855–158869, 2025. doi: 10.1109/ACCESS.2025.3605265.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

G. Pfurtscheller and F. H. Lopes da Silva. Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clinical Neurophysiology*, 110(11):1842–1857, 1999. doi: 10.1016/S1388-2457(99)00141-8.

Herbert Ramoser, Joachim Müller-Gerking, and Gert Pfurtscheller. Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE Transactions on Rehabilitation Engineering*, 8 (4):441–446, 2000. doi: 10.1109/86.895946.

Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H. Falk, and Jocelyn Faubert. Deep learning-based electroencephalography analysis: a systematic review. *Journal of Neural Engineering*, 16(5):051001, 2019. doi: 10.1088/1741-2552/ab260c.

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.

Sofia Serrano and Noah A. Smith. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2931–2951, 2019. doi: 10.18653/v1/P19-1282.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, volume 70, pp. 3145–3153, 2017.

Marco Siino, Giuseppe Bonomo, Rosario Sorbello, and Ilenia Tinnirello. RatioWaveNet: A learnable wavelet-based framework for robust and interpretable electroencephalogram motor imagery classification. *Array*, 30:100961, 2026. doi: 10.1016/j.array.2026.100961.

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, 2023. doi: 10.1109/TNSRE.2022.3230250.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, volume 70, pp. 3319–3328, 2017.

Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J. Miller, Gernot R. Müller-Putz, Guido Nolte, Gert Pfurtscheller, Hubert Preissl, Gerwin Schalk, Alois Schlögl, Carmen Vidaurre, Stephan Waldert, and Benjamin Blankertz. Review of the BCI competition IV. *Frontiers in Neuroscience*, 6, 2012. doi: 10.3389/fnins.2012.00055.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5797–5808, 2019. doi: 10.18653/v1/P19-1580.

Ziwei Wang, Siyang Li, Xiaoqing Chen, and Dongrui Wu. MVCNet: Multi-view contrastive network for motor imagery classification. *Knowledge-Based Systems*, 328:114205, 2025. doi: 10.1016/j.knosys.2025.114205.

Ziwei Wang, Hongbin Wang, Tianwang Jia, Xingyi He, Siyang Li, and Dongrui Wu. DBConformer: Dual-branch convolutional transformer for EEG decoding. *IEEE Journal of Biomedical and Health Informatics*, 30(5):4134–4147, 2026. doi: 10.1109/JBHI.2025.3622725.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 11–20, 2019. doi: 10.18653/v1/D19-1002.

Wei Zhao, Xiaolu Jiang, Baocan Zhang, Shixiao Xiao, and Sujun Weng. CTNet: a convolutional transformer network for EEG-based motor imagery classification. *Scientific Reports*, 14(1), 2024. doi: 10.1038/s41598-024-71118-7.

## A Implementation and training details

The local evidence extractor follows the configuration described in the main paper. For the three binary datasets, each channel is standardized over time within a trial and blended with the original signal,

$$\widetilde{X} = 0.75X + 0.25\frac{X - \mu_t(X)}{\max(\sigma_t(X), 0.25)}. \tag{16}$$

BCIC-IV-2a retains its original input scale. The nonlinear Key has width 64. The correction head uses hidden width $\max(24, 8C)$ and dropout 0.20, whereas the local Value adapter uses hidden width 64 and dropout 0.15. Class Queries are initialized orthogonally, and temporal carriers use distinct zero-mean sine or cosine harmonics.

Each SOAMNet instance is trained end to end as a single model under the corresponding seed; the evidence extractor and routing path are optimized jointly. AdamW uses a learning rate of $10^{-3}$, weight decay of $10^{-4}$, gradient clipping at 1.0, and cosine decay. BCIC-IV-2b is capped at 120 epochs with patience 35; the other datasets are capped at 100 epochs with patience 30. For BCIC-IV-2b and BNCI2015-001, the scalar $\beta$ in the activity-gated readout is selected per participant from $[-10, 10]$ at intervals of 0.02; the other datasets fix $\beta = 0$. Checkpoints and the two nonzero operating-point calibrations are selected by clean BA on the validation split.

## B Intervention definitions

All seven interventions are applied at inference time to the selected checkpoint; no model is retrained. At every audited target, the intervention changes only the attention weights and retains the current trial's Values and all model parameters. For SOAMNet, the audited target is the final class-conditioned route $a_{ct}$ in Eq. (13). `zero` sets the weights to zero, and `uniform` assigns $1/T$ to every temporal position. `canonical` uses the identity matrix for square self-attention and a fixed topology-specific one-hot anchor otherwise; SOAMNet uses the central temporal window. `shuffle` applies a fixed nontrivial permutation to the attention columns, whereas `reverse` reverses their order; neither operation reorders the Values.

`cross-sample` is a deterministic maximum-dissimilarity cyclic stress test. For the attention tensor supplied to each audited target, we average all axes except the sample and key-position axes to form one route signature per trial, evaluate every nonzero cyclic shift of the sample axis, and select the shift with the largest mean absolute signature distance. The resulting assignment is bijective and has no self-donor. Candidate donors come only from the same evaluation split; labels, logits, predictions, BA, and training or validation examples are not used to select the shift. The same rule is applied separately at every audited target, all of which are changed in the same forward pass, while each target retains the current trial's Values. `training-set mean` replaces each target with its attention mean estimated only from the training split. All audited attention targets in a checkpoint are changed jointly.

For SOAMNet, every replaced attention row is subsequently centered by its temporal mean. Zero and uniform therefore alter the weight tensor differently but produce the same analytic-null readout; neither intervention bypasses the readout or directly overwrites its output.

The intervention effect is clean BA minus intervened BA. Cross-sample and training-set-mean replacement form the primary intersection–union test: joint dependence is supported only when both participant-bootstrap 95% intervals have positive lower bounds. Their weaker mean effect is reported as $R_{\mathrm{core}}$. Uniform replacement is an analytic-null check for SOAMNet and a fixed-route control for the audited baselines; it does not enter the primary inference. Zero, uniform, canonical, shuffle, and reverse complete the descriptive seven-intervention matrix, while synchronized permutation is reported separately as an implementation control.

For a permutation matrix $P$, the implementation control uses the exact identity

$$(AP^{\top})(PV) = AV. \tag{17}$$

A prediction change under Eq. (17) would reveal an incorrect intervention or a hidden tensor mismatch.

Table S1: Functional attention audit. Each intervention is applied to a frozen checkpoint without retraining; only the audited attention weights change, while the current trial's Value tensors and all model parameters remain fixed.

| Property | Intervention | Question |
|---|---|---|
| Routed output | zero | Does a non-zero routed representation contribute? |
| Mixing structure | uniform, canonical | Does the learned mixing pattern matter relative to fixed routes? |
| Token alignment | shuffle, reverse keys | Does the learned route use temporal correspondence? |
| Sample adaptivity | cross-sample, training-set mean (core) | Does the route depend on the current trial? |
| Implementation control | synchronized permutation | Does a theoretically invariant transformation preserve $AV$ and predictions? |

## C BCIC-IV-2A FUNCTIONAL-AUDIT REPLICATION

We apply the same frozen-checkpoint protocol to the four-class BCIC-IV-2a task. Figure S1 reports the complete seven-intervention matrix for all eight model families, and Table S2 pairs clean performance with the structural uniform control and the two empirical replacement effects.

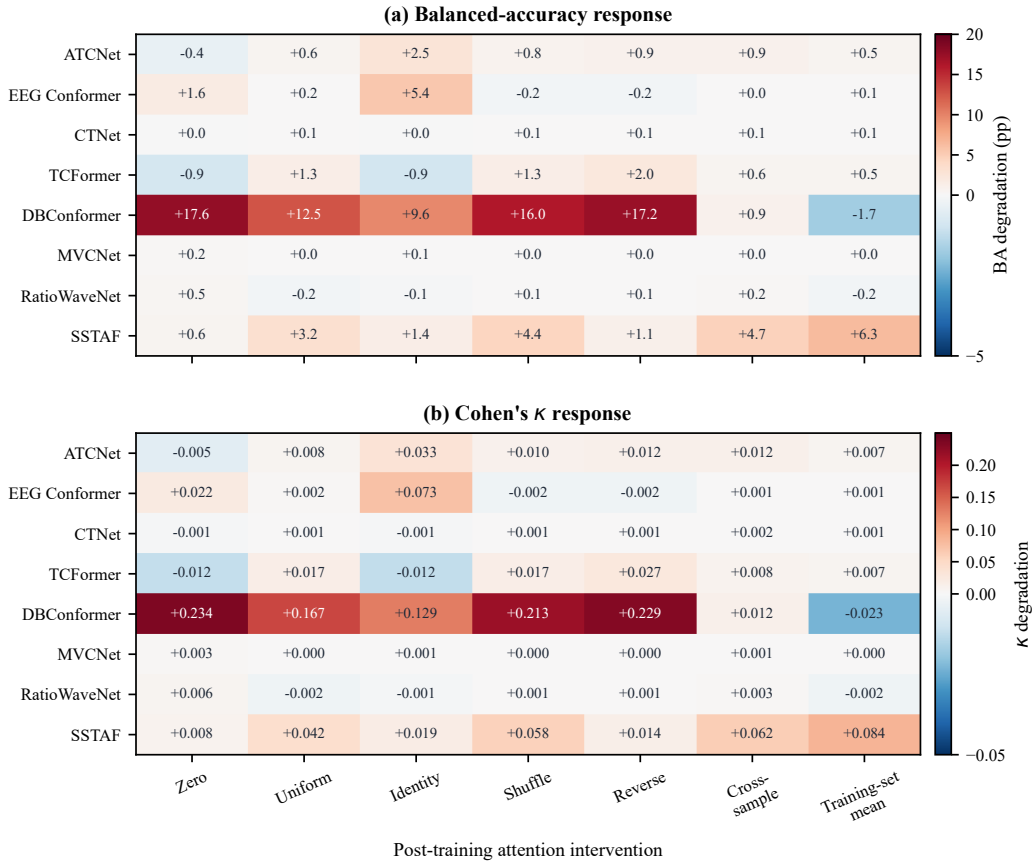



Figure S1: Seven-intervention frozen-checkpoint audit on BCIC-IV-2a at seed 2026. The upper panel reports mean BA degradation in percentage points and the lower panel reports Cohen's $\kappa$ degradation. Positive entries denote worse performance after replacement.

Table S2: Clean performance and route-replacement effects on BCIC-IV-2a. BA is a percentage; U, X, and M give mean BA degradation with participant-bootstrap 95% intervals under uniform, cross-sample, and training-set-mean replacement. U is a fixed-route control. $R_{\mathrm{core}}$ is the weaker mean of X and M. Only SSTAF has positive lower bounds for both empirical intervals on this dataset.

| Model | Clean BA | Clean $\kappa$ | U | X | M | $R_{\mathrm{core}}$ |
|---|---|---|---|---|---|---|
| ATCNet | 68.67 | 0.582 | 0.58 $[-0.93, 1.77]$ | 0.93 $[-0.39, 2.08]$ | 0.54 $[-0.77, 1.66]$ | 0.54 |
| EEG Conformer | 51.62 | 0.355 | 0.15 $[-0.58, 0.77]$ | 0.04 $[-0.66, 0.62]$ | 0.08 $[-0.62, 0.69]$ | 0.04 |
| CTNet | 65.55 | 0.541 | 0.08 $[-0.08, 0.27]$ | 0.12 $[0.00, 0.35]$ | 0.08 $[0.00, 0.19]$ | 0.08 |
| TCFormer | 67.17 | 0.562 | 1.31 $[0.04, 2.97]$ | 0.62 $[-0.35, 1.66]$ | 0.54 $[-0.54, 1.93]$ | 0.54 |
| DBConformer | 47.88 | 0.305 | 12.50 $[4.32, 21.91]$ | 0.93 $[-1.08, 2.89]$ | $-1.74$ $[-3.97, 0.15]$ | $-1.74$ |
| MVCNet | 45.91 | 0.279 | 0.00 $[0.00, 0.00]$ | 0.04 $[0.00, 0.12]$ | 0.00 $[0.00, 0.00]$ | 0.00 |
| RatioWaveNet | 67.40 | 0.565 | $-0.15$ $[-0.42, 0.19]$ | 0.23 $[0.00, 0.46]$ | $-0.15$ $[-0.42, 0.19]$ | $-0.15$ |
| SSTAF | 36.03 | 0.147 | 3.16 $[-0.04, 6.17]$ | 4.67 $[0.77, 8.76]$ | 6.33 $[1.93, 10.73]$ | 4.67 |

DBConformer, MVCNet, RatioWaveNet, and SSTAF are trained separately for each participant with seed 2026 and no more than 300 epochs. DBConformer uses Adam (batch 32, learning rate $10^{-4}$) for 200 epochs. MVCNet uses Adam (batch 64, learning rate $10^{-3}$): the component checkpoint is trained for 100 epochs, after which the fused model is optimized for at most 100 epochs, with a 20-epoch minimum and patience of 20. RatioWaveNet uses AdamW (batch 64, learning rate $10^{-3}$, weight decay $10^{-3}$, 20-epoch minimum, patience 20). SSTAF uses AdamW with batch 16, the same learning rate and weight decay, an 8–30 Hz passband, common-average reference, and normalization statistics estimated only from the training split. Checkpoints are selected by validation BA, with validation loss breaking ties. The frozen-checkpoint audit uses all 288 trials in each participant's official evaluation session.

The $R_{\mathrm{core}}$ values range from $-1.74$ to 4.67 pp. SSTAF has positive lower bounds for both empirical replacements on BCIC-IV-2a, but not on BCIC-IV-2b; none of the eight model families therefore shows jointly supported functional reliance across both datasets.

The mean attention maps in Figure S2 provide a complementary morphological view. EEG Conformer and CTNet approach uniform mixing, whereas ATCNet and TCFormer retain limited average structure. Nevertheless, the class-conditioned maps are nearly collinear in all four models. These summaries describe map morphology but do not determine functional reliance.

## D MODEL-SPECIFIC INTERVENTION DETAILS

DBConformer, MVCNet, RatioWaveNet, and SSTAF were trained independently on the common BCIC-IV-2b split with seed 2026. Table S3 records the optimization budget and the number of QK targets intervened jointly. Checkpoints were selected by validation BA, with validation loss breaking ties, before the frozen-checkpoint audit was applied.

Table S4 reports the paired BA/$\kappa$ audit for these models. All entries are nine-participant means on the held-out BCIC-IV-2b test sessions at seed 2026. Each intervention jointly replaces every QK-softmax route in a model while retaining the current Value tensor at every target.

DBConformer and SSTAF also contain vector-valued attention gates, for which an identity-matrix intervention is undefined. Table S5 therefore reports the six compatible interventions separately. Zeroing SSTAF's spectral and spatial gates reduces BA by 11.03 and 24.93 pp on BCIC-IV-2a and BCIC-IV-2b, whereas cross-sample and training-set-mean effects remain between $-0.12$ and

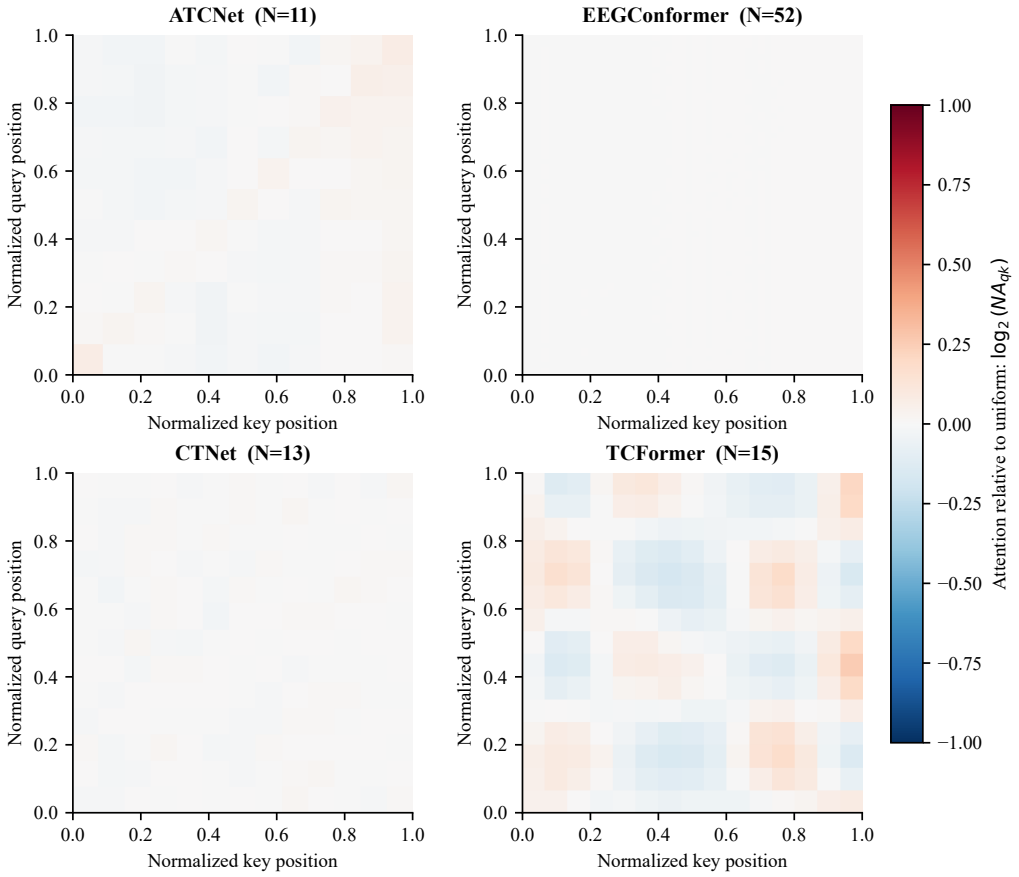



Figure S2: Test-set mean attention matrices for ATCNet, EEG Conformer, CTNet, and TCFormer on BCIC-IV-2b, averaged across participants, attention heads, and layers. Color gives the log-ratio relative to uniform attention. Map morphology is reported separately from the frozen-checkpoint functional audit.

Table S3: Training details for the four recent BCIC-IV-2b audits at seed 2026. "Cap/min/pat." gives maximum epochs, minimum epochs, and early-stopping patience. QK is the number of softmax route modules intervened jointly in every participant checkpoint.

| Model | QK | Batch | LR | WD | Cap/min/pat. |
|---|---|---|---|---|---|
| DBConformer | 4 | 32 | 1e−4 | 0 | 200/200/200 |
| MVCNet | 2 | 64 | 1e−3 | 0 | 100/20/20 |
| RatioWaveNet | 5 | 64 | 1e−3 | 1e−3 | 100/20/20 |
| SSTAF | 2 | 16 | 1e−3 | 1e−3 | 100/30/30 |

Table S4: Complete QK-softmax intervention audit for the four recent models. BA rows are degradation in percentage points; $\kappa$ rows are degradation on its native scale. Z: zero; U: uniform; I: identity; S: shuffle; R: reverse keys; X: cross-sample; M: training-set mean.

(a) Output removal and route-form controls

| Model | Metric | Z | U | I | S |
|---|---|---|---|---|---|
| DBConformer | BA (pp) | 13.74 | 0.19 | 0.24 | 0.45 |
| | $\kappa$ | 0.275 | 0.004 | 0.005 | 0.009 |
| MVCNet | BA (pp) | $-0.27$ | 0.03 | 0.05 | 0.04 |
| | $\kappa$ | $-0.005$ | 0.001 | 0.001 | 0.001 |
| RatioWaveNet | BA (pp) | 8.39 | 0.22 | 0.96 | 0.32 |
| | $\kappa$ | 0.166 | 0.004 | 0.019 | 0.006 |
| SSTAF | BA (pp) | 1.84 | 0.94 | 0.25 | 0.65 |
| | $\kappa$ | 0.037 | 0.019 | 0.005 | 0.013 |

(b) Nonzero route replacements

| Model | Metric | R | X | M |
|---|---|---|---|---|
| DBConformer | BA (pp) | 0.66 | 0.27 | 0.22 |
| | $\kappa$ | 0.013 | 0.005 | 0.004 |
| MVCNet | BA (pp) | $-0.01$ | 0.05 | 0.03 |
| | $\kappa$ | 0.000 | 0.001 | 0.001 |
| RatioWaveNet | BA (pp) | 0.27 | 0.14 | 0.27 |
| | $\kappa$ | 0.006 | 0.003 | 0.005 |
| SSTAF | BA (pp) | 0.76 | 0.89 | 0.93 |
| | $\kappa$ | 0.015 | 0.017 | 0.018 |

0.21 pp. The gate output contributes to prediction, but its trial-specific weighting remains replaceable.

Table S5: Model-specific vector-gate audit on BCIC-IV-2b. All native gates in a model are changed jointly. BA and $\kappa$ entries are test degradation; X is cross-sample replacement and M is training-set-mean replacement.

(a) Route-form controls

| Model | Metric | Zero | Uniform | Shuffle |
|---|---|---|---|---|
| DBConformer channel gate | BA (pp) | 0.17 | 0.00 | 0.00 |
| | $\kappa$ | 0.003 | 0.000 | 0.000 |
| SSTAF spectral/spatial gates | BA (pp) | 24.93 | 1.41 | 3.85 |
| | $\kappa$ | 0.497 | 0.028 | 0.078 |

(b) Nonzero gate replacements

| Model | Metric | Reverse | X | M |
|---|---|---|---|---|
| DBConformer channel gate | BA (pp) | 0.00 | 0.00 | 0.00 |
| | $\kappa$ | 0.000 | 0.000 | 0.000 |
| SSTAF spectral/spatial gates | BA (pp) | 1.11 | $-0.12$ | 0.21 |
| | $\kappa$ | 0.023 | $-0.002$ | 0.004 |

The integrity measurements in Table S6 establish that the small prediction effects do not reflect ineffective interventions. Uniform, cross-sample, and training-set-mean replacement change both the targeted weight tensors and the local $AV$ outputs in every architecture. Small changes in prediction therefore coexist with genuine changes inside the intervened routes.

Table S6: Internal intervention integrity for the four recent models, averaged across participants and all QK targets. $|\Delta A|$ is mean absolute weight change and $\|\Delta AV\|_2$ is the mean local routed-output change. U: uniform; X: cross-sample; M: training-set mean.

(a) Uniform replacement

| Model | $\lvert\Delta A\rvert$ | $\lVert\Delta AV\rVert_2$ |
|---|---|---|
| DBConformer | 0.00402 | 0.12521 |
| MVCNet | 0.09462 | 1.12139 |
| RatioWaveNet | 0.01552 | 0.80486 |
| SSTAF | 0.01593 | 3.98411 |

(b) Cross-sample

| Model | $\lvert\Delta A\rvert$ | $\lVert\Delta AV\rVert_2$ |
|---|---|---|
| DBConformer | 0.00482 | 0.15001 |
| MVCNet | 0.14000 | 1.57915 |
| RatioWaveNet | 0.02226 | 0.99505 |
| SSTAF | 0.02239 | 4.65726 |

(c) Training-set mean

| Model | $\lvert\Delta A\rvert$ | $\lVert\Delta AV\rVert_2$ |
|---|---|---|
| DBConformer | 0.00337 | 0.12150 |
| MVCNet | 0.09420 | 1.11158 |
| RatioWaveNet | 0.01552 | 0.79626 |
| SSTAF | 0.01598 | 3.71687 |

## E SEED-LEVEL SOAMNET INTERVENTION MATRIX

Table S7 reports every intervention effect for each dataset–seed pair. All core and auxiliary effects are positive throughout this descriptive matrix. The participant-level joint inference reported in the main paper, rather than treating seeds as independent observations, establishes the core route dependence; Cohen's $\kappa$ decreases throughout.

Table S7: Complete seed-level SOAMNet intervention audit. Panel (a) reports BA degradation in percentage points and panel (b) reports Cohen's $\kappa$ degradation. Z: zero; U: uniform; C: canonical one-hot center; S: shuffle; R: reverse; X: cross-sample; M: training-set mean. U is the structural null check; X and M are the empirical core tests.

(a) Balanced-accuracy degradation (pp)

| Dataset | Seed | Z | U | C | S | R | X | M |
|---|---|---|---|---|---|---|---|---|
| BCIC-IV-2b | 2026 | 29.81 | 29.81 | 30.29 | 23.80 | 26.99 | 33.93 | 30.25 |
| | 2036 | 30.26 | 30.26 | 30.26 | 18.07 | 24.54 | 28.86 | 27.39 |
| | 2046 | 30.23 | 30.23 | 26.85 | 27.36 | 25.34 | 31.32 | 27.75 |
| BCIC-IV-2a | 2026 | 45.45 | 45.45 | 45.25 | 17.55 | 34.38 | 43.21 | 41.55 |
| | 2036 | 44.33 | 44.33 | 44.29 | 17.21 | 33.06 | 42.71 | 42.98 |
| | 2046 | 43.56 | 43.56 | 43.44 | 17.86 | 30.71 | 39.27 | 38.31 |
| BNCI2015-001 | 2026 | 35.99 | 35.99 | 32.20 | 15.75 | 35.45 | 40.90 | 38.11 |
| | 2036 | 34.62 | 34.62 | 34.12 | 22.45 | 32.45 | 38.52 | 34.62 |
| | 2046 | 35.99 | 35.99 | 34.19 | 9.75 | 36.34 | 41.10 | 28.69 |
| Cho2017 | 2026 | 23.69 | 23.69 | 23.38 | 24.00 | 11.03 | 43.07 | 20.65 |
| | 2036 | 21.76 | 21.76 | 21.24 | 20.42 | 10.81 | 36.29 | 16.95 |
| | 2046 | 22.78 | 22.78 | 21.95 | 23.99 | 14.67 | 37.76 | 18.23 |

(b) Cohen's $\kappa$ degradation

| Dataset | Seed | Z | U | C | S | R | X | M |
|---|---|---|---|---|---|---|---|---|

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| BCIC-IV-2b | 2026 | 0.596 | 0.596 | 0.606 | 0.477 | 0.540 | 0.679 | 0.605 |
| | 2036 | 0.605 | 0.605 | 0.605 | 0.360 | 0.490 | 0.577 | 0.547 |
| | 2046 | 0.605 | 0.605 | 0.537 | 0.546 | 0.506 | 0.626 | 0.555 |
| BCIC-IV-2a | 2026 | 0.606 | 0.606 | 0.603 | 0.234 | 0.458 | 0.576 | 0.554 |
| | 2036 | 0.591 | 0.591 | 0.591 | 0.229 | 0.441 | 0.569 | 0.573 |
| | 2046 | 0.581 | 0.581 | 0.579 | 0.238 | 0.409 | 0.524 | 0.511 |
| BNCI2015-001 | 2026 | 0.720 | 0.720 | 0.644 | 0.315 | 0.709 | 0.818 | 0.762 |
| | 2036 | 0.692 | 0.692 | 0.682 | 0.449 | 0.649 | 0.770 | 0.692 |
| | 2046 | 0.720 | 0.720 | 0.684 | 0.195 | 0.727 | 0.822 | 0.574 |
| Cho2017 | 2026 | 0.474 | 0.474 | 0.468 | 0.480 | 0.221 | 0.861 | 0.413 |
| | 2036 | 0.435 | 0.435 | 0.424 | 0.408 | 0.216 | 0.725 | 0.339 |
| | 2046 | 0.456 | 0.456 | 0.439 | 0.480 | 0.293 | 0.756 | 0.365 |

# F DATA AND IMPLEMENTATION CONTROLS

The frozen four-dataset splits contain 24,240 clean trials: 5,267 from BCIC-IV-2b, 4,696 from BCIC-IV-2a, 4,795 from BNCI2015-001, and 9,482 from Cho2017. SHA-256 comparison of the preprocessed trials finds no duplicates across the training, validation, and test splits. Split indices are disjoint, all values are finite, and every split retains all classes.

As a negative control for the data pipeline, one fixed participant per dataset was trained after permuting only the training labels. Test BA was 47.7%, 28.8%, 50.0%, and 55.0%, respectively, close to the corresponding theoretical chance levels. Before any weight replacement, the instrumented implementation reproduces the original checkpoint logits within $10^{-6}$. For ATCNet, EEG Conformer, CTNet, and TCFormer, synchronized permutation of attention columns and Value rows preserves every prediction in all 72 audited dataset–model–participant cases; the largest absolute logit deviation is $1.91 \times 10^{-6}$. Separate integrity measurements for the 36 BCIC-IV-2b model–participant audits of DBConformer, MVCNet, RatioWaveNet, and SSTAF confirm nonzero changes in both the replaced QK weights and the corresponding local routed outputs (Table S6).

# G REPRODUCIBILITY DETAILS

## G.1 TRAINING-BUDGET AND SPLIT AUDIT

Table S8 reports the hyperparameter ranges used in the four-dataset neural comparison. Dataset size determines the epoch cap, minimum-epoch setting, and patience; no cap exceeds 300. Each of the eleven neural baselines is run with seeds 2026, 2036, and 2046, uses checkpoints selected by validation BA, and shares the same split hash for every dataset–seed–participant tuple. FBCSP is deterministic and therefore has no seed budget. Its BCIC-IV-2a configuration is selected on the fixed training-session validation split from eight candidates; the selected setting uses one-versus-rest CSP, the 1–4 s cue-relative window, two components per class and band, and the 24 highest-ranked mutual-information features. SOAMNet is optimized end to end as a single model; Table S9 reports the complete model.

Table S8: Optimization settings for the main neural comparison. Slash-separated entries list the distinct settings used across the four datasets. All models use AdamW with cosine decay, and checkpoints are selected by validation BA.

(a) Optimizer settings

| Method | Batch | LR | WD |
|---|---|---|---|
| EEGNet | 64 | 1e−3 | 0 |
| FBCNet | 64 | 1e−3 | 0 |
| EEG Conformer | 16/64 | 2e−4 | 0 |
| CTNet | 16/64 | 1e−3 | 1e−3 |
| TCFormer | 16/64 | 9e−4 | 1e−3 |
| ATCNet | 64 | 9e−4 | 1e−3 |
| DBConformer | 32 | 1e−3 | 0 |
| MVCNet | 64 | 1e−3/3e−4 | 0 |
| RatioWaveNet | 64 | 1e−3 | 1e−3 |
| SSTAF | 16 | 1e−3 | 1e−3 |
| EEG-CSANet | $16 \times 4$ | 9e−4 | 0 |
| SOAMNet | 64 | 1e−3 | 1e−4 |

(b) Training schedule

| Method | Epoch cap | Min. epoch | Patience |
|---|---|---|---|
| EEGNet | 200/300 | 30/100 | 30/50 |
| FBCNet | 200/300 | 30/100 | 30/50 |
| EEG Conformer | 100/200/300 | 20/30/100 | 20/30/50 |
| CTNet | 100/200/300 | 20/30/100 | 20/30/50 |
| TCFormer | 100/200/300 | 20/30/100 | 20/30/50 |
| ATCNet | 100/200/300 | 20/30/100 | 20/30/50 |
| DBConformer | 200/300 | 30/100 | 30/50 |
| MVCNet | 200/300 | 30/100 | 30/50 |
| RatioWaveNet | 200/300 | 30/100 | 30/50 |
| SSTAF | 100 | 30 | 30 |
| EEG-CSANet | 200/300 | 30/100 | 30/50 |
| SOAMNet | 100/120 | 30 | 30/35 |

For EEG-CSANet, the released multiscale convolution and centralized sparse-attention graph is retained. Only the spatial-kernel height and output-class count are parameterized to match each dataset; the temporal kernels and attention branches are unchanged. The $16 \times 4$ entry denotes physical batch 16 with four-step gradient accumulation, giving an effective batch of 64.

### G.2 End-to-end computational profile

Table S9 reports end-to-end inference cost under a common protocol for the eight audited hybrids, two attention-free neural controls, and SOAMNet. Measurements use a single RTX 2060 and BCIC-IV-2a inputs of shape $22 \times 1000$. Latency is the median of 100 batch-1 runs after 20 warm-ups; throughput and peak memory are measured with batch 64. SOAMNet contains 130.5 thousand parameters, all optimized jointly. Relative to ATCNet, total parameters increase by 14.7%, batch-1 latency by 28.4%, and peak batch-64 memory changes from 186.15 to 186.25 MiB.

All experiments record dataset identifiers, preprocessing hashes, participant split indices, model configurations, random seeds, and checkpoint-specific BA/$\kappa$ results. The reported SOAMNet evaluation spans four datasets and three seeds. Every post-training intervention reuses the corresponding single-model checkpoint without retraining.

Table S9: End-to-end inference cost on an RTX 2060 for BCIC-IV-2a inputs ($22 \times 1000$). Total parameters include every parameter in the inference graph; trainable parameters are reported separately.

| Method | Total (k) | Trainable (k) | Latency (ms) | Trials/s | Peak MiB |
|---|---|---|---|---|---|
| EEGNet | 4.0 | 4.0 | 0.37 | 11939 | 91.42 |
| FBCNet | 11.8 | 11.8 | 0.68 | 8633 | 272.93 |
| EEG Conformer | 789.6 | 789.6 | 2.73 | 7286 | 236.16 |
| CTNet | 152.7 | 152.7 | 3.06 | 11038 | 230.20 |
| TCFormer | 127.2 | 127.2 | 4.93 | 3179 | 1139.87 |
| ATCNet | 113.7 | 113.7 | 5.79 | 9772 | 186.15 |
| DBConformer | 93.8 | 93.8 | 4.01 | 11808 | 51.68 |
| MVCNet | 4114.7 | 4114.7 | 1.74 | 16758 | 79.45 |
| RatioWaveNet | 113.7 | 113.7 | 7.49 | 4783 | 192.45 |
| SSTAF | 201.4 | 201.4 | 2.17 | 24966 | 58.37 |
| SOAMNet | 130.5 | 130.5 | 7.44 | 7935 | 186.25 |

# H Extended quantitative results

## H.1 Complete BCIC-IV-2b functional-audit matrix

Table S10 provides the exact values underlying the functional-audit heatmap in the main paper. Its three panels report clean performance, all seven BA/$\kappa$ intervention effects, and participant-bootstrap inference for the structural uniform control and two empirical replacements.

Table S10: Complete BCIC-IV-2b audit under the common protocol at seed 2026. BA rows report clean percentages and intervention degradation in percentage points; $\kappa$ rows use its native scale. Z: zero; U: uniform; C: canonical fixed route (identity for square self-attention and a topology-specific one-hot anchor otherwise); S: shuffle; R: reverse keys; X: cross-sample; M: training-set mean.

(a) Clean performance and route-form controls

| Model | Metric | Clean | Z | U | C |
|---|---|---|---|---|---|
| ATCNet | BA | 75.68 | 11.19 | −0.28 | −0.64 |
| | $\kappa$ | 0.512 | 0.223 | −0.006 | −0.013 |
| EEG Conformer | BA | 71.27 | 20.81 | −0.04 | 0.46 |
| | $\kappa$ | 0.425 | 0.416 | −0.001 | 0.009 |
| CTNet | BA | 73.11 | 2.25 | 0.17 | 0.32 |
| | $\kappa$ | 0.460 | 0.045 | 0.003 | 0.006 |
| TCFormer | BA | 73.14 | 3.93 | −0.60 | 1.55 |
| | $\kappa$ | 0.461 | 0.078 | −0.012 | 0.031 |
| DBConformer | BA | 78.54 | 13.74 | 0.19 | 0.24 |
| | $\kappa$ | 0.570 | 0.275 | 0.004 | 0.005 |
| MVCNet | BA | 77.26 | −0.27 | 0.03 | 0.05 |
| | $\kappa$ | 0.544 | −0.005 | 0.001 | 0.001 |
| RatioWaveNet | BA | 75.24 | 8.39 | 0.22 | 0.96 |
| | $\kappa$ | 0.503 | 0.166 | 0.004 | 0.019 |
| SSTAF | BA | 74.93 | 1.84 | 0.94 | 0.25 |
| | $\kappa$ | 0.497 | 0.037 | 0.019 | 0.005 |

Table S10: Complete BCIC-IV-2b audit under the common protocol at seed 2026 (continued). BA rows report intervention degradation in percentage points; $\kappa$ rows use its native scale. S: shuffle; R: reverse keys; X: cross-sample; M: training-set mean.

(b) Nonzero route replacements

| Model | Metric | S | R | X | M |
|---|---|---|---|---|---|
| ATCNet | BA | −0.35 | 0.27 | 0.11 | −0.46 |
| | $\kappa$ | −0.007 | 0.005 | 0.002 | −0.009 |
| EEG Conformer | BA | −0.10 | −0.14 | −0.09 | −0.04 |
| | $\kappa$ | −0.002 | −0.003 | −0.002 | −0.001 |
| CTNet | BA | 0.02 | 0.18 | −0.18 | −0.18 |
| | $\kappa$ | 0.001 | 0.003 | −0.003 | −0.003 |
| TCFormer | BA | −0.55 | −0.39 | −0.20 | −0.51 |
| | $\kappa$ | −0.011 | −0.008 | −0.004 | −0.010 |
| DBConformer | BA | 0.45 | 0.66 | 0.27 | 0.22 |
| | $\kappa$ | 0.009 | 0.013 | 0.005 | 0.004 |
| MVCNet | BA | 0.04 | −0.01 | 0.05 | 0.03 |
| | $\kappa$ | 0.001 | 0.000 | 0.001 | 0.001 |
| RatioWaveNet | BA | 0.32 | 0.27 | 0.14 | 0.27 |
| | $\kappa$ | 0.006 | 0.006 | 0.003 | 0.005 |
| SSTAF | BA | 0.65 | 0.76 | 0.89 | 0.93 |
| | $\kappa$ | 0.013 | 0.015 | 0.017 | 0.018 |

Table S10: Complete BCIC-IV-2b audit under the common protocol at seed 2026 (continued). Cells give mean BA degradation in percentage points with participant-bootstrap 95% intervals below. U is a fixed-route control; X and M are the two empirical core replacements. $R_{\text{core}}$ is the weaker mean of X and M. No model has positive lower bounds for both empirical intervals.

(c) Joint inference and continuous core effect

| Model | U | X | M | $R_{\text{core}}$ |
|---|---|---|---|---|
| ATCNet | −0.28<br>[−1.06, 0.45] | 0.11<br>[−0.46, 0.69] | −0.46<br>[−1.17, 0.24] | −0.46 |
| EEG Conformer | −0.04<br>[−0.29, 0.17] | −0.09<br>[−0.28, 0.00] | −0.04<br>[−0.28, 0.17] | −0.09 |
| CTNet | 0.17<br>[−0.22, 0.71] | −0.18<br>[−0.51, 0.06] | −0.18<br>[−0.47, 0.00] | −0.18 |
| TCFormer | −0.60<br>[−1.54, 0.36] | −0.20<br>[−1.21, 0.80] | −0.51<br>[−1.55, 0.55] | −0.51 |
| DBConformer | 0.19<br>[−0.28, 0.67] | 0.27<br>[0.04, 0.50] | 0.22<br>[−0.05, 0.52] | 0.22 |
| MVCNet | 0.03<br>[−0.29, 0.32] | 0.05<br>[−0.45, 0.68] | 0.03<br>[−0.29, 0.32] | 0.03 |
| RatioWaveNet | 0.22<br>[−0.12, 0.56] | 0.14<br>[−0.27, 0.56] | 0.27<br>[−0.11, 0.67] | 0.14 |
| SSTAF | 0.94<br>[−0.21, 2.17] | 0.89<br>[−0.19, 2.13] | 0.93<br>[−0.26, 2.17] | 0.89 |

### H.2 COMPLETE INTERNAL MECHANISM DIAGNOSTICS

Tables S11 and S12 provide the complete numerical values summarized by the mechanism-diagnostic figure in the main paper. For every frozen checkpoint, linear probes are fitted on the training sessions, their regularization is selected on validation sessions, and test sessions are used only for evaluation. The consensus, disagreement, and attention-output features are respectively $\mu(V)$, $V_\perp$, and $AV$, flattened across heads and their remaining axes.

The leakage-safe consensus probe used in this analysis is

$$S_{\text{cons}} = \text{BA}(h_{\text{probe}}(\mu(V))) \,, \tag{18}$$

and is compared with probes fitted to the centered disagreement field $V_\perp$ and the routed output $AV$. For centered singular values $\{\sigma_j\}$, normalized effective rank is $\exp[-\sum_j p_j \log p_j]/\min(L-1, d_v)$ with $p_j = \sigma_j^2/\sum_k \sigma_k^2$. Consensus energy is $L\|\mu(V)\|_F^2/\|V\|_F^2$, and token correlation is the mean absolute off-diagonal cosine similarity between Value tokens. Prediction-flip rate is the fraction of held-out trials whose argmax class changes after cross-sample replacement; attention probe gain is the test BA of the $AV$ probe minus that of the consensus probe.

Table S11: Frozen-checkpoint representation diagnostics averaged across attention targets on BCIC-IV-2b. Probe entries are BA percentages. Gain is attention-output BA minus consensus BA in percentage points.

(a) Linear-probe BA and attention gain

| Model | Consensus | Disagreement | Attn. output | Gain |
|---|---|---|---|---|
| ATCNet | 74.31 | 55.91 | 74.27 | −0.04 |
| EEG Conformer | 72.31 | 51.84 | 71.66 | −0.65 |
| CTNet | 74.96 | 64.52 | 75.29 | 0.34 |
| TCFormer | 74.85 | 59.75 | 74.36 | −0.49 |
| DBConformer | 77.26 | 69.28 | 75.92 | −1.34 |
| MVCNet | 49.80 | 51.11 | 50.20 | 0.39 |
| RatioWaveNet | 75.32 | 53.42 | 74.48 | −0.83 |
| SSTAF | 74.66 | 59.72 | 74.51 | −0.15 |

(b) Value geometry

| Model | Eff. rank | Cons. energy | Token corr. |
|---|---|---|---|
| ATCNet | 0.506 | 0.490 | 0.564 |
| EEG Conformer | 0.689 | 0.953 | 0.961 |
| CTNet | 0.551 | 0.565 | 0.602 |
| TCFormer | 0.372 | 0.883 | 0.895 |
| DBConformer | 0.683 | 0.826 | 0.868 |
| MVCNet | 0.872 | 0.789 | 0.693 |
| RatioWaveNet | 0.579 | 0.322 | 0.430 |
| SSTAF | 0.457 | 0.262 | 0.447 |

Table S12: Route geometry and cross-sample effect propagation on BCIC-IV-2b. $H(A)$ is normalized attention entropy; $d(U)$ is mean distance from uniform. Relative changes are independently normalized by the corresponding clean tensor norms. The replacement uses the label-free maximally different cyclic pairing defined in Supplementary Section B.

| Model | $H(A)$ | $d(U)$ | $\Delta A$ | $\Delta AV$ | $\Delta$logit |
|---|---|---|---|---|---|
| ATCNet | 0.860 | 0.614 | 0.888 | 0.498 | 0.126 |
| EEG Conformer | 1.000 | 0.016 | 0.022 | 0.002 | 0.008 |
| CTNet | 0.995 | 0.130 | 0.119 | 0.048 | 0.007 |
| TCFormer | 0.940 | 0.477 | 0.451 | 0.090 | 0.069 |
| DBConformer | 0.999 | 0.037 | 0.041 | 0.021 | 0.021 |
| MVCNet | 0.967 | 0.229 | 0.344 | 0.142 | 0.013 |
| RatioWaveNet | 0.984 | 0.277 | 0.394 | 0.207 | 0.051 |
| SSTAF | 0.963 | 0.320 | 0.473 | 0.456 | 0.155 |

Table S13: Controls linking the proposed mechanisms to intervention response. Panel (a) reports participant-level Spearman associations across ATCNet, EEG Conformer, CTNet, TCFormer, and the controlled attention bottleneck, with model–participant cluster-bootstrap 95% intervals. Panel (b) applies equal-Frobenius-norm, row-sum-preserving perturbations to ATCNet: the null direction satisfies $\Delta A V = 0$, whereas the sensitive direction follows the leading left singular direction of centered Value.

(a) Representation geometry and prediction response

| Predictor | Response | $n$ | $\rho$ [95% CI] |
|---|---|---|---|
| Consensus energy | Prediction flips | 45 | $-0.761$ [$-0.883$,$-0.563$] |
| Token correlation | Prediction flips | 45 | $-0.719$ [$-0.835$,$-0.534$] |
| Attention probe gain | Prediction flips | 45 | $+0.428$ [$+0.106$,$+0.666$] |

(b) Matched Value-subspace perturbations

| Direction | $\Delta$BA [95% CI] (pp) | Flip rate | $\|\Delta A\|_F$ | $\|\Delta AV\|_F$ | $\|\Delta$logit$\|$ |
|---|---|---|---|---|---|
| Null space | 0.00 [0.00,0.00] | 0.000 | 0.479 | $4.41\text{e}{-7}$ | $5.43\text{e}{-8}$ |
| Value-sensitive | 0.43 [0.05,0.84] | 0.011 | 0.479 | 2.225 | 0.062 |

### H.3 Dataset roles and controlled factorial values

The first table separates the role of each dataset in the evidence chain; the second gives the exact residual-by-Value factorial values used to isolate bypass and consensus effects under a common task. Figure S3 reports the complementary route-by-TCN factorial.

Table S14: Datasets and their roles. Every SOAMNet result is reported over three seeds.

| Dataset | Classes | Subjects | Role |
|---|---|---|---|
| BCIC-IV-2b | 2 | 9 | Common functional audit, mechanism diagnosis, constructive-model evaluation |
| BCIC-IV-2a | 4 | 9 | Full functional-audit replication, constructive-model evaluation |
| BNCI2015-001 | 2 | 12 | Cross-session constructive-model evaluation |
| Cho2017 | 2 | 48 | Larger-population constructive-model evaluation |

Table S15: ATCNet residual-by-Value factorial on BCIC-IV-2b. Intervention columns report BA degradation in percentage points. The consensus probe reads the temporal mean of the Values presented to attention and is omitted when this mean is zero by construction.

(a) Factorial conditions and clean performance

| Condition | No res. | Center $V$ | BA | Cons. probe |
|---|---|---|---|---|
| Original ATCNet | no | no | 73.8 | 75.2 |
| No residual | yes | no | 74.4 | 75.6 |
| Centered Value | no | yes | 73.2 | — |
| Attention bottleneck | yes | yes | 75.4 | — |

(b) Attention-intervention BA degradation (pp)

| Condition | Uniform | Reverse | Cross-sample | Training-set mean |
|---|---|---|---|---|
| Original ATCNet | −0.0 | 0.2 | 0.1 | −0.1 |
| No residual | 0.2 | 0.5 | 0.4 | 0.2 |
| Centered Value | 0.1 | 1.8 | 1.3 | 0.3 |
| Attention bottleneck | 25.4 | 27.8 | 24.1 | 24.2 |

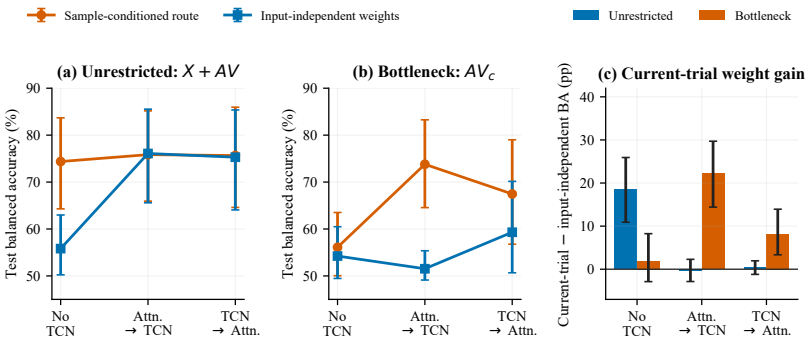



Figure S3: Current-trial versus input-independent attention weights crossed with TCN presence and module order on BCIC-IV-2b (seed 2026, nine participants). All conditions use the ATCNet frontend and matched routers. The no-TCN condition uses a parameter-matched pointwise MLP; the other conditions place the TCN before or after attention. The unrestricted topology permits downstream compensation, whereas the bottleneck requires evidence to pass through attention.

### H.4 TEMPORAL SCALE, CNN-TOKEN CONTROL, AND ERD PRESERVATION

As a positive architecture control, the CNN-token Transformer shows that the audit detects temporal attention dependence when it is present. A local multiscale CNN converts the EEG into a temporal token sequence processed by the same eight-layer Transformer, without a CNN classifier bypass or a post-attention TCN. It reaches 69.5% BA, compared with 67.1% and 62.8% for the pure point-wise and patch Transformers. Joint intervention across all eight layers reduces BA by 8.9 pp under uniform routes, 11.0 pp under reversed keys, 9.0 pp under cross-sample replacement, and 8.1 pp under training-set-mean replacement. Single-layer interventions localize the strongest dependence to the first layer.

A separate matched token-duration scan compares 4-ms point tokens with 25, 50, 100, and 200-ms patches using the same two-layer, width-64 Transformer and common split. Intervention windows are matched in physical duration by shifting each route by $\pm 200$ ms. Point tokens reach 66.6% BA and show a 13.1 pp response to this time-aligned shift. The 100 and 200 ms variants reach 65.2% and 64.5%, with shift effects of 8.7 and 8.8 pp. Held-out C3/C4 mu/beta summaries measure whether attention preserves event-related desynchronization: post/pre retention rises from 0.889 for point tokens to 0.899 for 200 ms patches, and post-attention class-contrast correlation rises from 0.197 to 0.499. No single scale dominates classification, intervention response, and ERD preservation (Fig. S4).

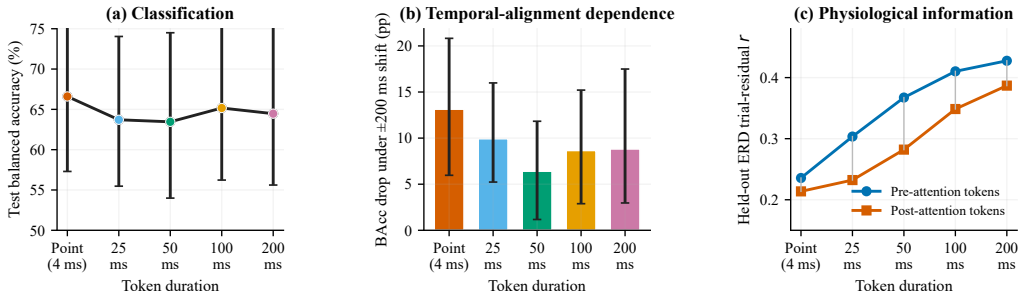



Figure S4: Point and patch tokenization on BCIC-IV-2b. The panels compare clean BA, time-aligned intervention response, and preservation of C3/C4 mu/beta ERD summaries. Token scale changes the balance among classification, route sensitivity, and ERD preservation.

### H.5 COMPLETE SOAMNET INTERVENTION MATRIX

Table S16: Seven-intervention audit of SOAMNet, averaged over three seeds. Each dataset spans paired BA and $\kappa$ rows. BA degradation is in percentage points; $\kappa$ degradation is on its native scale. U is the structural null check; asterisks mark the two empirical core tests.

(a) Output removal and route-form controls

| Dataset | Metric | Zero | Uniform | Canonical | Shuffle |
|---|---|---|---|---|---|
| BCIC-IV-2b | BA (pp) | 30.10 | 30.10 | 29.13 | 23.07 |
| | $\kappa$ | 0.602 | 0.602 | 0.583 | 0.461 |
| BCIC-IV-2a | BA (pp) | 44.44 | 44.44 | 44.33 | 17.54 |
| | $\kappa$ | 0.593 | 0.593 | 0.591 | 0.234 |
| BNCI2015-001 | BA (pp) | 35.53 | 35.53 | 33.50 | 15.98 |
| | $\kappa$ | 0.711 | 0.711 | 0.670 | 0.320 |
| Cho2017 | BA (pp) | 22.75 | 22.75 | 22.19 | 22.80 |
| | $\kappa$ | 0.455 | 0.455 | 0.444 | 0.456 |

(b) Nonzero route replacements

| Dataset | Metric | Reverse | Cross-sample* | Training-set mean* |
|---|---|---|---|---|
| BCIC-IV-2b | BA (pp) | 25.62 | 31.37 | 28.46 |
| | $\kappa$ | 0.512 | 0.627 | 0.569 |
| BCIC-IV-2a | BA (pp) | 32.72 | 41.73 | 40.95 |
| | $\kappa$ | 0.436 | 0.556 | 0.546 |
| BNCI2015-001 | BA (pp) | 34.75 | 40.17 | 33.81 |
| | $\kappa$ | 0.695 | 0.803 | 0.676 |
| Cho2017 | BA (pp) | 12.17 | 39.04 | 18.61 |
| | $\kappa$ | 0.243 | 0.781 | 0.372 |

### H.6 PARTICIPANT-PAIRED COMPARISON FOR TABLE 2

Table S17 reports every participant-paired comparison underlying the uncertainty summary in the main paper. For each neural method, participant-level BA and $\kappa$ are first averaged over seeds 2026, 2036, and 2046; FBCSP contributes its single deterministic value. Differences are SOAMNet minus the named comparator. Confidence intervals are percentile intervals from 100,000 participant-level bootstrap resamples. Both tests are two-sided, and Holm correction is applied across the 12 comparators separately within each dataset, metric, and test family.

Table S17: Participant-paired differences for the four-dataset comparison in Table 2. Each row stacks BA (upper) and $\kappa$ (lower); BA differences are percentage points and $\kappa$ differences use its native scale. $p_{t,H}$ and $p_{W,H}$ are Holm-adjusted two-sided paired $t$-test and Wilcoxon signed-rank $p$-values. Positive differences favor SOAMNet.

| Dataset | Comparator | $n$ | Difference [95% CI] (BA / $\kappa$) | $p_{t,H}$ (BA / $\kappa$) | $p_{W,H}$ (BA / $\kappa$) |
|---|---|---|---|---|---|
| BCIC-IV-2b | FBCSP | 9 | BA: +5.77 [+3.03, +8.26] | 0.018 | 0.047 |
| | | | $\kappa$: +0.116 [+0.063, +0.165] | 0.015 | 0.047 |
| BCIC-IV-2b | EEGNet | 9 | BA: +8.89 [+5.87, +12.48] | 0.009 | 0.047 |
| | | | $\kappa$: +0.179 [+0.119, +0.251] | 0.009 | 0.047 |
| BCIC-IV-2b | FBCNet | 9 | BA: +6.81 [+4.61, +8.90] | 0.004 | 0.047 |
| | | | $\kappa$: +0.137 [+0.093, +0.178] | 0.004 | 0.047 |
| BCIC-IV-2b | EEG Conformer | 9 | BA: +5.92 [+2.83, +9.13] | 0.018 | 0.047 |
| | | | $\kappa$: +0.119 [+0.057, +0.184] | 0.018 | 0.047 |

Continued on next page

Table S17 (continued)

| Dataset | Comparator | $n$ | Difference [95% CI] (BA / $\kappa$) | $p_{t,H}$ (BA / $\kappa$) | $p_{W,H}$ (BA / $\kappa$) |
|---|---|---|---|---|---|
| BCIC-IV-2b | CTNet | 9 | BA: +7.05 [+4.01, +10.43] | 0.018 | 0.047 |
| | | | $\kappa$: +0.142 [+0.082, +0.209] | 0.015 | 0.047 |
| BCIC-IV-2b | TCFormer | 9 | BA: +5.12 [+2.87, +7.17] | 0.015 | 0.047 |
| | | | $\kappa$: +0.104 [+0.059, +0.145] | 0.015 | 0.047 |
| BCIC-IV-2b | ATCNet | 9 | BA: +3.15 [+1.64, +4.75] | 0.018 | 0.047 |
| | | | $\kappa$: +0.064 [+0.033, +0.097] | 0.018 | 0.047 |
| BCIC-IV-2b | DBConformer | 9 | BA: +2.04 [-0.53, +4.71] | 0.190 | 0.301 |
| | | | $\kappa$: +0.042 [-0.010, +0.097] | 0.185 | 0.301 |
| BCIC-IV-2b | MVCNet | 9 | BA: +3.42 [+2.04, +4.87] | 0.015 | 0.047 |
| | | | $\kappa$: +0.070 [+0.041, +0.099] | 0.015 | 0.047 |
| BCIC-IV-2b | RatioWaveNet | 9 | BA: +4.93 [+3.51, +6.09] | 0.001 | 0.047 |
| | | | $\kappa$: +0.100 [+0.071, +0.124] | 0.001 | 0.047 |
| BCIC-IV-2b | SSTAF | 9 | BA: +7.18 [+4.85, +10.07] | 0.009 | 0.047 |
| | | | $\kappa$: +0.145 [+0.099, +0.202] | 0.008 | 0.047 |
| BCIC-IV-2b | EEG-CSANet | 9 | BA: +4.34 [+3.10, +5.63] | 0.002 | 0.047 |
| | | | $\kappa$: +0.088 [+0.062, +0.115] | 0.003 | 0.047 |
| BCIC-IV-2a | FBCSP | 9 | BA: +26.50 [+22.17, +31.61] | $< 0.001$ | 0.047 |
| | | | $\kappa$: +0.353 [+0.295, +0.421] | $< 0.001$ | 0.047 |
| BCIC-IV-2a | EEGNet | 9 | BA: +12.17 [+9.34, +14.84] | $< 0.001$ | 0.047 |
| | | | $\kappa$: +0.162 [+0.124, +0.198] | $< 0.001$ | 0.047 |
| BCIC-IV-2a | FBCNet | 9 | BA: +13.54 [+10.34, +16.78] | $< 0.001$ | 0.047 |
| | | | $\kappa$: +0.181 [+0.138, +0.224] | $< 0.001$ | 0.047 |
| BCIC-IV-2a | EEG Conformer | 9 | BA: +18.79 [+13.81, +23.68] | $< 0.001$ | 0.047 |
| | | | $\kappa$: +0.251 [+0.184, +0.316] | $< 0.001$ | 0.047 |
| BCIC-IV-2a | CTNet | 9 | BA: +4.58 [+2.49, +6.60] | 0.013 | 0.047 |
| | | | $\kappa$: +0.061 [+0.033, +0.088] | 0.013 | 0.047 |
| BCIC-IV-2a | TCFormer | 9 | BA: +2.37 [-0.37, +5.03] | 0.145 | 0.148 |
| | | | $\kappa$: +0.032 [-0.005, +0.067] | 0.145 | 0.148 |
| BCIC-IV-2a | ATCNet | 9 | BA: +1.52 [+0.69, +2.39] | 0.032 | 0.059 |
| | | | $\kappa$: +0.020 [+0.009, +0.032] | 0.032 | 0.059 |
| BCIC-IV-2a | DBConformer | 9 | BA: +22.96 [+16.62, +29.04] | $< 0.001$ | 0.047 |
| | | | $\kappa$: +0.306 [+0.221, +0.388] | $< 0.001$ | 0.047 |
| BCIC-IV-2a | MVCNet | 9 | BA: +18.38 [+13.70, +22.25] | $< 0.001$ | 0.047 |
| | | | $\kappa$: +0.245 [+0.183, +0.297] | $< 0.001$ | 0.047 |
| BCIC-IV-2a | RatioWaveNet | 9 | BA: +2.60 [+0.27, +4.90] | 0.144 | 0.148 |
| | | | $\kappa$: +0.035 [+0.004, +0.065] | 0.144 | 0.148 |
| BCIC-IV-2a | SSTAF | 9 | BA: +32.45 [+29.22, +35.98] | $< 0.001$ | 0.047 |
| | | | $\kappa$: +0.433 [+0.389, +0.480] | $< 0.001$ | 0.047 |
| BCIC-IV-2a | EEG-CSANet | 9 | BA: +20.20 [+16.62, +23.73] | $< 0.001$ | 0.047 |
| | | | $\kappa$: +0.269 [+0.222, +0.316] | $< 0.001$ | 0.047 |
| BNCI2015-001 | FBCSP | 12 | BA: +9.34 [+5.54, +13.42] | 0.005 | 0.006 |
| | | | $\kappa$: +0.187 [+0.110, +0.269] | 0.005 | 0.006 |
| BNCI2015-001 | EEGNet | 12 | BA: +12.68 [+7.05, +18.47] | 0.006 | 0.006 |
| | | | $\kappa$: +0.253 [+0.140, +0.370] | 0.006 | 0.006 |
| BNCI2015-001 | FBCNet | 12 | BA: +9.05 [+5.38, +12.90] | 0.005 | 0.006 |
| | | | $\kappa$: +0.181 [+0.107, +0.258] | 0.005 | 0.006 |
| BNCI2015-001 | EEG Conformer | 12 | BA: +9.03 [+6.19, +12.17] | 0.002 | 0.006 |
| | | | $\kappa$: +0.180 [+0.124, +0.243] | 0.002 | 0.006 |
| BNCI2015-001 | CTNet | 12 | BA: +11.15 [+7.28, +15.02] | 0.002 | 0.006 |
| | | | $\kappa$: +0.223 [+0.146, +0.300] | 0.002 | 0.006 |
| BNCI2015-001 | TCFormer | 12 | BA: +8.75 [+5.50, +12.13] | 0.004 | 0.006 |
| | | | $\kappa$: +0.175 [+0.110, +0.242] | 0.004 | 0.006 |
| BNCI2015-001 | ATCNet | 12 | BA: +4.42 [+2.55, +6.81] | 0.006 | 0.006 |
| | | | $\kappa$: +0.088 [+0.051, +0.137] | 0.006 | 0.006 |

Continued on next page

Table S17 (continued)

| Dataset | Comparator | $n$ | Difference [95% CI] (BA / $\kappa$) | $p_{t,H}$ (BA / $\kappa$) | $p_{W,H}$ (BA / $\kappa$) |
|---|---|---|---|---|---|
| | | | BA: +6.36 [+3.54, +9.44] | 0.006 | 0.006 |
| BNCI2015-001 | DBConformer | 12 | $\kappa$: +0.127 [+0.071, +0.188] | 0.006 | 0.006 |
| | | | BA: +7.96 [+4.40, +12.74] | 0.006 | 0.006 |
| BNCI2015-001 | MVCNet | 12 | $\kappa$: +0.159 [+0.088, +0.254] | 0.006 | 0.006 |
| | | | BA: +7.74 [+4.89, +10.71] | 0.004 | 0.006 |
| BNCI2015-001 | RatioWaveNet | 12 | $\kappa$: +0.155 [+0.098, +0.215] | 0.004 | 0.006 |
| | | | BA: +11.68 [+8.06, +15.12] | $< 0.001$ | 0.006 |
| BNCI2015-001 | SSTAF | 12 | $\kappa$: +0.234 [+0.162, +0.303] | $< 0.001$ | 0.006 |
| | | | BA: +6.69 [+4.05, +9.47] | 0.005 | 0.006 |
| BNCI2015-001 | EEG-CSANet | 12 | $\kappa$: +0.134 [+0.081, +0.189] | 0.005 | 0.006 |
| | | | BA: +11.65 [+9.03, +14.37] | $< 0.001$ | $< 0.001$ |
| Cho2017 | FBCSP | 48 | $\kappa$: +0.233 [+0.180, +0.287] | $< 0.001$ | $< 0.001$ |
| | | | BA: +15.38 [+12.41, +18.52] | $< 0.001$ | $< 0.001$ |
| Cho2017 | EEGNet | 48 | $\kappa$: +0.308 [+0.248, +0.370] | $< 0.001$ | $< 0.001$ |
| | | | BA: +5.37 [+3.21, +7.57] | $< 0.001$ | $< 0.001$ |
| Cho2017 | FBCNet | 48 | $\kappa$: +0.107 [+0.065, +0.151] | $< 0.001$ | $< 0.001$ |
| | | | BA: +4.62 [+3.00, +6.16] | $< 0.001$ | $< 0.001$ |
| Cho2017 | EEG Conformer | 48 | $\kappa$: +0.092 [+0.060, +0.123] | $< 0.001$ | $< 0.001$ |
| | | | BA: +11.78 [+9.10, +14.50] | $< 0.001$ | $< 0.001$ |
| Cho2017 | CTNet | 48 | $\kappa$: +0.236 [+0.182, +0.291] | $< 0.001$ | $< 0.001$ |
| | | | BA: +5.57 [+3.96, +7.15] | $< 0.001$ | $< 0.001$ |
| Cho2017 | TCFormer | 48 | $\kappa$: +0.111 [+0.079, +0.143] | $< 0.001$ | $< 0.001$ |
| | | | BA: +3.97 [+3.13, +4.81] | $< 0.001$ | $< 0.001$ |
| Cho2017 | ATCNet | 48 | $\kappa$: +0.079 [+0.062, +0.096] | $< 0.001$ | $< 0.001$ |
| | | | BA: +4.62 [+3.13, +6.19] | $< 0.001$ | $< 0.001$ |
| Cho2017 | DBConformer | 48 | $\kappa$: +0.092 [+0.062, +0.124] | $< 0.001$ | $< 0.001$ |
| | | | BA: +12.07 [+9.70, +14.49] | $< 0.001$ | $< 0.001$ |
| Cho2017 | MVCNet | 48 | $\kappa$: +0.241 [+0.194, +0.291] | $< 0.001$ | $< 0.001$ |
| | | | BA: +7.36 [+5.47, +9.27] | $< 0.001$ | $< 0.001$ |
| Cho2017 | RatioWaveNet | 48 | $\kappa$: +0.147 [+0.110, +0.185] | $< 0.001$ | $< 0.001$ |
| | | | BA: +13.50 [+10.69, +16.29] | $< 0.001$ | $< 0.001$ |
| Cho2017 | SSTAF | 48 | $\kappa$: +0.270 [+0.214, +0.325] | $< 0.001$ | $< 0.001$ |
| | | | BA: +5.35 [+3.59, +7.07] | $< 0.001$ | $< 0.001$ |
| Cho2017 | EEG-CSANet | 48 | $\kappa$: +0.107 [+0.072, +0.142] | $< 0.001$ | $< 0.001$ |

### H.7 MATCHED COMPONENT CONTROLS FOR SOAMNET

Seven variants isolate the contribution of the routed information path. The non-attention control replaces routing with a trainable full-trajectory pooling head. The local-only variant retains the temporally resolved evidence stream, whereas the global-only variant retains the class-conditioned carrier stream. The remaining controls remove temporal alignment, remove orthogonal query initialization and regularization, replace the uniform-null readout with a standard weighted sum, or add aggregated class evidence directly to the routed logits. Each variant is trained independently under the same splits, seeds, encoder architecture, optimization cap, and validation-based checkpoint selection as the complete model.

Table S18: Participant-paired classification changes of matched component controls relative to complete SOAMNet. Each cell is the variant-minus-complete mean change [95% $t$ confidence interval] after averaging the three seeds within each participant. BA changes are in percentage points and $\kappa$ changes are on the native scale; negative values indicate degradation. Every reported interval excludes zero.

(a) Balanced-accuracy change (pp)

| Variant | BCIC-IV-2b | BCIC-IV-2a | BNCI2015-001 | Cho2017 |
|---|---|---|---|---|
| Non-attention head | -3.51 [-5.23,-1.79] | -5.22 [-6.75,-3.69] | -4.45 [-6.68,-2.22] | -3.98 [-4.85,-3.12] |
| Local content only | -4.51 [-6.01,-3.00] | -7.57 [-9.32,-5.83] | -4.14 [-5.68,-2.61] | -4.29 [-5.35,-3.24] |
| Global carrier only | -2.38 [-3.77,-0.99] | -0.95 [-1.29,-0.62] | -2.29 [-3.75,-0.83] | -2.45 [-2.95,-1.95] |
| No temporal alignment | -2.42 [-3.58,-1.25] | -1.12 [-1.56,-0.67] | -3.31 [-5.08,-1.53] | -2.46 [-3.04,-1.88] |
| No orthogonal queries | -2.26 [-3.76,-0.77] | -0.63 [-1.06,-0.20] | -2.85 [-4.81,-0.89] | -2.13 [-2.69,-1.57] |
| Standard readout | -2.85 [-4.37,-1.32] | -1.04 [-1.39,-0.69] | -3.00 [-4.71,-1.29] | -2.44 [-2.99,-1.89] |
| Direct logit bypass | -2.65 [-4.11,-1.19] | -0.68 [-0.84,-0.52] | -2.97 [-4.66,-1.29] | -2.63 [-3.22,-2.05] |

(b) Cohen's $\kappa$ change

| Variant | BCIC-IV-2b | BCIC-IV-2a | BNCI2015-001 | Cho2017 |
|---|---|---|---|---|
| Non-attention head | -.072 [-.107,-.036] | -.070 [-.090,-.049] | -.089 [-.134,-.044] | -.080 [-.097,-.062] |
| Local content only | -.091 [-.122,-.060] | -.101 [-.124,-.078] | -.083 [-.114,-.052] | -.086 [-.107,-.065] |
| Global carrier only | -.048 [-.076,-.020] | -.013 [-.017,-.008] | -.046 [-.075,-.017] | -.049 [-.059,-.039] |
| No temporal alignment | -.049 [-.072,-.025] | -.015 [-.021,-.009] | -.066 [-.102,-.031] | -.049 [-.061,-.038] |
| No orthogonal queries | -.046 [-.077,-.015] | -.008 [-.014,-.003] | -.057 [-.096,-.018] | -.042 [-.054,-.031] |
| Standard readout | -.058 [-.089,-.026] | -.014 [-.019,-.009] | -.060 [-.094,-.026] | -.049 [-.060,-.038] |
| Direct logit bypass | -.054 [-.084,-.024] | -.009 [-.011,-.007] | -.059 [-.093,-.026] | -.053 [-.064,-.041] |

To locate functional reliance within the complete model, we also froze every selected checkpoint and decomposed the attention scores and Values into their content and carrier terms. Each frozen condition removes only the named score or Value term; the one-path forwards retain the corresponding score and Value stream, together with the positional bias. "Direct aggregated evidence" uses $m$ as logits before routing. The controlled swaps either restrict the attention-distance change to that of the evidence-matched condition or select, without labels, a donor with similar pre-routing aggregated evidence.

Table S19: Functional reliance of the independently trained controls and frozen pathway attribution within complete SOAMNet. Panel (a) reports $R_{\mathrm{core}}$, the weaker mean BA degradation under cross-sample and training-set-mean replacement. Panel (b) reports frozen-checkpoint BA degradation relative to the unmodified forward pass as mean [participant-bootstrap 95% interval], with positive values denoting worse performance. Seeds are averaged within participant before aggregation. A dagger indicates that both intervals entering $R_{\mathrm{core}}$ have positive lower bounds.

(a) Functional-reliance strength, $R_{\mathrm{core}}$ (pp)

| Variant | BCIC-IV-2b | BCIC-IV-2a | BNCI2015-001 | Cho2017 |
|---|---|---|---|---|
| Complete SOAMNet | 28.46$^{\dagger}$ | 40.95$^{\dagger}$ | 33.81$^{\dagger}$ | 18.61$^{\dagger}$ |
| Non-attention head | — | — | — | — |
| Local content only | 4.76$^{\dagger}$ | 3.96$^{\dagger}$ | 9.24$^{\dagger}$ | 0.60 |
| Global carrier only | 25.38$^{\dagger}$ | 41.22$^{\dagger}$ | 32.23$^{\dagger}$ | 20.61$^{\dagger}$ |
| No temporal alignment | 24.16$^{\dagger}$ | 40.97$^{\dagger}$ | 25.19$^{\dagger}$ | 14.67$^{\dagger}$ |
| No orthogonal queries | 23.55$^{\dagger}$ | 37.45$^{\dagger}$ | 27.79$^{\dagger}$ | 14.68$^{\dagger}$ |
| Standard readout | 6.25$^{\dagger}$ | 30.72$^{\dagger}$ | 8.76$^{\dagger}$ | 7.45$^{\dagger}$ |
| Direct logit bypass | 0.47 | 0.46$^{\dagger}$ | $-0.75$ | 1.38$^{\dagger}$ |

(b) Frozen pathway attribution: BA degradation (pp)

| Condition | BCIC-IV-2b | BCIC-IV-2a | BNCI2015-001 | Cho2017 |
|---|---|---|---|---|
| Remove content Q–K scores | 1.82 [0.64, 3.25] | 1.29 [0.95, 1.68] | 1.08 [0.64, 1.57] | 4.29 [3.10, 5.64] |
| Remove carrier scores | 25.69 [14.20, 37.21] | 13.94 [9.81, 18.18] | 8.52 [5.17, 12.12] | 10.96 [8.41, 13.62] |
| Remove carrier Values | 18.10 [11.63, 24.47] | 27.55 [23.60, 32.12] | 29.86 [22.98, 36.76] | 9.51 [7.56, 11.57] |
| Carrier path only | 1.96 [0.80, 3.34] | 1.45 [0.94, 2.03] | 1.49 [0.87, 2.17] | 4.56 [3.25, 6.09] |
| Local path only | 22.31 [15.73, 28.58] | 34.00 [29.28, 39.06] | 29.92 [21.63, 38.14] | 13.25 [10.79, 15.80] |
| Direct aggregated evidence | 2.78 [1.58, 4.14] | 1.53 [1.03, 2.04] | 1.90 [1.20, 2.68] | 4.49 [3.17, 6.02] |
| Distance-controlled swap | 3.24 [1.08, 6.14] | 4.32 [3.10, 5.56] | 1.60 [0.97, 2.33] | 6.16 [4.88, 7.55] |
| Evidence-matched swap | 2.04 [0.86, 3.40] | 1.50 [1.04, 1.93] | 1.32 [0.63, 2.21] | 3.12 [2.39, 3.90] |