NEUROINTENTFORMER: EVIDENCE-CONDITIONED TEMPORAL AGGREGATION FOR MOTOR IMAGERY DECODING IN STROKE REHABILITATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Decoding left- and right-hand motor imagery from stroke recordings can provide intention signals for rehabilitation-oriented interaction, but discriminative evidence varies across patients and trials. Beyond learning temporal representations, a decoder must determine how their content and class predictions should inform a joint decision. We formulate this step as evidence-conditioned temporal aggregation and introduce NeuroIntentFormer. The model constructs complementary temporal evidence and lets the evidence content and predictions jointly condition source aggregation and decision refinement. On seven held-out participants from a subject-disjoint 49-participant stroke cohort, NeuroIntentFormer achieves 85.53% subject-mean balanced accuracy (BAcc), with a participant-level bootstrap 95% confidence interval of [81.10%, 89.84%]. A parameter-matched ablation shows that evidence conditioning with joint decision refinement improves BAcc across all test trials by 1.76 percentage points (pp) over context-only routing. On OpenBMI, the same architecture reaches 82.03% subject-mean BAcc, with a 1.65 pp difference over MSVTNet. These results support the joint decision mechanism for offline motor imagery decoding.

1 INTRODUCTION

Restoring hand function is a central goal of post-stroke rehabilitation. Brain-computer interfaces (BCIs) can couple neural activity with intention-contingent feedback for rehabilitation training (Biasiucci et al., 2018). Figure 1 illustrates how decoding the imagined hand can provide intention-related information for therapist-supervised repetitive practice, while confidence-aware decisions can support review when the estimated intention is uncertain. We study subject-disjoint decoding of visually guided left- and right-hand motor imagery (MI) from multichannel stroke recordings, asking whether a model trained on one patient group can recover intention from unseen patients rather than memorize patient-specific patterns.

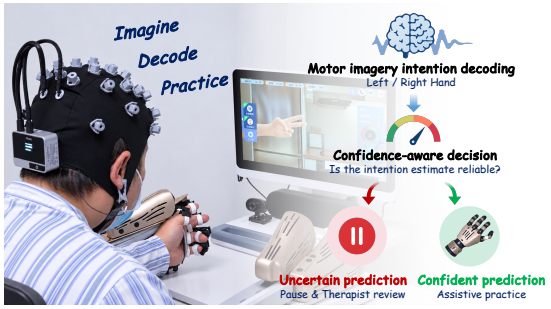


Figure 1: Motivation and potential use of motor-imagery decoding in stroke rehabilitation. The photograph on the left depicts an actual hospital-based stroke rehabilitation training setting, in which a patient performs hand motor imagery while multichannel electrophysiological signals are recorded. The conceptual pathway on the right illustrates left- versus right-hand intention decoding followed by a confidence-aware decision.

In patients with stroke, lesions and motor impairment may alter the spatial and temporal expression of motor imagery, leading to heterogeneous decoding evidence across patients (Xu et al., 2021; Liu et al., 2024a). Across patients and trials, class-relevant evidence may occur at different temporal locations and may be captured by different inductive biases. Several evidence sources can support the same prediction, emphasize different aspects of the signal, or provide conflicting class preferences. A decoder must

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.

therefore determine not only what evidence is present, but also how the current evidence should be combined into one decision. This motivates constructing complementary temporal views and adapting their aggregation to the evidence available in each trial.

Convolutional, attention-based, and Transformer-based EEG decoders have substantially improved temporal and spatial representation learning (Lawhern et al., 2018; Song et al., 2023; Altaheri et al., 2023; Liu et al., 2024b), while pretrained EEG models target transfer across datasets and recording configurations (Yang et al., 2023; Jiang et al., 2024; Wang et al., 2025; El Ouahidi et al., 2025). Multi-scale interaction, mixture-of-experts routing, and prediction-level stacking provide flexible ways to combine representations or outputs (Liu et al., 2024b;c; Zhu et al., 2025). These developments motivate a more specific question for MI decoding. How should the latent content and class predictions of complementary temporal sources jointly inform their aggregation and final decision?

To address this question, we introduce NeuroIntentFormer, which learns how temporal evidence should be combined into a joint decision. Our contributions are threefold. First, we formulate motor imagery decoding in stroke patients as evidence-conditioned temporal aggregation, enabling the decoder to adaptively combine complementary temporal evidence across heterogeneous patients and trials. Second, we develop NeuroIntentFormer, an end-to-end decoder that learns complementary temporal sources and couples source weighting with residual decision refinement. Source representations and class predictions jointly inform aggregation, while their prediction relationships guide refinement. Third, subject-disjoint evaluations yield subject-mean BAcc of 85.53% on Stroke and 82.03% on OpenBMI. Parameter-matched ablations further support the interaction between evidence-conditioned routing and refinement, linking the observed performance to the proposed decision mechanism.

2 Related Work

2.1 Motor Imagery Decoding for Stroke Rehabilitation

Motor-imagery-based BCIs associate intended movement with rehabilitation feedback and assistance (Biasiucci et al., 2018). Early decoding pipelines relied on common spatial patterns (CSP), filter-bank common spatial patterns (FBCSP), linear discriminant analysis (LDA), and support vector machines (SVMs) (Ramoser et al., 2000; Ang et al., 2008). Later work introduced task-specific convolutional networks and transfer learning to improve representation quality and cross-participant generalization (Schirrmeister et al., 2017; Lawhern et al., 2018; Mane et al., 2021; Xu et al., 2021). Public stroke datasets now support subject-disjoint evaluation of visually guided left- and right-hand imagery in larger patient cohorts (Liu et al., 2024a). These studies establish the clinical motivation and the need for patient-level evaluation. They primarily address feature extraction or transfer, while our work studies how several learned sources of temporal evidence should be combined for each trial.

2.2 Temporal EEG Representation Learning

Modern EEG decoders combine convolution, attention, and Transformer architectures to capture local patterns, long-range dependencies, and multiple temporal scales (Vaswani et al., 2017; Song et al., 2023; Altaheri et al., 2023; Zhao et al., 2024; Liu et al., 2024b). Multi-scale models construct several spatiotemporal representations and explicitly model interactions among them. Latent cross-attention provides a general mechanism for summarizing long input sequences through a compact latent representation (Jaegle et al., 2021), while related EEG architectures use temporal attention or contextual feature interaction to adapt representations to the current input (Hu et al., 2023; Yuan et al., 2026). These methods motivate the use of complementary temporal representations, but the presence of multiple representations does not by itself specify how their predictions should jointly determine the final decision.

Pretrained EEG models pursue a complementary objective. BIOT, LaBraM, CBraMod, and REVE learn reusable representations across datasets, tasks, and recording configurations (Yang et al., 2023; Jiang et al., 2024; Wang et al., 2025; El Ouahidi et al., 2025). Their contribution concerns the scale and transferability of the encoder. NeuroIntentFormer instead starts from a set of learned evidence sources and addresses the decision that follows. The question is how the content and prediction of each source should influence the aggregation of all sources for a particular trial.

2.3 CONDITIONAL AGGREGATION

Attention, adaptive fusion, and mixture-of-experts models assign input-dependent importance to features or expert functions. This principle has been applied to MI decoding through sparse expert routing and cross-participant expert specialization (Liu et al., 2024c; Xu et al., 2025). Prediction-level stacking follows a different route and learns a meta-classifier over outputs produced by feature-specific models (Zhu et al., 2025). These approaches provide learned alternatives to fixed averaging, but they generally derive the combination rule from the input, a shared state, or predictions after the source representations have been formed.

Recent source-free methods adapt trained EEG decoders using unlabeled target data through manifold optimization or gradient-free Bayesian calibration (Li et al., 2025; Luo et al., 2026). Our setting instead studies within-trial evidence aggregation by a fixed decoder, without target-time adaptation.

NeuroIntentFormer explicitly conditions dense source weights on both source representations and class logits, together with the encoded signal context. Its distinction lies in this joint information dependency and its end-to-end residual decision refinement, rather than in adaptive weighting alone. Section 3 provides the corresponding mathematical formulation.

3 METHOD

3.1 OVERVIEW AND PROBLEM FORMULATION

NeuroIntentFormer treats motor imagery decoding as a decision over a learned set of temporal evidence sources. Given a multichannel recorded epoch $X \in \mathbb{R}^{C\times T}$, where X is the input signal matrix, C is the number of input signal columns, and T is the number of temporal samples, the model first constructs K complementary sources. A binary vector $m \in \{0,1\}^C$ indicates the available input signals. Each source produces a latent representation z_k and class logits $\ell_k \in \mathbb{R}^2$ for left and right motor imagery. Together they form the evidence set

$$\mathcal{E}(X,m) = \{(z_k, \ell_k)\}_{k=1}^{K},$$

where the current model uses $K = 4$ sources with complementary temporal and representational biases. A raw descriptor $r(X,m)$ summarizes the current observation and channel availability, and its embedding c serves as the encoded signal context. The router conditions on both the evidence set and this context to produce sample-specific source weights. Their output and the resulting aggregated source logits are

$$g = G(\mathcal{E}, c) = [g_1, \ldots, g_K], \qquad \ell_{\text{agg}} = \sum_{k=1}^{K} g_k \ell_k.$$

Each component g_k is the normalized weight assigned to source k.

A deterministic evidence vector e summarizes the aggregated prediction, source predictions, and routing weights. The jointly trained readout R maps this vector to a residual correction scaled by α. The final logits are

$$\ell = \ell_{\text{agg}} + \alpha R(e), \qquad \hat{y} = \arg\max_{j\in\{0,1\}} [\ell]_j.$$

Here, $[\ell]_j$ is the class-j logit. Classes 0 and 1 denote left- and right-hand motor imagery, respectively.

The central idea is that the same source evidence determines how source predictions are aggregated and how the aggregate is refined. NeuroIntentFormer therefore learns evidence construction, evidence-conditioned aggregation, and decision refinement as one end-to-end function. Figure 2 summarizes this computation from the recorded signal epoch to the final left or right prediction.

3.2 COMPLEMENTARY TEMPORAL EVIDENCE

Evidence construction follows $X \to H \to H_k \to (z_k, \ell_k)$. We write $\widetilde{X} = X \odot m$ for the available input, where $\odot$ denotes element-wise multiplication and the signal mask m is broadcast across

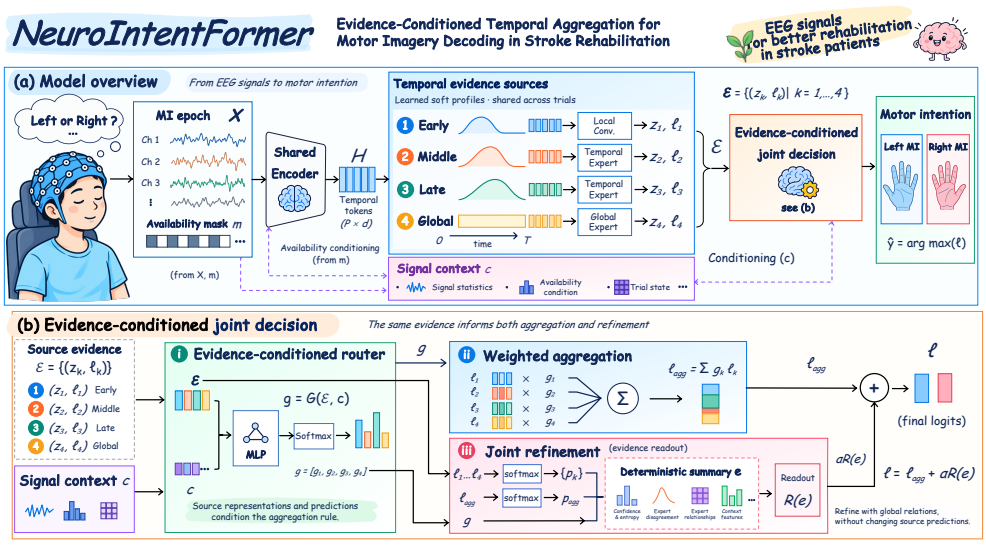


Figure 2: Overview of NeuroIntentFormer. A recorded signal epoch is encoded into four complementary temporal evidence sources. Their representations and logits, together with signal context c, condition the router $G(\mathcal{E}, c)$. Weighted source logits and the residual readout produce the left- or right-hand prediction; solid and dashed arrows denote prediction and conditioning paths.

time. The shared encoder F_{shared} maps this input to $H = F_{\text{shared}}(\widetilde{X})$. Here, $H \in \mathbb{R}^{P\times d}$ contains P temporally ordered tokens with hidden dimension d. The shared encoder combines temporal convolution, filtering across input signals, temporal downsampling, token projection, and positional encoding.

The model derives three temporally localized views from H. Let $u_p = p/(P-1)$ for $p = 0, \ldots, P-1$ denote normalized token positions in $[0, 1]$. For localized source $k \in \{1, 2, 3\}$, a Gaussian profile with learnable center μ_k and width σ_k is sharpened by temperature τ, peak-normalized, and assigned a floor ρ

$$q_{k,p} = \exp\left[-\frac{1}{2}\left(\frac{u_p - \mu_k}{\sigma_k}\right)^2\right], \qquad M_{k,p} = \rho + (1-\rho)\frac{q_{k,p}^{1/\tau}}{\max_{0\le j<P} q_{k,j}^{1/\tau}}. \tag{1}$$

The profile is broadcast over the hidden dimension to form $H_k = H \odot M_k$. These profiles are global model parameters shared across samples. They reweight the complete sequence instead of applying hard temporal crops. A fourth source uses $M_{K,p} = 1$ and receives the full sequence as global context. This design allows the model to learn where useful evidence tends to occur while preserving information outside each emphasized region. Appendix B gives the center and width parameterization and the temporal-profile hyperparameters.

The four views are processed by source-specific encoders. The early source emphasizes local temporal structure, the middle and late sources use Transformer encoders, and the global source summarizes the full sequence through prototype-based cross-attention. Their different inductive biases encourage complementary evidence without creating separate training pipelines.

The mask encoder F_m maps channel availability to the embedding $q_m = F_m(m)$, which is provided to every source. Each source encoder E_k then returns its latent representation and class logits

$$(z_k, \ell_k) = E_k(H_k, q_m).$$

The condition informs each source about which input signals are present. Learnable prototypes summarize class-related and global information inside each source. Appendix B provides the encoder layouts, prototype construction, and tensor dimensions.

3.3 EVIDENCE-CONDITIONED AGGREGATION

The usefulness of an evidence source can vary across trials. NeuroIntentFormer therefore computes its aggregation rule only after observing all source representations and predictions. A raw descriptor $r(X, m)$ is computed from the masked signal $\widetilde{X}$ and availability mask m. It summarizes temporal signal magnitude, coarse variation across input groups, global scale, and channel availability, and is encoded as signal context $c = F_{\text{state}}(r)$. The input groups are defined by column indices and are not assigned hemispheric or lesion-related meaning.

The router jointly conditions on all source representations, source logits, and the current signal context

$$g = \text{softmax}\left(\text{MLP}_{\text{route}}\left[z_1 \| \cdots \| z_K \| \ell_1 \| \cdots \| \ell_K \| c\right]\right).$$

Here, $\|$ denotes vector concatenation in the fixed early, middle, late, and global source order, and $\text{MLP}_{\text{route}}$ is the router multilayer perceptron. Unlike fixed averaging, the source weights are estimated separately for each trial from the evidence available in that trial. All sources are evaluated, so the router performs dense aggregation rather than sparse expert selection.

The weighted source predictions form the aggregated source logits

$$\ell_{\text{agg}} = \sum_{k=1}^{K} g_k \ell_k.$$

Because z_k and ℓ_k also condition the router, g_k measures the direct mixing weight of a source rather than its complete influence on the decision. We therefore interpret routing weights together with source ablations and temporal intervention analyses.

3.4 JOINT DECISION REFINEMENT AND LEARNING

The aggregated source decision does not explicitly preserve how strongly the sources agree or disagree. NeuroIntentFormer therefore applies a lightweight residual refinement. Let $p_{\text{agg}} = \text{softmax}(\ell_{\text{agg}})$ denote the aggregated source probability vector and $p_k = \text{softmax}(\ell_k)$ denote the probability vector of source k. The evidence vector and final logits are

$$e = \phi\left(p_{\text{agg}}, \{p_k\}_{k=1}^{K}, g\right), \qquad \ell = \ell_{\text{agg}} + \alpha R(e),$$

where ϕ deterministically constructs a 20-dimensional summary from the same forward pass. It contains four aggregated-posterior statistics, one cross-source disagreement statistic, twelve source-specific entries comprising each source's gate, right-class probability, and confidence, and three pairwise source contrasts. The evidence readout network R maps this summary to a residual correction, and α controls the correction magnitude.

The readout is optimized jointly with the shared encoder, evidence sources, and router. It is therefore part of the end-to-end decision function rather than a post hoc classifier. Its inputs are deterministic functions of the model state, are not detached, and are not presented as independent observations or calibrated uncertainty estimates. Appendix B and Table 5 specify the readout architecture, feature ordering, and definitions.

Training uses the original trial together with channel-drop and time-shift views. The complete objective is written as

$$L = L_{\text{cls}} + L_{\text{con}} + L_{\text{reg}}.$$

Here, L_{cls} supervises the full and perturbed views, L_{con} aligns each perturbed prediction with the stop-gradient full-view prediction, and L_{reg} preserves latent variability, supervises probability quality and source-level discriminability, and regularizes routing behavior. All components are optimized jointly. Appendix B gives the full expansion, coefficients, augmentation magnitudes, and optimization settings.

4 EXPERIMENTS

We organize the experiments around four questions: subject-disjoint stroke decoding, evidence-conditioned decision making, complementary temporal evidence, and utility beyond the primary patient comparison.

4.1 EXPERIMENTAL SETUP

Datasets. We evaluate left- versus right-hand MI on the acute-stroke dataset of Liu et al. (2024a), termed *Stroke*, and OpenBMI (Lee et al., 2019). Stroke includes 49 usable patients with a 35/7/7 training/validation/test split and 32 recorded signals. OpenBMI includes 54 healthy participants with a 38/8/8 split and 62 EEG channels. Models are trained or adapted independently within each dataset using subject-disjoint partitions. Appendix A details exclusions, epoch construction, and preprocessing; Appendices C and F.1 specify adaptation and training settings.

The complete performance comparison uses the same methods on both datasets. Diagnostic figures use question-specific representative subsets. Participant-paired figures compare NeuroIntentFormer with DeepConvNet, EEG-Conformer, MSVTNet, TCFormer (Altaheri et al., 2025), and REVE-LP, covering distinct model families and the strongest patient Transformer baseline.

Evaluation. NeuroIntentFormer checkpoints are selected using validation pooled balanced accuracy, and its prediction rule is argmax. Baseline selection and decision rules follow the protocols in Appendix C. Balanced accuracy is the primary metric, with $\mathrm{BAcc}_s = \frac{1}{2}(\mathrm{Recall}_{L,s} + \mathrm{Recall}_{R,s})$ for participant s, where $\mathrm{Recall}_{L,s}$ and $\mathrm{Recall}_{R,s}$ are the left- and right-class recalls. We report subject-mean BAcc with sample standard deviation across participants and participant-level bootstrap 95% confidence intervals (CIs), together with pooled BAcc, macro-averaged F1 score (Macro-F1), and area under the receiver operating characteristic curve (AUROC).

Implementation. NeuroIntentFormer is optimized end to end with AdamW, cosine learning-rate decay, and validation-based early stopping. Checkpoints are selected by validation pooled BAcc under argmax, and the test set is evaluated only after selection. Appendix B reports the complete architecture, objective, and optimization settings.

4.2 RQ1: SUBJECT-DISJOINT STROKE DECODING

Table 1 compares classical pipelines, task-specific deep models, pretrained EEG references, and NeuroIntentFormer. Subject-mean BAcc weights patients equally, while pooled BAcc summarizes all test trials. We first examine Stroke and return to OpenBMI in Section 4.5.

NeuroIntentFormer leads the Stroke comparison in both subject-mean and pooled BAcc, exceeding MSVTNet by 3.62 pp and 3.10 pp. Its leading Macro-F1 and AUROC in Appendix D further support the advantage in class decisions and score ranking. The consistent ordering across these summaries motivates examining the contributions of evidence construction and decision learning.

Figure A1a shows positive mean gains over all five representative baselines. Against MSVTNet, the mean gain is 3.62 pp with a paired 95% confidence interval from 0.92 to 5.98 pp, and six of seven patients benefit. These paired observations show that the aggregate advantage is accompanied by improvements across multiple patients.

4.3 RQ2: EVIDENCE-CONDITIONED DECISION MAKING

Evidence conditioning and decision refinement. We test aggregation conditioned on the evidence set and signal context, $G(\mathcal{E}, c)$, against aggregation conditioned only on signal context, $G(c)$. The context-only router uses the encoded signal context without source representations or logits. Its context embedding is repeated to match the full router's input width and parameter count. We enable or disable the readout under each conditioning rule while retaining the remaining architecture and training protocol. With the readout enabled, evidence conditioning improves pooled test BAcc by 1.76 pp. Appendix E records the four configurations and their aggregate results.

Table 1: Held-out motor imagery decoding results on Stroke and OpenBMI (%).

	Stroke		OpenBMI
Method	Subject BAcc	Pooled BAcc	Subject BAcc
Classical feature methods			
CSP + LDA (Ramoser et al., 2000)	56.92 ± 6.32	57.78	59.91 ± 9.94
CSP + SVM (Ramoser et al., 2000)	58.10 ± 7.84	58.68	60.03 ± 9.53
FBCSP + LDA (Ang et al., 2008)	56.53 ± 4.25	57.54	67.22 ± 13.82
Task-specific deep models			
EEGNet (Lawhern et al., 2018)	80.24 ± 7.93	80.84	79.19 ± 10.95
DeepConvNet (Schirrmeister et al., 2017)	76.34 ± 11.25	78.65	80.69 ± 9.49
FBCNet (Mane et al., 2021)	75.05 ± 9.30	74.95	76.00 ± 10.31
EEG-Conformer (Song et al., 2023)	79.70 ± 9.90	80.18	76.03 ± 9.91
ATCNet (Altaheri et al., 2023)	81.00 ± 8.13	82.14	78.91 ± 9.29
CTNet (Zhao et al., 2024)	80.15 ± 10.13	81.46	78.97 ± 10.57
MSVTNet (Liu et al., 2024b)	81.91 ± 6.37	82.85	80.38 ± 10.50
TCFormer (Altaheri et al., 2025)	79.61 ± 6.95	80.46	80.00 ± 9.33
Pretrained EEG references			
BIOT (Yang et al., 2023)	63.07 ± 6.92	61.93	59.91 ± 12.16
LaBraM (Jiang et al., 2024)	74.56 ± 14.50	76.29	69.94 ± 8.98
CBraMod (Wang et al., 2025)	75.54 ± 8.43	76.67	51.00 ± 1.94
REVE-LP (El Ouahidi et al., 2025)	75.66 ± 5.57	77.82	72.06 ± 10.05
REVE-LP+LoRA (El Ouahidi et al., 2025)	75.60 ± 7.91	77.64	74.00 ± 7.46
Proposed method			
NeuroIntentFormer	**85.53** ± 6.53	**85.95**	**82.03** ± 9.50

Subject BAcc is reported as mean ± SD across seven Stroke patients or eight OpenBMI participants. Pooled BAcc aggregates all test trials. Bold indicates the highest BAcc in each column. Model protocols are detailed in Appendix C.

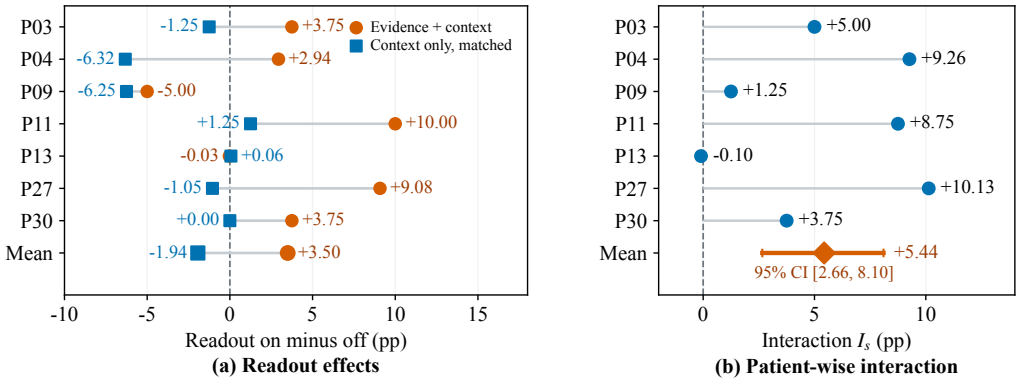


Figure 3: Patient-wise interaction between evidence conditioning and refinement. Panel a shows BAcc changes from enabling the readout under each router for the same seven held-out patients. Panel b shows their difference I_s in pp. The diamond and interval denote the patient mean and paired participant-bootstrap 95% CI. All checkpoints are fixed and validation-selected. Positive interaction indicates a more favorable readout effect under evidence-conditioned routing.

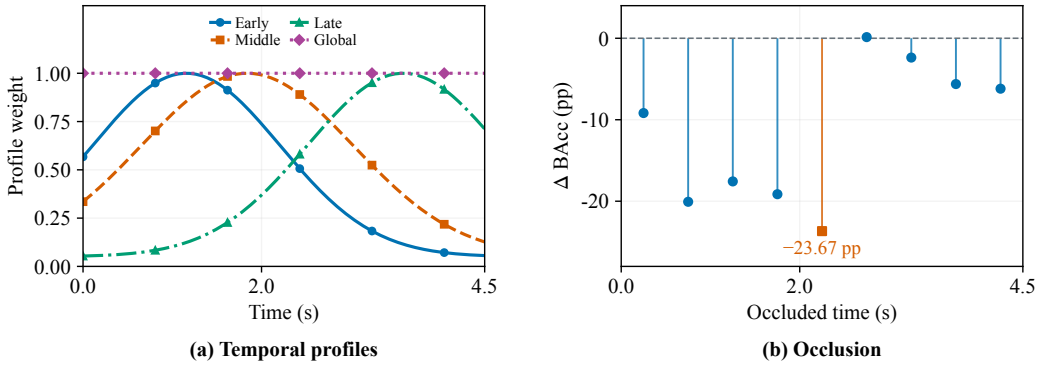
To examine whether the interaction extends across patients, let $B_{r,u,s}$ denote BAcc in percent for patient s, router condition $r \in \{E, S\}$, and readout setting $u \in \{\text{on}, \text{off}\}$. Here, E denotes evidence-conditioned routing and S the matched context-only control. We compute

$$I_s = (B_{E,\text{on},s} - B_{E,\text{off},s}) - (B_{S,\text{on},s} - B_{S,\text{off},s}).$$

Positive I_s indicates that refinement has a more favorable effect under evidence-conditioned routing.

Table 2: Evidence-construction and training ablations on the Stroke dataset. Values are percentages except for Δ, the test BAcc difference from the full model in pp. Each variant is trained independently and selected using validation BAcc. The interaction between routing and readout is shown separately in Figure 3.

Design	Val BAcc	Test BAcc	Δ (pp)
Full model	84.39	85.95	+0.00
Fixed temporal profiles	83.67	84.30	-1.65
Homogeneous Transformers	81.49	84.26	-1.69
Uniform mean fusion	80.94	84.48	-1.47
No channel/time consistency	82.39	81.71	-4.24



(a) Temporal profiles **(b) Occlusion**

Figure 4: Temporal evidence on Stroke. Panel a shows learned trial-shared soft profiles and the fixed global profile. Panel b reports changes in pooled BAcc after independently zeroing each 0.5 s interval. More negative values indicate greater sensitivity to the occluded interval.

Figure 3 compares the effect of refinement under the two routing conditions and then isolates their interaction. The refinement effect is 5.44 pp more favorable under evidence-conditioned routing, with positive interaction in six of seven patients and a paired 95% confidence interval from 2.66 to 8.10 pp. This supports a coupled decision process in which the value of refinement depends on the information used to aggregate the sources.

Evidence construction and training design. Table 2 supports learned temporal emphasis, heterogeneous source encoders, and sample-specific fusion over their respective controls. Removing channel and time consistency produces the largest decrease of 4.24 pp in test BAcc. This suggests that learning from perturbed views contributes substantially to generalization alongside the architectural choices.

4.4 RQ3: Complementary Temporal Evidence

Figure 4a shows overlapping temporal profiles with distinct emphasis, alongside the fixed global profile. These shared profiles provide complementary views of each epoch, whose evidence is combined by the sample-specific router.

Figure 4b probes the complete decoder by zeroing one interval at a time. Occluding 2.0 to 2.5 s reduces pooled BAcc from 85.95% to 62.28%, whereas the change for 2.5 to 3.0 s is close to zero. This contrast links temporal evidence construction to the uneven temporal sensitivity of the final prediction.

4.5 RQ4: Complementary Validation and Practicality

Finally, we examine whether the same architecture can be trained on healthy-participant MI data and assess its computational and confidence-based operating characteristics. OpenBMI provides

complementary within-dataset validation. The efficiency and review analyses characterize different aspects of offline decoder use.

Healthy-participant validation. All methods in the OpenBMI column of Table 1 are fitted or adapted independently on the fixed 38/8/8 participant split. Task-specific neural models are trained from scratch, whereas pretrained models retain their respective initialization and adaptation procedures. Test predictions use argmax after validation-based selection.

On OpenBMI, NeuroIntentFormer achieves the highest subject-mean BAcc at 82.03%, compared with 80.69% for DeepConvNet and 80.38% for MSVTNet. Independent training on healthy-participant recordings extends the empirical support for the architecture beyond the patient cohort. Figure A1b shows the participant-level distribution of these gains.

Model efficiency. NeuroIntentFormer uses 1.903 million parameters and an estimated 0.566 GFLOPs per trial, with a median forward latency of 5.80 ms on an RTX 4090 D using AMP FP16. It achieves the leading Stroke BAcc with more parameters and higher latency than compact task-specific models, but fewer parameters and lower latency than REVE-LP under the measured protocols. Appendix F.3 details the trade-offs in parameters, latency, throughput, and memory.

Offline review option. Appendix G examines how validation-selected confidence thresholds balance accepted prediction quality and review workload. This complements forced-choice accuracy by characterizing which predictions are retained for automatic output.

5 Discussion and Limitations

The results support learning evidence construction, aggregation, and refinement as a coupled prediction process. Patient-level gains establish its empirical value, while matched ablations show that refinement is more effective when routing has access to source evidence. Temporal and training analyses connect this benefit to the construction and learning of complementary evidence.

OpenBMI extends the evaluation through independent training, and the efficiency and review analyses characterize computational cost and automatic prediction quality. The evidence concerns offline decoding with a single-center clinical cohort. Broader cohorts and prospective integration of review policies are natural next steps toward rehabilitation applications.

6 Conclusion

We have introduced NeuroIntentFormer for evidence-conditioned temporal aggregation in motor imagery decoding. Evaluations on Stroke and OpenBMI have supported its decoding performance, while matched ablations have demonstrated a more favorable refinement effect under evidence-conditioned routing. These findings have provided empirical support for jointly learning evidence construction and decision making in rehabilitation-oriented MI research.

Artificial Intelligence Use Statement

Generative artificial intelligence (AI) tools assisted with research planning, code organization, experiment-record consolidation, figure preparation, and language editing. The authors remain responsible for the study design, source verification, numerical results, interpretation, and final manuscript.

Reproducibility Statement

Appendices A, B, and C document preprocessing, model specifications, and evaluation protocols. Reported results are traceable to their source records. We plan to release the implementation and experiment configurations as open-source code in the future.

References

Hamdi Altaheri, Ghulam Muhammad, and Mansour Alsulaiman. Physics-informed attention temporal convolutional network for EEG-based motor imagery classification. *IEEE Transactions on Industrial Informatics*, 19(2):2249–2258, Feb. 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), Sep. 2025. doi: 10.1038/s41598-025-16219-7.

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

A. Biasiucci, R. Leeb, I. Iturrate, S. Perdikis, A. Al-Khodairy, T. Corbet, A. Schnider, T. Schmidlin, H. Zhang, M. Bassolino, D. Viceic, P. Vuadens, A. G. Guggisberg, and J. d. R. Millán. Brain-actuated functional electrical stimulation elicits lasting arm motor recovery after stroke. *Nature Communications*, 9(1), Jun. 2018. doi: 10.1038/s41467-018-04673-z.

Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. REVE: A foundation model for EEG - adapting to any setup with large-scale pretraining on 25,000 subjects. In *Advances in Neural Information Processing Systems 38*, pp. 25583–25619, San Diego, California, USA and Mexico City, Mexico, 2025. Neural Information Processing Systems Foundation, Inc. (NeurIPS). doi: 10.52202/085713-0760.

Lingyan Hu, Weijie Hong, and Lingyu Liu. MSATNet: multi-scale adaptive transformer network for motor imagery classification. *Frontiers in Neuroscience*, 17, Jun. 2023. doi: 10.3389/fnins.2023.1173778.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4651–4664. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/jaegle21a.html.

Wei-Bang Jiang, Liming Zhao, and Bao-liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 16405–16426, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/47393e8594c82ce8fd83adc672cf9872-Paper-Conference.pdf.

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, Oct. 2018. doi: 10.1088/1741-2552/aace8c.

Min-Ho Lee, O-Yeon Kwon, Yong-Jeong Kim, Hong-Kyung Kim, Young-Eun Lee, John Williamson, Siamac Fazli, and Seong-Whan Lee. EEG dataset and OpenBMI toolbox for three BCI paradigms: An investigation into BCI illiteracy. *GigaScience*, 8(5):giz002, May 2019. doi: 10.1093/gigascience/giz002.

Shanglin Li, Motoaki Kawanabe, and Reinmar J. Kobler. SPDIM: Source-free unsupervised conditional and label shift adaptation in EEG. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/fc6cd575b8340be6ee70beaed9ca4015-Abstract-Conference.html.

Haijie Liu, Penghu Wei, Haochong Wang, Xiaodong Lv, Wei Duan, Meijie Li, Yan Zhao, Qingmei Wang, Xinyuan Chen, Gaige Shi, Bo Han, and Junwei Hao. An EEG motor imagery dataset for brain computer interface in acute stroke patients. *Scientific Data*, 11(1), Jan. 2024a. doi: 10.1038/s41597-023-02787-8.

Ke Liu, Tao Yang, Zhuliang Yu, Weibo Yi, Hong Yu, Guoyin Wang, and Wei Wu. MSVTNet: Multi-scale vision transformer neural network for EEG-based motor imagery decoding. *IEEE Journal of Biomedical and Health Informatics*, 28(12):7126–7137, Dec. 2024b. doi: 10.1109/jbhi.2024.3450753.

Ming Liu, Yanbing Liu, Weiyou Shi, Yitai Lou, Yuan Sun, Qi Meng, Dezheng Wang, Fangzhou Xu, Yang Zhang, Lei Zhang, and Jiancai Leng. EMPT: a sparsity transformer for EEG-based motor imagery recognition. *Frontiers in Neuroscience*, 18, Apr. 2024c. doi: 10.3389/fnins.2024.1366294.

Huan Luo, Na Lu, Xiaopeng Wang, and Xu Niu. Bayesian test-time adaptation via dirichlet feature projection and GMM-driven inference for motor imagery EEG decoding. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/36eeec256f89eba9ed60cbca73ea92ed-Abstract-Conference.html.

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, 2021. URL https://arxiv.org/abs/2104.01233.

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

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, Nov. 2017. doi: 10.1002/hbm.23730.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBraMod: A criss-cross brain foundation model for EEG decoding. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 75310–75346, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/bbbd6d915cb90be21c1254a82d45cedd-Paper-Conference.pdf.

Fangzhou Xu, Yunjing Miao, Yanan Sun, Dongju Guo, Jiali Xu, Yuandong Wang, Jincheng Li, Han Li, Gege Dong, Fenqi Rong, Jiancai Leng, and Yang Zhang. A transfer learning framework based on motor imagery rehabilitation for stroke. *Scientific Reports*, 11(1), Oct. 2021. doi: 10.1038/s41598-021-99114-1.

Jingzhou Xu, Haoyu Wu, Yong Yue, and Jun Qi. A mixture-of-expert model for cross-subject motor imagery decoding. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 6149–6153, Wuhan, China, Dec. 2025. IEEE. doi: 10.1109/bibm66473.2025.11357140.

Chaoqi Yang, M Westover, and Jimeng Sun. BIOT: Biosignal transformer for cross-data learning in the wild. In *Advances in Neural Information Processing Systems 36*, pp. 78240–78260, New Orleans, Louisiana, USA, 2023. Neural Information Processing Systems Foundation, Inc. (NeurIPS). doi: 10.52202/075280-3420.

Jiahui Yuan, Rui Zhang, Yazhou Zhao, Weidong Zhou, Lan Tian, and Guoyang Liu. PG-MCTFormer: A prior-guided multi-scale convolutional transformer for interpretable motor imagery EEG classification. *Biomimetics*, 11(6):377, May 2026. doi: 10.3390/biomimetics11060377.

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), Aug. 2024. doi: 10.1038/s41598-024-71118-7.

Xianglong Zhu, Ming Meng, Zewen Yan, and Zhizeng Luo. Motor imagery EEG classification based on multi-domain feature rotation and stacking ensemble. *Brain Sciences*, 15(1):50, Jan. 2025. doi: 10.3390/brainsci15010050.

APPENDIX

A DATA AND TASK PROTOCOL

A.1 STROKE COHORT

The dataset introduced by Liu et al. (2024a) contains 50 continuous recordings, each represented by an $N \times 34$ matrix, where N is the number of samples in that recording. Columns 1–32 contain the recorded signal channels from the acquisition protocol, column 33 contains sample time points, and column 34 contains sparse event markers. Marker 1 denotes left-hand motor imagery onset, marker 2 denotes right-hand motor imagery onset, and marker 3 terminates the current trial and begins rest. Each nonzero marker occupies one sample. Zeros are ordinary continuous samples without an event trigger and are not treated as independently labeled guidance samples.

We map marker 1 to class 0 and marker 2 to class 1. Their respective counts are 1842 and 1951. One recording has no marker 1 or 2 and is excluded, leaving 49 participants and 3793 labeled epochs. The fixed participant-disjoint partition contains 35 training, 7 validation, and 7 test participants, corresponding to 2728, 551, and 514 epochs. Every recording from one participant remains in one partition.

Continuous signals are filtered with a fourth-order zero-phase Butterworth bandpass from 4 to 45 Hz. Each event onset anchors the half-open interval from 0 to 4.5 s, producing an input of 32×2250 samples at 500 Hz. Filtering is applied before epoch extraction. Each epoch is normalized per signal using the median and median absolute deviation, then clipped to $[-20, 20]$. NeuroIntentFormer and the task-specific baselines use this same 32-signal, 4.5 s input protocol. Electrode identities and physical units are not encoded in the recording matrices and are not inferred in this study.

A.2 OPENBMI

OpenBMI contains 54 participants and two sessions per participant. We use all 62 EEG channels and the 4 s left- and right-hand motor imagery interval. Each continuous run is filtered from 4 to 45 Hz, resampled from 1000 to 500 Hz, and normalized with the same per-signal robust procedure. The fixed partition contains 38 training, 8 validation, and 8 test participants. Both sessions and all runs from each participant remain in one partition. This produces 15200 training, 3200 validation, and 3200 test epochs.

B MODEL AND TRAINING SPECIFICATION

B.1 ARCHITECTURE

Table 3 follows the stroke model from the recorded signal epoch to the left/right logits. It lists each stage's configuration and output shape; the learning objective is specified below.

For localized source k, the center and width in Equation 1 are parameterized as $\mu_k = \mathrm{sigmoid}(a_k)$ and $\sigma_k = \sigma_{\min} + (1 - 2\sigma_{\min})\,\mathrm{sigmoid}(b_k)$, where a_k and b_k are unconstrained learned scalars. Here, $\sigma_{\min}$ is the minimum allowed width, τ controls profile sharpness, and ρ preserves a minimum contribution outside the emphasized interval. The global source receives the complete token sequence. In the stroke configuration, $\sigma_{\min} = 0.03$, $\tau = 0.85$, and $\rho = 0.05$. No hard temporal crop or ordering constraint is used.

Each source uses four learnable prototypes. Two summarize class-oriented evidence and two summarize global evidence. The source representation averages the class and global summaries, while the source logits combine class-prototype scores with a weak auxiliary prediction. The router receives four 128-dimensional source representations, four pairs of source logits, and one 128-dimensional encoding of an eight-dimensional signal descriptor. The signal descriptor summarizes coarse temporal energy, variation across input index groups, retained signal ratio, and global scale.

The 20-dimensional readout input is a deterministic summary of the same prediction process. Let p^R denote the right-class probability, $\mathrm{conf}(p) = \max_j p_j$, and $\mathrm{ent}(p) = -\sum_j p_j \log p_j$. The binary-class margin is $|p^R - p^L|$, and source disagreement is the standard deviation of $\{p_k^R\}_{k=1}^K$.

Table 3: NeuroIntentFormer architecture for the stroke experiment, with 1,903,200 trainable parameters. Shapes omit the batch dimension. All source encoders use hidden dimension 128, two attention layers, four heads, FFN dimension 512, dropout 0.24, and four learnable prototypes per source.

Stage	Configuration	Output
Shared representation		
Input	32 signals, 2250 samples, 500 Hz	32×2250
Convolutional stem	Kernels 64/16; $F_1 = 32$, $F_2 = 64$; spatial multiplier 2; two temporal pools of 4	140×64
Token projection	Linear $64 \to 128$, LayerNorm, GELU; learned positions	140×128
Availability encoder	Mask + retained ratio; MLP $33 \to 128 \to 128$	128
Temporal evidence sources		
Soft profiles	Three learned masks, one full-window mask; $\tau = 0.85$, floor 0.05	$4 \times 140 \times 128$
Early source	Depthwise ($k = 5$) + pointwise convolution + Transformer	$z_1 : 128,\ \ell_1 : 2$
Middle/late sources	Separate Transformers, $k \in \{2, 3\}$	$z_k : 128,\ \ell_k : 2$
Global source	Prototype queries, full-sequence cross-attention	$z_4 : 128,\ \ell_4 : 2$
Evidence-conditioned decision		
Signal context	Eight statistics; MLP $8 \to 128 \to 128$	$c : 128$
Router	Source latents, logits, context; MLP $648 \to 128 \to 4$, softmax	$g : 4$
Weighted logits	$\ell_{\text{agg}} = \sum_{k=1}^{4} g_k \ell_k$	2
Evidence readout	LayerNorm; MLP $20 \to 64 \to 2$; dropout 0.08	$R(e) : 2$
Final logits	$\ell_{\text{agg}} + \alpha R(e), \quad \alpha = 0.45$	2

Table 4: Notation used throughout the paper. Symbols are grouped according to the computation from input signals to the final decision.

Symbol	Meaning	Role
X	Input signal matrix	$C \times T$ epoch
C, T	Number of input signals and samples	Input dimensions
m	Binary signal-availability mask	Missing-signal condition
$\widetilde{X}, \odot$	Masked input and element-wise product	Available-signal input
H, H_k	Shared and source-specific token sequences	Evidence construction
P, d	Token count and hidden dimension	Representation size
M_k	Soft temporal profile for source k	Token reweighting
K	Number of evidence sources	Source index range
z_k, ℓ_k	Source latent and class logits	Source evidence
$F_{\text{shared}}, F_m, E_k$	Shared encoder, mask encoder, and source encoder	Representation functions
q_m	Embedded signal-availability condition	Source conditioning
$\mathcal{E}$	Set of all (z_k, ℓ_k) pairs	Router evidence input
r, F_{state}, c	Raw signal descriptor, context encoder, and encoded signal context	Router conditioning
$G, \text{MLP}_{\text{route}}$	Aggregation function and its router network	Weight generation
g_k, g	Source weight and weight vector	Aggregation rule
$\ell_{\text{agg}}, p_{\text{agg}}$	Aggregated source logits and probabilities	Weighted source decision
ϕ, e	Deterministic evidence construction and its vector	Readout input
$R, R(e)$	Evidence readout network and its residual logits	Decision refinement
α	Readout scale	Residual strength
$\ell, [\ell]_j$	Final logit vector and its class-j component	Final decision scores
μ_k, σ_k	Temporal-mask center and width	Soft time emphasis
a_k, b_k	Unconstrained learned mask parameters	Center and width parameterization
$\sigma_{\min}, \tau, \rho$	Minimum width, temperature, and mask floor	Temporal-profile hyperparameters
p_k	Source-k probability vector	Evidence summary
f, cd, ts	Full, channel-drop, and time-shift views	Training views
$\ell_v, p_v, \mathcal{L}_{\text{CE}}^{(v)}$	Final logits, probabilities, and cross-entropy loss for view v	Multi-view supervision
$L_{\text{cls}}, L_{\text{con}}, L_{\text{reg}}$	Classification, consistency, and regularization losses	Learning objective
$L_{\text{var}}, \mathcal{R}_v$	Latent variance and view-specific source/router regularization	Auxiliary objectives
$\lambda_\bullet$	Scalar loss coefficients	Values stated below equations
$\mathcal{L}_{\text{CE}}, D_{\text{KL}}, B, \mathcal{H}$	Cross-entropy, KL divergence, Brier loss, entropy	Loss terms
$y, \hat{y}, j$	Target label, predicted label, and class index	Class 0 is left and class 1 is right

Table 5: Ordered input features of the evidence readout. Early, middle, late, and global correspond to source indices 1 through 4. Each source contributes its routing weight, right-class probability, and confidence.

Indices	Group	Ordered contents
1–4	Aggregated posterior	p^R_{agg}, confidence, entropy, margin
5	Source disagreement	Standard deviation of the four source right-class probabilities
6–17	Per-source evidence	$(g_k, p^R_k, \text{conf}(p_k))$ for early, middle, late, and global, in that order
18	Signed contrast	$p^R_{\text{late}} - p^R_{\text{global}}$
19	Absolute contrast	$\lvert p^R_{\text{late}} - p^R_{\text{global}} \rvert$
20	Absolute contrast	$\lvert p^R_{\text{early}} - p^R_{\text{middle}} \rvert$
Total		20 deterministic features

Table 5 gives the exact feature order. No additional normalization is applied before the readout's LayerNorm. The features remain in the computation graph, and the readout is trained jointly rather than fitted as a separate classifier.

B.2 Learning Objective

Let f, cd, and ts denote the full, channel-drop, and time-shift views. For a batch of N trials, $p_{v,i} = \text{softmax}(\ell_{v,i})$ is the final class-probability vector for trial i under view v. The patient training objective is

$$L = L_{\text{cls}} + L_{\text{con}} + L_{\text{reg}}. \tag{2}$$

Classification. Supervision is applied to the full and both perturbed views,

$$L_{\text{cls}} = \mathcal{L}^{(f)}_{\text{CE}} + \lambda_{\text{cd}}\mathcal{L}^{(\text{cd})}_{\text{CE}} + \lambda_{\text{ts}}\mathcal{L}^{(\text{ts})}_{\text{CE}}. \tag{3}$$

Each $\mathcal{L}^{(v)}_{\text{CE}}$ is the batch-reduced cross-entropy of the final logits, using training-set class weights and label smoothing 0.02.

View consistency. The full-view prediction supplies the detached reference distribution,

$$L_{\text{con}} = \lambda_{\text{con,cd}} D_{\text{KL}}(\text{sg}(p_f) \| p_{\text{cd}}) + \lambda_{\text{con,ts}} D_{\text{KL}}(\text{sg}(p_f) \| p_{\text{ts}}) \,. \tag{4}$$

Here, sg stops gradients through the reference. For either perturbed view v, the divergence is $N^{-1}\sum_{i=1}^{N}\sum_{j=0}^{1} p_{f,i,j}\log(p_{f,i,j}/p_{v,i,j})$. Thus the KL direction is from the detached full-view distribution to the perturbed prediction, with a sum over classes and a mean over trials.

Regularization. Latent variability, probability supervision, and source/router regularization contribute

$$\begin{aligned} L_{\text{reg}} = \lambda_{\text{var}} L_{\text{var}} + \lambda_B \left(B_f + \lambda_{B,\text{cd}} B_{\text{cd}}\right) \\ + \mathcal{R}_f + \lambda_{\mathcal{R},\text{cd}}\mathcal{R}_{\text{cd}} + \lambda_{\mathcal{R},\text{ts}}\mathcal{R}_{\text{ts}}. \end{aligned} \tag{5}$$

The Brier term uses the unsmoothed one-hot target,

$$B_v = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=0}^{1}\left(p_{v,i,j} - \mathbf{1}[y_i = j]\right)^2, \qquad v \in \{f, \text{cd}\}. \tag{6}$$

It is evaluated on the full and channel-drop views only. For the variance term, each row of Z_v is the fused latent $z_{v,i} = \sum_{k=1}^{K} g_{v,i,k} z_{v,i,k}$ for trial i under view v. Stack the three batches as $Z = [Z_f; Z_{\text{cd}}; Z_{\text{ts}}] \in \mathbb{R}^{3N \times d_z}$, where d_z is the latent dimension. Then

$$L_{\text{var}} = \frac{1}{d_z}\sum_{j=1}^{d_z}\max\left(0, \gamma - \sqrt{\text{Var}(Z_{:,j}) + \epsilon}\right), \qquad \gamma = 0.5, \quad \epsilon = 10^{-4}. \tag{7}$$

The variance is computed over the $3N$ stacked entries using the sample-variance denominator $3N - 1$. This penalizes latent standard deviations below the specified floor.

For each view v, source and router regularization is

$$\mathcal{R}_v = \lambda_{\text{aux}} \frac{1}{K} \sum_{k=1}^{K} \mathcal{L}_{\text{CE}}(\ell_{v,k}, y) - \lambda_{\text{ent}} \mathcal{H}(g_v) + \lambda_{\text{late}} \overline{g}_{v,\text{late}}. \tag{8}$$

The auxiliary source loss encourages discriminative source predictions, the entropy term encourages dispersed routing weights, and the final term penalizes the mean late-source weight. The auxiliary cross-entropy uses label smoothing 0.02 without class weights. Routing entropy is the batch mean $\mathcal{H}(g_v) = -N^{-1} \sum_{i=1}^{N} \sum_{k=1}^{K} g_{v,i,k} \log g_{v,i,k}$, and $\overline{g}_{v,\text{late}} = N^{-1} \sum_{i=1}^{N} g_{v,i,\text{late}}$ is the mean late-source weight.

Here, $\ell_{v,k}$ denotes the batch of source-k logits under view v, and y denotes its target labels. Each λ is a scalar coefficient. The coefficient values are $\lambda_{\text{cd}} = 0.10$, $\lambda_{\text{ts}} = 0.06$, $\lambda_{\text{con,cd}} = 0.06$, $\lambda_{\text{con,ts}} = 0.04$, $\lambda_{\text{var}} = 0.0005$, $\lambda_B = 0.03$, $\lambda_{B,\text{cd}} = 0.5$, $\lambda_{\mathcal{R},\text{cd}} = 0.70$, $\lambda_{\mathcal{R},\text{ts}} = 0.15$, $\lambda_{\text{aux}} = 0.14$, $\lambda_{\text{ent}} = 0.001$, and $\lambda_{\text{late}} = 0.018$. The gate-prior weight is zero, so no target source distribution is imposed. The channel-drop view retains one of 32, 30, 28, 24, or 20 randomly selected input signals. The time-shift view applies an independently sampled integer displacement from −50 to 50 samples with zero padding.

The patient model uses AdamW with learning rate 2.2×10^{-4}, weight decay 0.018, cosine decay, batch size 32, and gradient clipping at 1.0. Training runs for at most 120 epochs with patience 30 under AMP. The first checkpoint attaining the maximum validation pooled BAcc under argmax is retained.

C Baseline Protocols

Patient comparison protocol. All patient experiments use the fixed 35/7/7 participant partition. Classical and task-specific methods use the same 32 recorded signals over 4.5 s at 500 Hz. Task-specific neural models use suitable architecture-specific objectives and otherwise follow the common validation-pooled BAcc and argmax evaluation protocol. Pretrained models retain the input conversion and adaptation required by their official backbones and are treated as native-protocol references.

Table 6: Initialization, checkpoint selection, and test decisions for the patient comparison.

Method group	Initialization	Selection and test decision
Classical feature pipelines	Train-set fit without pretraining	Native class decision; no epoch selection
Task-specific deep models	Supervised training from scratch	Validation pooled BAcc under argmax; argmax test
Pretrained EEG references	Official pretrained initialization and task adaptation	Validation-based native selection and recorded decision
NeuroIntentFormer	End-to-end supervised training from scratch	Validation pooled BAcc under argmax; argmax test

The task-specific models in Table 1 use argmax evaluation. EEGNet, FBCNet, EEG-Conformer, MSVTNet, and TCFormer share the validation-argmax selection rule. TCFormer (Altaheri et al., 2025) retains its published architecture, adapted to the dataset input and binary classification task. It is trained from scratch for up to 80 epochs with patience 20, without segmentation-and-reconstruction augmentation. Patient BIOT, LaBraM, and CBraMod use validation-selected decision thresholds, while both REVE modes use argmax. These foundation-model rows are native-protocol references. On OpenBMI, all neural checkpoints are selected by validation argmax BAcc, and all test decisions use argmax. The NeuroIntentFormer patient configuration is specified in Appendix B; OpenBMI uses independent within-dataset training.

D Complete Performance Statistics

Tables 7 and 8 supplement the subject means and SDs in Table 1. Subject intervals resample participants with equal weight. Pooled intervals resample participants together with all their trials and then recompute BAcc. Each percentile interval uses 10,000 bootstrap draws. These intervals characterize the held-out participants for fixed trained models rather than repeated-training variability. Macro-F1 and AUROC are computed over all test trials.

Table 7: Supplementary stroke test statistics for seven participants and 514 trials. All values are percentages. The two BAcc intervals use participant-mean and pooled cluster-bootstrap estimands, respectively. Initialization and decision protocols are specified in Appendix C.

Method	Subject 95% CI	Pooled 95% CI	Macro-F1	AUROC
Classical feature methods				
CSP + LDA	[52.87, 61.57]	[52.81, 62.75]	55.31	63.64
CSP + SVM	[52.70, 63.34]	[52.53, 64.34]	58.63	59.63
FBCSP + LDA	[53.61, 59.33]	[53.51, 61.10]	57.40	63.77
Task-specific deep models				
EEGNet	[74.50, 85.45]	[75.30, 85.67]	80.52	89.94
DeepConvNet	[68.48, 84.17]	[71.18, 85.00]	78.94	90.47
FBCNet	[68.64, 81.21]	[68.02, 81.24]	73.87	86.22
EEG-Conformer	[72.72, 86.00]	[73.62, 86.20]	80.39	89.14
ATCNet	[75.64, 86.73]	[75.62, 87.70]	82.37	90.02
CTNet	[73.34, 87.44]	[74.87, 87.98]	81.52	90.17
MSVTNet	[77.53, 86.24]	[78.71, 86.54]	82.91	90.62
TCFormer	[74.50, 84.05]	[75.75, 84.43]	80.41	88.70
Pretrained EEG references				
BIOT	[58.50, 67.98]	[57.48, 66.92]	61.69	67.97
LaBraM	[64.88, 84.36]	[66.70, 85.44]	76.28	83.06
CBraMod	[69.70, 81.09]	[71.07, 81.49]	76.25	84.18
REVE-LP	[71.85, 79.47]	[74.51, 80.24]	77.71	87.00
REVE-LP+LoRA	[69.62, 80.64]	[72.77, 81.05]	77.64	87.20
Proposed method				
NeuroIntentFormer	[81.10, 89.84]	[81.61, 89.91]	85.75	93.41

Macro-F1 and AUROC support the main BAcc comparison through class decisions and score ranking. On OpenBMI, NeuroIntentFormer also achieves the highest Macro-F1 and AUROC among the evaluated methods, complementing its leading BAcc.

E Supplementary Mechanism Evidence

E.1 Matched Router and Readout Configurations

Table 9 specifies the configurations underlying Figure 3. The context-only router repeats the encoded context to match the full router's input width and parameter count.

Table 9 provides the absolute scores behind the interaction analysis, with the complete evidence-conditioned model achieving the highest test BAcc. The pooled interaction is 4.29 pp, compared with the patient-mean interaction of 5.44 pp because the latter gives equal weight to patients. Both summaries favor learning refinement together with evidence-conditioned routing.

The patient-level analysis uses the same seven patients and 514 trials in all four conditions. Its percentile 95% CI uses 10,000 bootstrap draws, resampling patients jointly across the four fixed models. This interval describes participant variation for the selected checkpoints rather than variability across repeated training runs.

Joint training and fixed-checkpoint refinement. Retraining ablations assess refinement during joint learning, whereas a fixed-checkpoint intervention isolates its immediate effect on predictions.

Table 8: Supplementary OpenBMI test statistics for eight participants and 3200 trials (%). The BAcc interval applies to both subject-mean and pooled BAcc, which coincide because every participant has the same number of trials in each class and both intervals use the same participant-bootstrap draws. All test decisions use argmax.

Method	BAcc 95% CI	Macro-F1	AUROC
Classical feature methods			
CSP + LDA	[54.28, 66.94]	59.28	65.49
CSP + SVM	[54.44, 66.75]	59.19	63.80
FBCSP + LDA	[58.69, 76.59]	67.09	73.98
Task-specific deep models			
EEGNet	[72.34, 86.31]	79.13	87.92
DeepConvNet	[74.75, 86.81]	80.67	89.84
FBCNet	[69.62, 82.91]	76.00	84.17
EEG-Conformer	[69.97, 82.59]	76.00	85.36
ATCNet	[72.88, 84.69]	78.90	87.58
CTNet	[72.34, 85.91]	78.92	87.96
MSVTNet	[73.97, 87.22]	80.32	88.69
TCFormer	[74.06, 86.00]	80.00	88.68
Pretrained EEG references			
BIOT	[52.72, 68.25]	59.50	64.93
LaBraM	[64.38, 75.94]	69.94	76.55
CBraMod	[49.72, 52.22]	50.21	50.69
REVE-LP	[65.97, 78.78]	71.84	80.07
REVE-LP+LoRA	[69.53, 79.03]	74.00	80.88
Proposed method			
NeuroIntentFormer	[75.78, 87.97]	82.01	89.96

Table 9: Exact validation and test BAcc (%) for the parameter-matched router–readout experiment. Each checkpoint is selected by validation BAcc.

Router conditioning	Readout	Val BAcc	Test BAcc
Evidence + context	Off	81.12	82.34
Evidence + context	On	84.39	85.95
Context only, matched	Off	83.12	84.87
Context only, matched	On	82.94	84.19

Base and final predictions achieve 86.02% and 85.95% BAcc, respectively. These analyses distinguish learning the decision function from refining its output at inference.

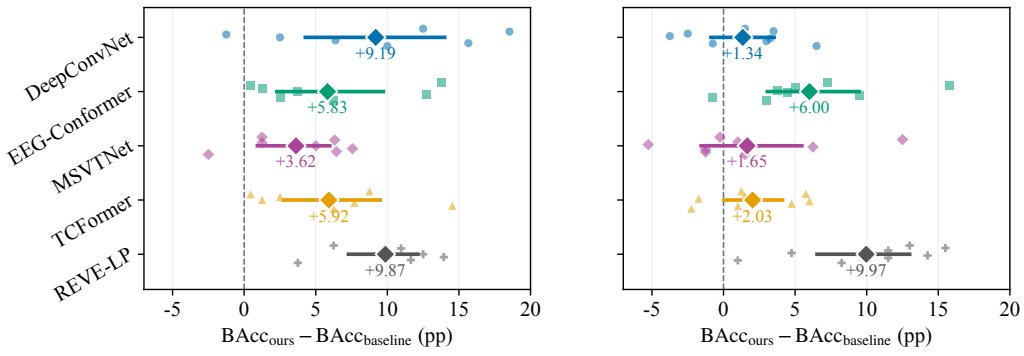
F Supplementary Evaluation Details

F.1 OpenBMI Configuration

The OpenBMI model keeps the same evidence sources, router, readout, hidden dimensions, and objective terms. Its input contains 62 signals and 2000 samples. Channel retention choices are scaled to 62, 58, 54, 47, and 39 signals. Training uses batch size 64 and a late-gate penalty of 0.005. All other optimization settings match the patient model.

F.2 Participant-Level Paired Comparisons

Figure A1 presents paired BAcc differences on the same held-out participants within each dataset. On Stroke, NeuroIntentFormer improves performance for six of seven patients relative to MSVTNet. On OpenBMI, positive mean gains across all five comparisons complement the aggregate results. The participant-level distributions show how the gains vary across individuals.



(a) Stroke paired gains **(b) OpenBMI paired gains**

Figure A1: Participant-paired gains on Stroke in panel a and OpenBMI in panel b. Values are NeuroIntentFormer minus each baseline in BAcc, expressed in pp, with identical scales and method encodings. Points represent participants, labeled diamonds are means, and intervals are paired participant-bootstrap 95% CIs. Positive values favor NeuroIntentFormer.

F.3 Efficiency Protocol

Table 10: Stroke model-forward efficiency. Input shapes omit the batch dimension. Params M is active deployment parameters in millions. B1 is median latency, B32/s is trials per second at batch 32, and MiB is peak allocated memory at batch 32.

Method	Input	Prec.	Params M	GFLOPs	B1 ms	B32/s	MiB
EEGNet	32×2250	FP16	0.008	0.154	0.34	35402	153.9
DeepConvNet	32×2250	FP16	0.294	0.201	0.41	26722	267.9
FBCNet	32×2250	FP16	0.020	0.043	1.24	9610	183.0
EEG-Conformer	32×2250	FP16	1.616	0.412	1.52	15867	367.4
ATCNet	32×2250	FP16	0.114	0.168	4.41	7208	149.9
CTNet	32×2250	FP16	0.114	0.211	1.72	17952	198.4
MSVTNet	1×32×2250	FP16	0.073	0.319	1.39	12632	89.3
TCFormer	32×2250	FP16	0.080	0.576	2.47	5735	853.0
BIOT	16×800	FP16	3.186	0.726	4.02	7709	59.1
LaBraM	32×4×200	FP16	5.820	1.650	4.58	6857	98.9
CBraMod	32×4×200	FP16	4.885	1.274	7.01	4509	64.9
REVE-LP	32×800	FP16	69.388	17.719	8.51	3498	490.8
REVE-LP+LoRA	32×800	FP16	70.469	18.008	10.82	2762	356.6
NeuroIntentFormer	32×2250	FP16	1.903	0.566	5.80	5156	310.6

Efficiency is measured on an RTX 4090 D without concurrent workloads or model compilation, using 100 warmup iterations and five blocks of 200 timed iterations. Batch-1 latency is the median model-forward time. Batch-32 throughput is computed from mean latency, and peak allocated memory is measured independently for each model. AMP FP16 is used when its BAcc drop relative to FP32 is at most 0.002, otherwise FP32 is reported separately. Forward measurements exclude data loading, offline preprocessing, host-to-device transfer, and actuator control.

Parameters count the deployment model, including the encoder for linear probing. FLOPs are estimated from FP32 matrix operations and convolutions with two operations per fused multiply-add. Fused attention, filtering, FFT, reductions, and other unsupported operations may not be counted fully. Foundation models use the same input transformations as in their classification evaluations.

Table 10 reports the efficiency of the selected Stroke models under this measurement protocol. MSVTNet excludes training-only branch heads during inference while preserving its final classification output.

The table separates model size from measured inference cost. NeuroIntentFormer combines fewer parameters than REVE-LP with higher Stroke BAcc, while compact task-specific models offer lower forward latency.

G Offline Decoding with a Review Option

This analysis considers the 514 held-out stroke trials known to contain left- or right-hand MI. The decoder predicts a hand and can withhold a low-confidence prediction for hypothetical review. Confidence is $|p(\text{right}) - 0.5|$. Each model's thresholds are selected on validation data at target automatic-decision rates of 70, 80, 90, and 100% and applied unchanged to test trials. Actual test acceptance rates differ because thresholds are fixed on validation data.

Figure A2a reports BAcc among accepted predictions, while Figure A2b counts incorrect automatic outputs per 100 original trials. The curves separate accepted prediction quality from residual errors and review workload. The comparison uses the same validation and test trials for NeuroIntentFormer, EEGNet, MSVTNet, EEG-Conformer, and TCFormer.

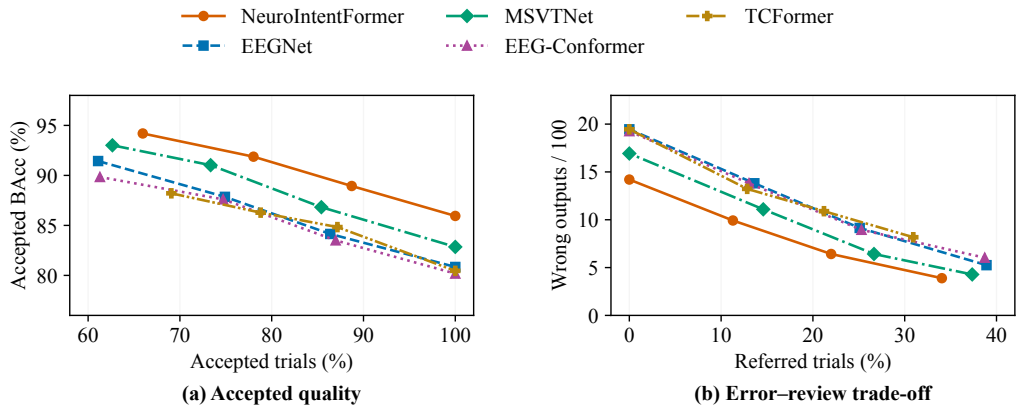


Figure A2: Offline left/right MI decoding with a review option. Panel a reports BAcc among accepted predictions versus test acceptance rate. Panel b reports incorrect automatic outputs per 100 original trials versus the fraction referred for review. Points use validation-selected thresholds at 70, 80, 90, and 100% target acceptance. Review flags withhold predictions rather than correct them.

At the 80% validation target, NeuroIntentFormer accepts 78.0% of test trials with 91.9% BAcc, compared with 73.3% and 91.0% for MSVTNet. Both produce 6.42 incorrect outputs per 100 original trials, so NeuroIntentFormer retains more correct predictions for automatic output while referring fewer trials for review. This shows how its classification advantage is retained when lower-confidence predictions are withheld.