Predicting Only from Selected Evidence: A Tempered Product-of-Experts Bottleneck for Auditable EEG Diagnosis

Yinghao WANG[1], Shujian YU[2,*], Duc Han LE[1]
Zhikai YU[3], Changming WANG[3,*], Van-Tam NGUYEN[1]

[1]LTCI, Télécom Paris, Institut Polytechnique de Paris, Palaiseau, France
[2]Department of Artificial Intelligence, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
[3]Beijing Preclinical Testing Platform for Brain-Computer Interface Products,
Zhongguancun Industrial Park, Beijing, China
[*]*Co-corresponding authors.*
Shujian Yu: s.yu3@vu.nl; Changming Wang: superwcm@163.com

Abstract

Pretrained EEG backbones improve transfer performance, yet downstream diagnostic heads remain difficult to audit: predictions are typically computed from unrestricted hidden states, while explanations are generated only after the decision. We introduce tPoE-EIB, an *evidence-information bottleneck* head that adapts frozen EEG backbones under an evidence-only prediction constraint. tPoE-EIB extracts gated temporal and channel summaries, maps them to Gaussian experts over a shared latent variable, and integrates them via a closed-form tempered product-of-experts (tPoE) posterior. Because the classifier conditions exclusively on this latent variable, the forward decision path is explicit and rate-limited by the expected posterior KL divergence, while axis-specific tempering mitigates overconfident fusion of correlated evidence views. We evaluate tPoE-EIB across six diagnostic settings—event-type classification, abnormality detection, seizure detection, cognitive-decline staging, depression screening, and cerebrovascular-disease classification—spanning public benchmarks and clinical cohorts with sparse and dense montages. Across frozen backbones, tPoE-EIB preserves competitive balanced accuracy while exposing both evidence selection and posterior fusion to direct intervention. Using five complementary audits—insertion–deletion, expert-drop necessity, gate causality, posterior reliance, and expert disagreement—we show that tPoE-EIB separates aggregate axis necessity from per-channel ranking stability, resolves causal gate regimes, and tests when diagnostic predictions recover or depart from canonical EEG biomarkers.

1 Introduction

Deep EEG modeling has shifted from compact task-specific networks toward pretrained foundation-style backbones. Architectures such as EEGNet and EEG Conformer introduced EEG-specific convolutional and attention biases [1, 2], while recent models such as LaBraM, CBraMod, and CodeBrain learn reusable representations through large-scale pretraining and structured temporal–channel modeling [3–5]. As representation quality improves, however, the downstream diagnostic head becomes the primary bottleneck for auditability. We use *auditability* in a concrete, operational sense: a diagnostic head is auditable when its forward evidence path is explicit enough that both evidence selection and evidence integration can be directly intervened upon and measured. A frozen backbone followed by standard pooling and a task head may perform well, but the classifier remains free to exploit any predictive regularity in the unrestricted hidden state—including nuisance variation, subject signatures, or artifact-correlated cues—while explanations are produced only after the prediction has been fixed.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Post-hoc explainers such as Integrated Gradients, Dynamask, and TIMING identify salient inputs after a prediction [6–8], but they do not constrain the information used to compute that prediction. This separation between explanation generation and model optimization creates a faithfulness gap: a saliency map need not reflect the computation that actually drove the decision [9]. In high-stakes domains, this gap has motivated architectures whose interpretability constraints are built into the decision path itself [10]. The requirement is particularly natural in EEG diagnosis, where clinicians inspect evidence along two primary axes: *when* a relevant waveform occurs and *where* on the scalp it appears. A model that spreads support diffusely across time, or concentrates weight on anatomically implausible channels, may remain accurate while offering little mechanistic transparency.

Adding a learnable gate alone is not sufficient. A sparse or augmentation-stable gate is not automatically faithful: a selector that assigns the same low-mass pattern to every input can satisfy sparsity and consistency penalties while selecting no sample-specific evidence. Beyond selecting temporal and channel support, an adaptation head must therefore restrict what the selected evidence can carry to the classifier and expose the resulting pathway to intervention-based audits [11].

We propose **tPoE-EIB**, a downstream adaptation head that places evidence selection and probabilistic fusion directly on the forward decision path (Fig. 1). Given a frozen backbone representation, tPoE-EIB learns marginal temporal and channel gates, pools the gated support into axis-specific evidence summaries, and maps those summaries to Gaussian experts over a shared latent variable. A tempered product-of-experts posterior integrates the experts, and the classifier receives only a sample from this latent during training and its posterior mean at evaluation. We call this structural restriction an *evidence-only prediction constraint*: after the frozen backbone, the classifier cannot access pooled hidden states or concatenated evidence summaries, and any feature that influences the prediction must pass through the gates, the evidence summaries, and the tPoE posterior. TheKL rate term controls how much selected evidence reaches the classifier, while the audit suite in Sec. 4 tests how each stage of the pathway responds to perturbation and whether the selected channels align with established EEG biomarkers.

We evaluate tPoE-EIB on a frozen CodeBrain backbone, with cross-backbone validation on CBraMod and LaBraM, across six EEG diagnostic settings: event-type classification, abnormality detection, seizure detection, cognitive-decline staging, depression screening, and cerebrovascular-disease classification. The evaluation spans public benchmarks and in-house clinical cohorts, binary screening and four-way staging regimes, and sparse (16–22) to dense (58–129) channel montages. Code is available at `https://github.com/W-Yinghao/tPoe_EIB`.

Our contributions are as follows:

- **Evidence-only EEG adaptation.** We formulate downstream adaptation of pretrained EEG backbones under an *evidence-only prediction constraint*: the classifier has no bypass to unrestricted backbone states and predicts exclusively from gated, rate-limited evidence along the forward graph $H \to E \to Z \to \hat{Y}$.
- **Tempered PoE evidence bottleneck.** We instantiate this constraint with a tempered product-of-experts bottleneck in which temporal and channel summaries parameterize Gaussian experts over a shared latent variable and are fused via precision-weighted posterior aggregation rather than feature concatenation.
- **Theory and two-stage audit protocol.** We derive the closed-form tPoE posterior, bound the evidence-IB objective via a variational KL rate penalty, and motivate tempering through an idealized Gaussian calibration model for correlated experts. We pair the architecture with five complementary audits targeting gate faithfulness, expert necessity, gate causality, posterior reliance, and expert disagreement.
- **Empirical validation and mechanistic insights.** Across public benchmarks, clinical cohorts, and three frozen EEG backbones, tPoE-EIB remains competitive with unconstrained probes while enabling stage-wise interventions that separate aggregate axis necessity from per-channel ranking stability and test alignment with regional EEG priors.

2 Related Work

Pretrained EEG backbones and adaptation heads. Compact task-specific networks such as EEGNet and EEG Conformer established convolutional and attention biases for multichannel EEG decoding [1, 2]. Recent foundation-style models shift the focus to reusable representations: LaBraM

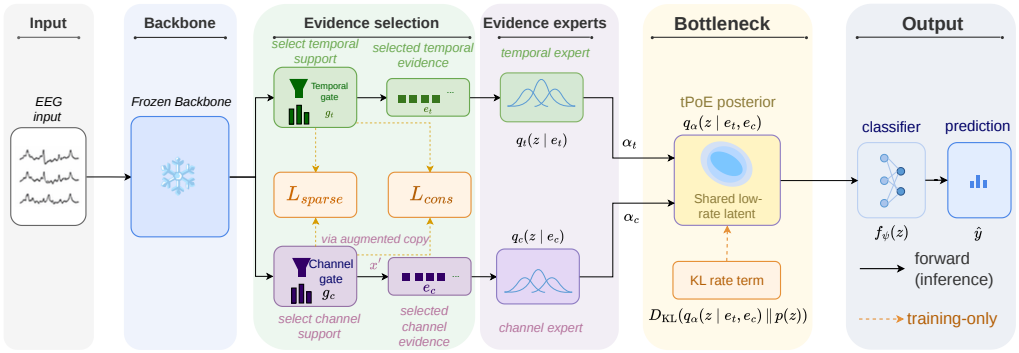


Figure 1: Architecture of tPoE-EIB. Solid arrows denote the inference path; orange dashed arrows denote training-only regularization. A frozen EEG backbone produces a patch-grid representation H. Temporal and channel gates g_t, g_c select evidence summaries e_t and e_c, which two Gaussian experts map to beliefs $q_t(z \mid e_t)$ and $q_c(z \mid e_c)$ over a shared latent variable z. These are integrated by a tempered product-of-experts posterior $q_\alpha(z \mid e_t, e_c) \propto p(z)\, q_t^{\alpha_t} q_c^{\alpha_c}$ with axis-specific temperatures $\alpha_t, \alpha_c \in [0, 1]$, and the classifier f_ψ observes only the latent z. Two selection-stage regularizers shape the gates: $\mathcal{L}_{\text{sparse}}$ penalizes gate mass to prevent trivial all-pass solutions, and $\mathcal{L}_{\text{cons}}$ encourages stable gates across a lightly augmented copy x' (augmented branch omitted for clarity). By construction, the classifier never receives concatenated evidence $[e_t; e_c]$ or unrestricted hidden states H; all classification information flows through the rate-limited latent z.

learns generic neural representations through masked tokenization, CBraMod models criss-cross temporal–channel dependencies across heterogeneous formats, CodeBrain decouples time–frequency tokenization from multi-scale representation learning, and concurrent work explores seq2seq pretraining, unified representation learning, and spatial parameter-efficient adaptation such as EGA [3–5, 12, 13]. While parameter-efficient adapters tune internal backbone weights, standard downstream heads—linear probes, MLPs, and attention pooling—still feed unrestricted hidden states to the classifier. We address this final interface by constraining how frozen backbone representations enter the diagnostic decision head.

EEG interpretability and intrinsic constraints. EEG interpretability often relies on attention maps or post-hoc attribution. Integrated Gradients, Dynamask, and TIMING identify salient time–channel regions after prediction, but do not prevent the classifier from relying on features outside the displayed attribution [6–9]. Intrinsic approaches instead constrain the forward computation: Right for the Right Reasons regularizes input gradients, Concept Bottleneck Models route predictions through intermediate concepts, and rationale models select sufficient input subsets [14–20]. Related designs in EEG and multimodal learning include discrete Gumbel-Softmax channel selection [21, 22], sparse dynamic graph learning for dementia interpretation in EEG-SeeGraph [23], and evidence-grounded search in Tree-of-Evidence [24]. Complementing discrete sensor selection and graph-topology explainers, tPoE-EIB enforces a dual-axis (temporal and channel) evidence interface coupled with a rate-limited shared posterior. Because learned selectors can still degenerate into input-independent templates [11], we audit both the selection gates and the posterior that integrates their summaries.

Information bottlenecks and posterior fusion. The Information Bottleneck (IB) principle compresses representations while preserving label-relevant information, made tractable in deep networks via variational KL bounds [25–27]. Multi-view and multimodal extensions study how multiple sources should be compressed and fused [28–32]. Simple feature concatenation obscures which axis drives a given prediction. Product-of-experts (PoE) fusion provides an explicit precision-weighted aggregation rule, but standard PoE treats experts as conditionally independent. Because temporal and channel summaries are correlated projections of the same EEG representation, we temper the PoE posterior to discount duplicated support across the two axes.

3 tPoE-EIB

3.1 Problem setup and evidence-only prediction constraint

Let $x \in \mathbb{R}^{C \times S \times P}$ denote an EEG window with C channels, S temporal segments, and P samples per segment. A frozen pretrained backbone B_θ maps x to a patch-grid representation

$H = B_\theta(x) \in \mathbb{R}^{C\times S\times D}$. Standard adaptation pools H and trains a classifier on the resulting unrestricted representation. tPoE-EIB instead inserts a structured evidence interface between the backbone and classifier:

$$H = B_\theta(x), \qquad (g_t, g_c, z) = E_\phi(H), \qquad \hat{y} = f_\psi(z), \tag{1}$$

where $g_t \in [0,1]^S$ and $g_c \in [0,1]^C$ are temporal and channel gates and $z \in \mathbb{R}^{d_z}$ is the decision latent. The classifier f_ψ observes only z: neither the gates nor the intermediate summaries e_t, e_c defined below are exposed to f_ψ. We refer to this forward-graph structure ($H \to E = (e_t, e_c) \to Z \to \hat{Y}$) as the *evidence-only prediction constraint*. It guarantees forward-path faithfulness to the selected evidence interface: any signal that affects $\hat{y}$ must pass through the gates, the axis summaries, and the posterior bottleneck. A compact glossary of notation is provided in Appendix A.

3.2 Evidence-information bottleneck objective and training surrogate

Let $E = (e_t, e_c)$ collect the temporal and channel evidence summaries extracted from H (Sec. 3.3), and let $q_\alpha(z \mid E)$ denote the stochastic posterior encoder mapping E to z (Sec. 3.4). We optimize the *evidence-information bottleneck* objective

$$\mathcal{J}_{\mathrm{EIB}} = I(Y;Z) - \beta\, I(E;Z), \tag{2}$$

which replaces the raw input in the classic IB formulation [25, 26] with the selected evidence E.

Why compress E rather than H? Because the classifier never accesses the full hidden state H directly, the implemented decision graph is $H \to E \to Z \to \hat{Y}$. Compressing $I(E;Z)$ regulates information flow along the exact pathway available to the classifier. Under the deterministic map $H \mapsto E(H)$, penalizing $I(E;Z)$ coincides with the path-wise hidden-state rate and upper-bounds $I(\hat{Y};X)$ by data processing, while making explicit that the bottleneck governs the selected evidence summaries rather than the entire frozen backbone tensor.

Proposition 1 (Variational lower bound on $\mathcal{J}_{\mathrm{EIB}}$)**.** *For any variational decoder $r_\psi(y \mid z)$ and prior $p(z) = \mathcal{N}(0, I)$,*

$$\mathcal{J}_{\mathrm{EIB}} \;\geq\; \mathbb{E}_{p(x,y)}\mathbb{E}_{q_\alpha(z|E)}[\log r_\psi(y \mid z)] - \beta\, \mathbb{E}_{p(x)} D_{\mathrm{KL}}(q_\alpha(z \mid E)\|p(z)) + H(Y). \tag{3}$$

The proof is given in Appendix C.3. Maximizing Eq. (3) is equivalent to minimizing the classification loss $\mathcal{L}_{\mathrm{cls}} = -\mathbb{E}[\log r_\psi(y \mid z)]$ plus the KL rate penalty $\mathcal{L}_{\mathrm{KL}} = \mathbb{E}_{p(x)} D_{\mathrm{KL}}(q_\alpha(z \mid E)\|p(z))$. Combining this bound with two selection-stage gate regularizers yields the training objective

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{cls}} + \beta_{\mathrm{eff}}(t)\, \mathcal{L}_{\mathrm{KL}} + \eta\, \mathcal{L}_{\mathrm{sparse}} + \gamma\, \mathcal{L}_{\mathrm{cons}}, \tag{4}$$

with $\mathcal{L}_{\mathrm{sparse}}, \mathcal{L}_{\mathrm{cons}}$ defined in Eqs. (6)–(7) and $\beta_{\mathrm{eff}}(t) = \beta \min(1, t/T_{\mathrm{warm}})$ linearly warming up the target coefficient β over the first T_{warm} epochs. By the data-processing inequality on $E \to Z \to \hat{Y}$, $I(\hat{Y};X) \leq \mathbb{E}_{p(E)} D_{\mathrm{KL}}(q_\alpha(z \mid E)\|p(z))$, so the expected KL divergence upper-bounds the input information transmitted along the stochastic training path (Corollary 1 in Appendix B).

3.3 Dual-axis evidence selection

From H, tPoE-EIB forms marginal temporal and channel views by averaging across the complementary dimension, $H_t[s] = \frac{1}{C}\sum_c H_{c,s,:}$ and $H_c[j] = \frac{1}{S}\sum_s H_{j,s,:}$, yielding $H_t \in \mathbb{R}^{S\times D}$ and $H_c \in \mathbb{R}^{C\times D}$. Two lightweight MLPs predict continuous sigmoid gates $g_t = \sigma(\mathrm{MLP}_t(H_t))$ and $g_c = \sigma(\mathrm{MLP}_c(H_c))$, broadcast across feature dimensions, to produce the gated summaries

$$e_t = \frac{1}{S}\sum_{s=1}^{S} g_t[s] H_t[s], \qquad e_c = \frac{1}{C}\sum_{j=1}^{C} g_c[j] H_c[j]. \tag{5}$$

This marginal factorization provides a compact $(S + C)$ evidence interface aligned with clinical inspection axes rather than parameterizing a full $C \times S$ joint mask; because each branch averages over the complementary axis prior to gating, the no-bypass guarantee operates at the summary interface $E = (e_t, e_c)$. Whereas an unrestricted concatenation head feeds $[e_t; e_c]$ directly to f_ψ, tPoE-EIB routes both summaries through the shared posterior bottleneck. Two regularizers shape the gates during training:

$$\mathcal{L}_{\mathrm{sparse}} = \mathbb{E}_{p(x)}\left[\tfrac{1}{2}\left(\tfrac{1}{S}\textstyle\sum_s g_t(x)[s] + \tfrac{1}{C}\sum_j g_c(x)[j]\right)\right], \tag{6}$$

$$\mathcal{L}_{\mathrm{cons}} = \mathbb{E}_{p(x)}\left[\tfrac{1}{2}\left(\tfrac{1}{S}\|g_t(x) - g_t(x')\|_2^2 + \tfrac{1}{C}\|g_c(x) - g_c(x')\|_2^2\right)\right], \tag{7}$$

where $x' = \mathrm{Aug}(x)$ is a lightly augmented view of x. $\mathcal{L}_{\mathrm{sparse}}$ penalizes total gate mass, while the posterior KL penalty in Eq. (4) limits the information rate reaching the classifier. Lemma 1 (Appendix B) shows that $\mathcal{L}_{\mathrm{sparse}}$ and $\mathcal{L}_{\mathrm{cons}}$ alone can be satisfied by a constant gate, motivating the rate-limited bottleneck and the intervention audits in Sec. 4.

3.4 Tempered Product-of-Experts posterior

The temporal and channel summaries parameterize two diagonal-Gaussian evidence experts over the shared latent variable $z \in \mathbb{R}^{d_z}$:

$$q_t(z \mid e_t) = \mathcal{N}(\mu_t(e_t), \mathrm{diag}(\sigma_t^2(e_t))), \qquad q_c(z \mid e_c) = \mathcal{N}(\mu_c(e_c), \mathrm{diag}(\sigma_c^2(e_c))). \tag{8}$$

Both experts share the prior $p(z) = \mathcal{N}(0, I)$. Because e_t and e_c are projections of the same EEG representation H, standard PoE fusion can double-count correlated evidence and overestimate posterior precision; Proposition 2 formalizes this effect in a Gaussian calibration model. We therefore integrate the experts via a *tempered* product-of-experts posterior:

$$q_\alpha(z \mid e_t, e_c) \propto p(z)\, q_t(z \mid e_t)^{\alpha_t} q_c(z \mid e_c)^{\alpha_c}, \qquad \alpha_t, \alpha_c \in [0, 1]. \tag{9}$$

We use learnable $(\alpha_t, \alpha_c) \in [0, 1]^2$ as the task-adaptive default, $\alpha_t = \alpha_c = 1$ as the untempered standard PoE ablation, and fixed $\alpha_t = \alpha_c = 0.5$ as the conservative $\rho \to 1$ reference (Appendix D).

Proposition 2 (Tempering corrects overconfidence under correlated experts)**.** *Let $z \in \mathbb{R}$ have prior $p(z) = \mathcal{N}(0, 1)$, and let two experts observe noisy views $y_t = z + \varepsilon_t$ and $y_c = z + \varepsilon_c$, where $(\varepsilon_t, \varepsilon_c)$ are jointly Gaussian with zero mean, marginal variances σ^2, and correlation $\rho \in [0, 1)$. Treating the experts as Gaussian* likelihood factors*, $q_t(z; y_t) \propto \exp\big(-(z - y_t)^2/(2\sigma^2)\big)$ and $q_c(z; y_c) \propto \exp\big(-(z - y_c)^2/(2\sigma^2)\big)$, the true Bayesian posterior $p(z \mid y_t, y_c) \propto p(z)\, p(y_t, y_c \mid z)$ has precision $\tau^* = 1 + 2/(\sigma^2(1 + \rho))$. The tempered PoE with $\alpha_t = \alpha_c = \alpha$ has posterior precision $\tau_\alpha = 1 + 2\alpha/\sigma^2$, and the unique α^* matching the true posterior precision is*

$$\alpha^* = \frac{1}{1 + \rho} \in \left(\tfrac{1}{2}, 1\right]. \tag{10}$$

Thus standard PoE ($\alpha = 1$) is precision-calibrated only when $\rho = 0$ and is overconfident whenever $\rho > 0$; for any $\rho > 0$, a tempered configuration $\alpha^* \in (1/2, 1)$ recovers the true posterior precision.

Remark 1. Proposition 2 provides an idealized posterior-precision motivation under 1D, equal-variance Gaussian likelihood factors with observation-noise correlation ρ. Because the deployed neural experts are multidimensional nonlinear parameterizations whose means, variances, and temperatures are optimized jointly, Eq. (10) motivates the tempered fusion family $\alpha \in [0, 1]$ (with 0.5 as the limiting $\rho \to 1$ reference) rather than serving as a direct estimator of ρ.

The proof is in Appendix C.4. Under diagonal Gaussian experts, the tempered posterior is available in closed form.

Proposition 3 (Closed-form tempered PoE posterior)**.** *Under the experts in Eq.* (8) *with diagonal covariance and $p(z) = \mathcal{N}(0, I)$, $q_\alpha(z \mid e_t, e_c) = \mathcal{N}(\mu_\alpha, \mathrm{diag}(\sigma_\alpha^2))$ with coordinate-wise precision, variance, and mean*

$$\tau_{\alpha,j} = 1 + \alpha_t \sigma_{t,j}^{-2} + \alpha_c \sigma_{c,j}^{-2}, \quad \sigma_{\alpha,j}^2 = \tau_{\alpha,j}^{-1}, \tag{11}$$

$$\mu_{\alpha,j} = \sigma_{\alpha,j}^2 \left(\alpha_t \sigma_{t,j}^{-2} \mu_{t,j} + \alpha_c \sigma_{c,j}^{-2} \mu_{c,j}\right). \tag{12}$$

The proof is in Appendix C.2. Equation (12) shows that each coordinate of the posterior mean is a precision-weighted convex combination of the prior mean 0 and the two expert means:

$$w_{t,j} = \frac{\alpha_t \sigma_{t,j}^{-2}}{\tau_{\alpha,j}}, \quad w_{c,j} = \frac{\alpha_c \sigma_{c,j}^{-2}}{\tau_{\alpha,j}}, \quad w_{0,j} = \frac{1}{\tau_{\alpha,j}}, \quad w_{t,j} + w_{c,j} + w_{0,j} = 1. \tag{13}$$

An axis dominates the posterior only when its expert is high-precision and its temperature is not suppressed; these weights define our posterior-reliance diagnostic in Sec. 4. During training we sample via reparameterization,

$$z = \mu_\alpha + \sigma_\alpha \odot \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \tag{14}$$

and at evaluation we pass the deterministic posterior mean $z = \mu_\alpha$ to the MLP classifier f_ψ, preserving the no-bypass graph $H \to E \to Z \to \hat{Y}$ while the mutual-information bound in Corollary 1 characterizes the stochastic training path (Appendix B).

4 Evaluation Protocol

We use *audit* in an intervention sense: each diagnostic perturbs or inspects a specific component of the forward evidence path and measures the response of the trained model. The five metrics target distinct mechanisms across evidence selection and posterior fusion and are not interchangeable.

(i) Gate faithfulness. For axis $a \in \{t, c\}$, temporal segments or channels are ranked once by descending gate value on the clean input and held fixed across 11 uniformly spaced mask fractions $\{0, 0.1, \ldots, 1.0\}$. Under zero-masking, insertion progressively restores top-ranked units to a fully zeroed input and deletion progressively zeroes them in the original input, rerunning the full model at every fraction to record the target-class score. The faithfulness gap is $\mathrm{gap}_a = \mathrm{AUC}_{\mathrm{ins}}^{(a)} - \mathrm{AUC}_{\mathrm{del}}^{(a)}$.

(ii) Expert necessity. Setting $\alpha_c = 0$ at evaluation collapses the channel expert into the prior ($q_\alpha \propto p(z)\, q_t^{\alpha_t}$); setting $\alpha_t = 0$ drops the temporal expert. We report the balanced accuracy drop $\Delta_a = M_{\mathrm{full}} - M_{\alpha_a=0}$, measuring the necessity of the aggregate expert branch.

(iii) Gate causality. For each axis and sweep value $v \in [0, 1]$, we replace $g_a(x)$ with the constant vector $v\mathbf{1}$ at evaluation time while keeping the trained experts intact, recording $M(v)$. A flat response indicates an insensitive gate regime; a sharp cliff identifies a causally required gate threshold.

(iv) Posterior reliance. Averaging the precision weights in Eq. (13) across dimensions gives $\rho_a = \frac{1}{d_z} \sum_j w_{a,j}$ on the simplex $\rho_t + \rho_c + \rho_0 = 1$, partitioning posterior precision into temporal, channel, and prior contributions: gates indicate *where* evidence is selected, whereas reliance quantifies *how much* posterior precision each axis contributes after selection.

(v) Expert disagreement. $D_{\mathrm{SKL}}(q_t, q_c) = \frac{1}{2}[D_{\mathrm{KL}}(q_t \| q_c) + D_{\mathrm{KL}}(q_c \| q_t)]$ measures divergence between the two Gaussian experts. To make its scale directly comparable across latent dimensions, we report both raw D_{SKL} and per-dimension D_{SKL}/d_z ($d_z = 64$), treating it as an inter-axis conflict diagnostic rather than predictive uncertainty. Detailed protocols are in Appendix F.

5 Experiments

We organize the empirical evaluation around four questions. Q1 (Sec. 5.2) tests whether the evidence-only constraint remains predictively feasible across tasks and frozen backbones (CodeBrain, CBraMod, and LaBraM). Q2 (Sec. 5.3) ablates the adaptation head variants, tempering rules, and gate-shaping regularizers. Q3 (Sec. 5.4) applies the five mechanistic audits to examine how evidence selection and posterior fusion interact. Q4 (Sec. 5.5) tests whether the selected channels recover literature-derived regional EEG biomarkers.

5.1 Experimental setup

Setup. We evaluate six EEG diagnostic tasks: event-type classification (TUEV), abnormality detection (TUAB), and seizure detection (TUSZ) [33], alongside cognitive-decline staging (AD/MCI/SCD/HC), depression screening, and cerebrovascular disease (CVD). The in-house clinical cohorts use 58–129-channel Electrical Geodesics, Inc. (EGI) HydroCel and BrainVision montages (Appendix E), providing a substantially denser spatial grid than the 16–22-channel TUH benchmarks. All data splits are strictly subject-disjoint prior to windowing. Our primary frozen backbone is CodeBrain [5], complemented by cross-backbone evaluations on frozen CBraMod [4] and LaBraM [3] encoders. We compare four heads that differ only in how temporal and channel information is integrated: *Linear probe*, *Axis-Concat*, *PoE-EIB* (α_t=α_c=1, untempered PoE bottleneck), and *tPoE-EIB* (learnable $\alpha_t, \alpha_c \in [0, 1]$). Default hyperparameters are d_z=64, β=10^{-4} (T_{warm}=5 epochs), η=10^{-3}, γ=10^{-2}, and 50 epochs of AdamW (lr 10^{-3}, patience 15) across 3 seeds (Appendices D–E).

5.2 Q1: Performance under the evidence-only constraint

Table 1 reports 3-seed balanced accuracy (BAcc) and matched temporal/channel faithfulness gaps $\mathrm{gap}_t, \mathrm{gap}_c$ across CNN, self-supervised, and pretrained EEG baselines. The comparison addresses two coupled questions: (i) whether restricting the classifier to the evidence-only interface costs diagnostic accuracy, and (ii) whether this behavior transfers across distinct frozen backbones.

Linear vs. tPoE-EIB across three frozen backbones. Replacing the linear probe with tPoE-EIB improves balanced accuracy on every public-benchmark cell for both CodeBrain and CBraMod, with gains of $+1.7/+4.3$ pts on TUAB, $+2.1/+2.7$ on TUEV, and $+2.2/+2.7$ on TUSZ (CodeBrain / CBraMod), alongside $+1.2$ to $+7.5$ pts across the three clinical cohorts. On an architecturally distinct frozen LaBraM encoder (Appendix H.3, Table 10), tPoE-EIB matches the Linear probe on TUAB ($79.7\pm0.7\%$ vs. $79.6\pm0.2\%$) and TUEV ($46.6\pm1.1\%$ vs. $46.9\pm0.3\%$) while improving

Table 1: Cross-model comparison across six EEG diagnosis tasks, split into two row-blocks (public benchmarks above, in-house clinical cohorts below). Each task reports 3-seed balanced accuracy (%) and matched temporal/channel faithfulness gaps $\mathrm{gap}_t, \mathrm{gap}_c$ (integrated gradients on the posterior, uniform across methods). Best BAcc per task is in **bold**.

	TUAB			TUEV			TUSZ		
Method	BAcc	gap_t	gap_c	BAcc	gap_t	gap_c	BAcc	gap_t	gap_c
EEGNet	79.4±0.6	0.129	−0.005	27.8±2.1	0.020	−0.012	50.2±0.0	0.001	0.001
ContraWR	78.3±1.5	0.159	0.140	47.9±2.7	0.151	0.066	61.7±3.8	0.014	0.068
BENDR	63.3±1.3	0.055	0.064	16.7±0.0	−0.026	0.003	50.0±0.0	0.000	0.000
BIOT	71.6±0.1	0.074	0.003	37.4±0.1	0.087	−0.055	60.9±6.2	0.015	0.002
LaBraM	73.2±0.1	0.176	−0.019	40.6±0.3	0.092	−0.028	55.4±0.6	0.017	−0.006
CodeBrain-Linear	79.0±0.3	0.058	0.085	46.7±0.6	0.150	−0.052	74.2±0.6	0.127	−0.025
CBraMod-Linear	68.2±0.6	0.042	−0.255	30.4±2.4	0.093	−0.087	57.6±1.2	0.027	−0.008
CodeBrain-tPoE-EIB	**80.7±0.2**	0.169	0.066	**48.8±0.7**	0.152	−0.003	**76.4±0.5**	0.098	−0.010
CBraMod-tPoE-EIB	72.5±0.2	0.020	−0.119	33.1±1.7	0.118	−0.097	60.3±1.7	0.031	−0.006

	AD staging			Depression			Cerebrovascular		
Method	BAcc	gap_t	gap_c	BAcc	gap_t	gap_c	BAcc	gap_t	gap_c
EEGNet	22.6±6.4	0.030	0.011	63.0±4.2	0.024	0.011	56.0±0.4	−0.004	−0.029
ContraWR	23.6±3.3	0.180	−0.058	**65.7±1.7**	0.215	0.046	**64.2±1.2**	−0.035	−0.049
BENDR	25.4±0.4	−0.002	0.001	56.3±1.0	0.110	0.013	50.1±0.1	−0.004	0.001
BIOT	**38.5±2.0**	0.036	0.008	59.2±0.8	0.173	0.254	61.4±0.2	−0.034	−0.007
LaBraM	36.0±1.3	0.074	−0.018	61.2±0.1	0.184	0.042	54.8±0.4	0.033	0.027
CodeBrain-Linear	31.0±1.6	0.222	−0.078	54.2±0.9	0.276	−0.091	56.8±0.8	0.184	−0.008
CBraMod-Linear	25.2±0.2	0.215	−0.032	50.5±0.4	0.085	0.059	54.8±0.8	0.294	−0.088
CodeBrain-tPoE-EIB	**38.5±4.1**	0.340	−0.072	58.8±2.0	0.221	0.365	58.0±0.8	0.290	−0.047
CBraMod-tPoE-EIB	26.1±5.4	0.023	0.001	54.3±0.4	−0.033	0.052	57.1±1.4	−0.028	0.014

Table 2: Head ablation: balanced accuracy (%, mean ± std over 3 seeds) on EEG diagnosis tasks with a frozen CodeBrain backbone. Best per task is in **bold**; second best is underlined.

Dataset	Linear probe	Axis-Concat	PoE-EIB ($\alpha = 1$)	tPoE-EIB (learnable α)
TUAB	79.0±0.3	80.4±0.5	80.3±0.4	**80.7±0.2**
TUEV	46.7±0.6	46.8±1.4	47.1±0.5	**48.8±0.7**
TUSZ	74.2±0.6	75.1±0.9	75.9±0.6	**76.4±0.5**
AD staging	31.0±1.6	31.3±3.1	34.0±1.6	**38.5±4.1**
AD binary	75.2±1.6	75.6±3.6	**77.5±2.6**	77.4±4.8
Depression	54.2±0.9	58.2±2.0	**60.4±1.1**	58.8±2.0
Cerebrovascular	56.8±0.8	57.4±0.8	57.6±1.0	**58.0±0.8**

TUSZ by +4.7 pts (78.6±0.3% vs. 73.9±0.5%), demonstrating that the evidence-only interface transfers cleanly across pretrained backbones.

Comparison to external methods. CodeBrain-tPoE-EIB achieves the highest BAcc on all three public benchmarks: TUAB (80.7%), TUEV (48.8%), and TUSZ (76.4%). On the in-house clinical cohorts, ContraWR leads on Depression (65.7%) and Cerebrovascular (64.2%), while BIOT ties with CodeBrain-tPoE-EIB on AD staging (38.5%). Across every dataset and backbone, tPoE-EIB consistently outperforms its matched linear probe, showing that enforcing an explicit, rate-limited decision path does not sacrifice frozen-backbone adaptability.

5.3 Q2: Ablations on heads, tempering, and regularizers

Head ablation. Table 2 completes the four-head sweep on a frozen CodeBrain backbone with the intermediate variants Axis-Concat and untempered PoE-EIB ($\alpha = 1$). tPoE-EIB ranks first on five of seven tasks (TUAB, TUEV, TUSZ, AD staging, Cerebrovascular) and second on AD binary and Depression, where untempered PoE ($\alpha = 1$) is highest. Among the three gated heads, the per-task spread is at most 2.0 pts on the public benchmarks and 7.2 pts on the clinical cohorts, indicating that dual-axis gate selection and rate-limited posterior integration contribute on the same scale. Calibration is task-dependent (Appendix H.2): tPoE-EIB attains the lowest NLL across all three public benchmarks (TUAB 0.427, TUEV 0.882, TUSZ 0.227) and the lowest ECE on TUEV (0.117) and TUSZ (0.018), whereas Axis-Concat yields lower ECE on TUAB (0.018 vs. 0.029).

Temperature ablation. Comparing fixed $\alpha = 0.5$, standard PoE ($\alpha = 1.0$), and learnable $\alpha \in [0, 1]$ (Appendix H.4, Table 11), learnable α achieves the highest BAcc on all three public benchmarks (80.7/48.8/76.4) and the lowest ECE on TUEV and TUSZ, while standard PoE ($\alpha = 1.0$) achieves

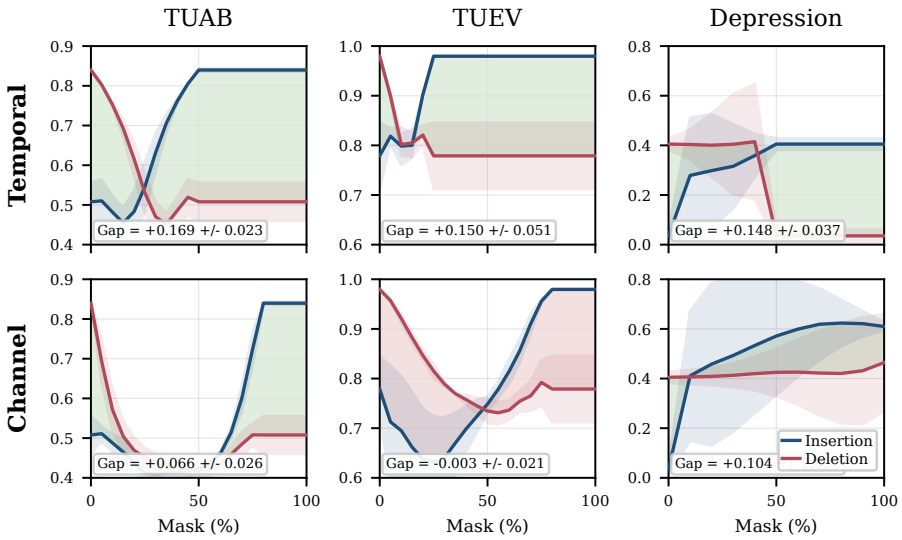


Figure 2: Insertion–deletion curves on TUAB, TUEV, and Depression computed from the learned tPoE-EIB gates. The shaded area between insertion (top) and deletion (bottom) curves quantifies gate faithfulness on each axis.

the lowest ECE on TUAB (0.028) and leads on Depression (60.4%). Tempering thus acts as an adaptive fusion mechanism that recovers both accuracy and calibration on imbalanced, event-centric tasks without forcing uniform down-weighting across every setting.

Loss and gate-shaping ablations. The effect of gate regularization scales with electrode density (Appendix H.5): on 16-channel TUAB, removing single regularizers leaves $\mathrm{gap}_c \in [+0.051, +0.061]$ and BAcc within 0.5 pts; on dense-channel Depression, removing $\mathcal{L}_{\text{sparse}}$, $\mathcal{L}_{\text{cons}}$, or $\mathcal{L}_{\text{KL}}$ flips gap_c from $+0.104$ to between -0.113 and -0.190, confirming that explicit channel compression is essential at high channel counts. Without the posterior bottleneck, the Axis-Concat temporal gate range on TUAB contracts from 0.316 to 0.137 (Appendix H.6).

5.4 Q3: Auditing selected evidence and posterior integration

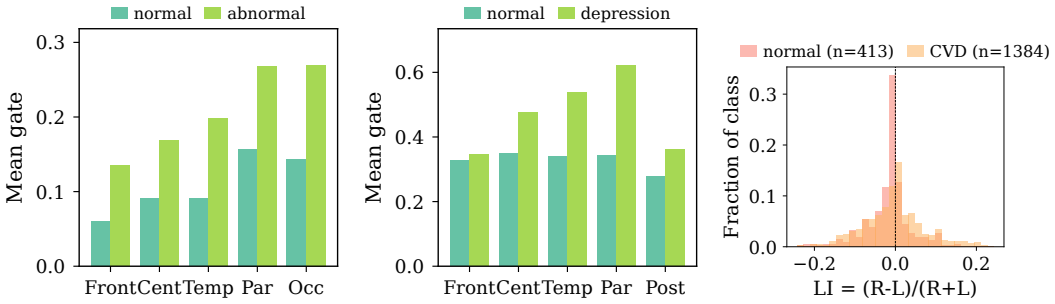
We next apply our five audits—gate faithfulness, expert necessity, posterior reliance, expert disagreement, and gate causality—to examine how selection and fusion interact across tasks (Table 3 and Appendix H.7, Table 14).

Decoupling expert necessity from marginal channel-ranking stability. Temporal gate faithfulness gap_t is positive across all six tasks ($+0.098$ to $+0.340$), and channel faithfulness gap_c is positive on TUAB ($+0.061$; 3-seed aggregate 0.066 ± 0.026, Fig. 2) and Depression ($+0.104$). By contrast, gap_c is near zero on TUEV (-0.003) and TUSZ (-0.010), and negative on CVD (-0.040) and four-way AD staging (-0.172). Jointly inspecting selection faithfulness (gap_c) and integration-stage expert necessity (Δ_c) separates two fundamentally different mechanisms. On TUEV and TUSZ, dropping the channel expert ($\alpha_c = 0$) reduces balanced accuracy by 22.4 and 26.9 points ($\Delta_c = 0.224$ and 0.269), proving that the aggregate channel summary e_c is critical for prediction even when individual channels do not admit a stable deletion ordering under zero-masking (reflecting diffuse spatial recruitment, unmodeled channel–time interactions, or zero-masking perturbation shifts). In four-way AD staging, however, $\mathrm{gap}_c = -0.172$ co-occurs with channel gates that saturate near 1.0 (Fig. 7d) and minimal expert-drop sensitivity ($\Delta_c = 0.026$), diagnosing genuine selector saturation at four-way granularity. Because the audit protocol probes both selection and fusion, it distinguishes an unused branch from an indispensable branch whose unit-level ranking is diffuse under masking.

Posterior reliance, expert disagreement, gate causality, and nuisance attenuation. Posterior reliance shows $\rho_0 \in [0.52, 0.62]$ on the public benchmarks, confirming active rate compression on the simplex $\rho_t + \rho_c + \rho_0 = 1$, and aligns positively with gap_c across reliance tertiles (Appendix H.1). Reporting expert disagreement both as raw D_{SKL} and normalized per latent dimension (D_{SKL}/d_z, $d_z = 64$) reveals wide task separation, from 6,953/dim on TUAB to 31.3/dim on Depression and CVD; on TUAB, correctly classified windows carry higher D_{SKL} than misclassified ones, showing that D_{SKL} measures temporal–channel conflict rather than predictive uncertainty. Gate-causality sweeps (Appendix H.8, Fig. 6) further resolve near-constant gates, separating a flat override response

Table 3: Audit summary on representative public tasks (best-validation checkpoint of tPoE-EIB, learned-gate ranking). Posterior reliance satisfies $\rho_t + \rho_c + \rho_0 = 1$; expert disagreement is reported both as raw D_{SKL} and normalized per latent dimension (D_{SKL}/d_z, $d_z = 64$). Full six-task table in Appendix H.7.

Dataset	gap_t	gap_c	Δ_t	Δ_c	ρ_t	ρ_c	ρ_0	D_{SKL}	Approx. D_{SKL}/d_z
TUAB	$+0.161$	$+0.061$	0.226	0.301	0.14	0.24	0.62	445k	6,953
TUEV	$+0.150$	-0.003	0.145	0.224	0.18	0.23	0.59	68k	1,063
TUSZ	$+0.098$	-0.010	0.130	0.269	0.18	0.30	0.52	42k	656



(a) TUAB: correct × class (b) Depression: correct × class (c) Cerebrovascular: signed LI

Figure 3: Regional alignment visualizations on three tasks (correctly classified samples, split by ground-truth class). (a) TUAB peaks at parieto-occipital channels with abnormal recordings doubling normal gate magnitude. (b) Depression elevates gate mass at parietal, central, and temporal sites in the depressive class. (c) Cerebrovascular disease exhibits a broader signed lateralization-index distribution than normal controls. Full per-task grids are in Appendix H.9.

on CVD (7.3 pt drop) from a sharp threshold cliff on Depression (44.1 pt drop between $v = 0.5$ and 0.6). Finally, linear probes for subject identity on TUAB (Appendix H.6, Table 13) show that subject-ID decoding attenuates from 30.9% at the channel gate to 21.1% at the latent mean and 1.1% at the logits (chance 0.4%).

5.5 Q4: Alignment of selected evidence with regional EEG priors

Whereas Q3 evaluates internal mechanistic behavior, Q4 tests whether the learned channel gates recover task-specific anatomical biomarkers reported in the clinical EEG literature. For each task we pre-register a directional regional hypothesis from textbook EEG signatures, average the per-sample channel gate within canonical 10–20 region groups (frontal, central, temporal, parietal, posterior/occipital), and run a one-sided Wilcoxon test of the predicted target–control contrast (or a Mann–Whitney test on the unsigned lateralization magnitude $|\mathrm{LI}| = |(L - R)/(L + R)|$ for focal slowing in CVD) across three seeds. Table 4 reports the test outcomes and Fig. 3 visualizes three representative tasks on correctly classified samples split by ground-truth class.

Prior recovered (TUAB, Depression, CVD, TUEV). On TUAB, the selector concentrates on parieto-occipital sites with the abnormal class doubling the normal-class gate magnitude, consistent with Posterior Dominant Rhythm (PDR) disruption [34, 35] (d=1.75±0.37, p<10^{-3}). Depression elevates central, parietal, and temporal weight in the depressive class with frontal exceeding posterior in the directional Wilcoxon contrast (d=2.06±1.36), extending beyond classical Frontal Alpha Asymmetry (FAA) [36–38]. CVD shows pronounced lateralization (|LI| effect d=4.92±2.67, p<10^{-3}) and a posterior shift specific to the disease class, consistent with focal lesion-side slowing [39, 40]. On TUEV, among correctly predicted events, ocular activity (EYEM) peaks at frontal channels (0.78 vs. 0.46–0.74 elsewhere), periodic discharges (GPED/PLED) concentrate at parietal/central sites, and artifacts (ARTF) receive the lowest weight across all regions (Fig. 8f).

Prior rejected, accuracy preserved (TUSZ, AD staging). TUSZ reaches 76.4±0.5% BAcc, but the focal-onset hypothesis (temporal/frontal > central/parietal) is rejected in every seed (d= − 0.41±0.19); the seizure-class profile (Fig. 8d) recruits channels broadly with a central/occipital peak across the heterogeneous TUSZ event mix. On four-way AD staging, tPoE-EIB ties for the highest BAcc (38.5±4.1%, chance 25%), yet the posterior–frontal contrast reverses direction (d= − 6.54±4.46) as the per-region gate saturates near 1.0 (Fig. 7d). Reformulating the same

Table 4: Alignment of learned channel gates against literature-derived EEG biomarkers. Cohen's d is the directional target–control effect size (or unsigned |LI| effect size for CVD), reported as 3-seed mean $\pm$ std on the held-out test split. *Aligned* summarizes the one-sided Wilcoxon decision across seeds.

Task	Anchor / predicted gate pattern	d (3-seed)	Aligned
TUAB	PDR: par./occ. $>$ front.	$+1.75\pm0.37$	✓ (3/3)
TUSZ	Focal onset: temp./front. $>$ cent./par.	-0.41 ± 0.19	$\times$ (3/3)
TUEV	Per-event (EYEM $\rightarrow$ front.; GPED/PLED $\rightarrow$ cent.)	per-class	EYEM front. (0.78)
AD staging	Post. $\alpha\downarrow/\theta\delta\uparrow$: post. $>$ front.	-6.54 ± 4.46	$\times$ (3/3)
Depression	FAA: front. $>$ post.	$+2.06\pm1.36$	✓ (3/3)
Cerebrovascular	Lesion-side slowing: $\lvert\text{LI}\rvert\uparrow$ in disease class	$+4.92\pm2.67$	✓ (3/3)

cohort as binary AD vs. NC recovers strong spatial discriminability—a logistic regression on the 58-dimensional channel gate reaches 88.8–97.9% BAcc and reveals a consistent fronto-parietal gate-correlation reversal (Appendix H.10)—showing that the four-way saturation stems from label granularity rather than absent AD signal, with a distributed fronto-parietal pattern rather than a simple "posterior $>$ frontal" contrast [41, 42]. Predictive accuracy and literature-prior alignment can therefore decouple, and an explicit channel interface makes this distinction directly testable.

6 Conclusion

We introduced tPoE-EIB, an evidence-information bottleneck head that adapts frozen EEG foundation backbones under a strict evidence-only prediction constraint ($H \rightarrow E \rightarrow Z \rightarrow \hat{Y}$), combining dual-axis gate selection with a closed-form tempered product of Gaussian experts over a shared low-rate latent. Across six public and clinical EEG tasks and three frozen backbones, tPoE-EIB maintains competitive diagnostic accuracy while enabling separate forward-path interventions on selection and fusion—allowing our audit protocol to distinguish aggregate expert necessity from unit-level ranking stability and to test when accurate predictions recover or diverge from textbook regional priors. Because the current interface is intentionally factorized into marginal temporal and channel gates, it does not directly parameterize localized channel-by-segment interactions (such as focal seizure propagation), and complementing internal mechanistic audits with demographic/medication confound controls and blinded clinician reader studies remains an important step toward clinical translation. Three technical directions naturally extend this framework while preserving the evidence-only restriction: rate-adaptive bottlenecks that allocate latent capacity by task granularity, hierarchical experts that decompose multi-stage staging into nested decisions, and structured joint time–channel selectors that capture localized spatiotemporal events.

Acknowledgments and Disclosure of Funding

Yinghao Wang is supported by the China Scholarship Council (CSC).

References

[1] 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.

[2] Yunfa Song, Qian Zheng, Bin 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.

[3] Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *International Conference on Learning Representations (ICLR)*, 2024.

[4] 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 *International Conference on Learning Representations (ICLR)*, 2025.

[5] Jingying Ma, Feng Wu, Qika Lin, Yucheng Xing, Chenyu Liu, Ziyu Jia, and Mengling Feng. CodeBrain: Bridging decoupled tokenizer and multi-scale architecture for EEG foundation model. In *International Conference on Learning Representations (ICLR)*, 2025.

[6] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning (ICML)*, 2017.

[7] Jonathan Crabbé and Mihaela van der Schaar. Explaining time series predictions with dynamic masks. In *International Conference on Machine Learning (ICML)*, 2021.

[8] Hyeongwon Jang, Changhun Kim, and Eunho Yang. TIMING: Temporality-aware integrated gradients for time series explanation. In *International Conference on Machine Learning (ICML)*, 2025.

[9] Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4198–4205, 2020.

[10] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[11] Mo Yu, Shiyu Chang, Yang Zhang, and Tommi Jaakkola. Rethinking cooperative rationalization: Introspective extraction and complement control. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[12] Chenyu Liu, Yuqiu Deng, Tianyu Liu, Jinan Zhou, Xinliang Zhou, Ziyu Jia, and Yi Ding. ECHO: Toward contextual seq2seq paradigms in large EEG models. *arXiv preprint arXiv:2509.22556*, 2025.

[13] Zhisheng Chen, Yingwei Zhang, Qizhen Lan, Tianyu Liu, Huacan Wang, Yi Ding, Ziyu Jia, Ronghao Chen, Kun Wang, and Xinliang Zhou. Uni-NTFM: A unified foundation model for EEG signal representation learning. *arXiv preprint arXiv:2509.24222*, 2025.

[14] Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

[15] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning (ICML)*, 2020.

[16] Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, Pietro Lio, and Mateja Jamnik. Concept embedding models: Beyond the accuracy-explainability trade-off. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[17] Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *International Conference on Learning Representations (ICLR)*, 2023.

[18] Tuomas Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *International Conference on Learning Representations (ICLR)*, 2023.

[19] Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

[20] Jasmijn Bastings, Wilker Aziz, and Ivan Titov. Interpretable neural predictions with differentiable binary variables. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

[21] Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations (ICLR)*, 2017.

[22] Thomas Strypsteen and Alexander Bertrand. End-to-end learnable EEG channel selection for deep neural networks with Gumbel-softmax. *Journal of Neural Engineering*, 18(4):0460a9, 2021.

[23] Feng Wu, Z. Song, G. Xu, K. Hu, Z. Wang, Y. Guo, and Z. Zhang. EEG-SeeGraph: Interpreting functional connectivity disruptions in dementias via sparse-explanatory dynamic EEG-graph learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6996–7000, 2026.

[24] M. C. Nnamdi, B. L. Marteau, Y. Zhong, J. B. Tamo, and M. D. Wang. Tree-of-evidence: Efficient "system 2" search for faithful multimodal grounding. *arXiv preprint arXiv:2604.07692*, 2026.

[25] Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 2000.

[26] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations (ICLR)*, 2017.

[27] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

[28] Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. In *International Conference on Learning Representations (ICLR)*, 2020.

[29] Yaochen Zhu, Jiayi Xie, and Zhenzhong Chen. Predicting the popularity of micro-videos with multimodal variational encoder-decoder framework. *arXiv preprint arXiv:2003.12724*, 2020.

[30] Jingqi Song, Yuanjie Zheng, Jing Wang, Muhammad Zakir Ullah, and Wanzhen Jiao. Multicolor image classification using the multimodal information bottleneck network (MMIB-Net) for detecting diabetic retinopathy. *Optics Express*, 29(14):22732–22748, 2021.

[31] Zongbo Han, Fan Yang, Junzhou Huang, Changqing Zhang, and Jianhua Yao. Multimodal dynamics: Dynamical fusion for trustworthy multimodal classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[32] Ziyu Jia, Tingyu Du, Zhengyu Tian, Hongkai Li, Yong Zhang, and Chenyu Liu. A multimodal BiMamba network with test-time adaptation for emotion recognition based on physiological signals. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[33] Iyad Obeid and Joseph Picone. The Temple University Hospital EEG data corpus. *Frontiers in Neuroscience*, 10:196, 2016.

[34] Shaun S. Lodder and Michel J. A. M. van Putten. Automated EEG analysis: Characterizing the posterior dominant rhythm. *Journal of Neuroscience Methods*, 200(1):86–93, 2011.

[35] Lara V. Marcuse, Marc Schneider, Katherine A. Mortati, Karen M. Donnelly, Vanessa Arnedo, and Arthur C. Grant. Quantitative analysis of the EEG posterior-dominant rhythm in healthy adolescents. *Clinical Neurophysiology*, 119(8):1778–1781, 2008.

[36] Jeffrey B. Henriques and Richard J. Davidson. Left frontal hypoactivation in depression. *Journal of Abnormal Psychology*, 100(4):535–545, 1991.

[37] John J. B. Allen and Samantha J. Reznik. Frontal EEG asymmetry as a promising marker of depression vulnerability: Summary and methodological considerations. *Current Opinion in Psychology*, 4:93–97, 2015.

[38] Nikita van der Vinne, Madelon A. Vollebregt, Michel J. A. M. van Putten, and Martijn Arns. Frontal alpha asymmetry as a diagnostic marker in depression: Fact or fiction? A meta-analysis. *NeuroImage: Clinical*, 16:79–87, 2017.

[39] Simon Finnigan and Michel J. A. M. van Putten. EEG in ischaemic stroke: Quantitative EEG can uniquely inform (sub-)acute prognoses and clinical management. *Clinical Neurophysiology*, 124(1):10–19, 2013.

[40] Simon Finnigan, Andrew Wong, and Stephen Read. Defining abnormal slow EEG activity in acute ischaemic stroke: Delta/alpha ratio as an optimal QEEG index. *Clinical Neurophysiology*, 127(2):1452–1459, 2016.

[41] Jaeseung Jeong. EEG dynamics in patients with Alzheimer's disease. *Clinical Neurophysiology*, 115(7):1490–1505, 2004.

[42] Roberta Lizio, Fabrizio Vecchio, Giovanni B. Frisoni, Raffaele Ferri, Guido Rodriguez, and Claudio Babiloni. Electroencephalographic rhythms in Alzheimer's disease. *International Journal of Alzheimer's Disease*, 2011:927573, 2011.

A Conceptual overview and symbol glossary

For reference, Fig. 4 contrasts a standard post-hoc downstream head with the evidence-only forward path used by tPoE-EIB, and Table 5 summarizes the core notation used throughout the paper.

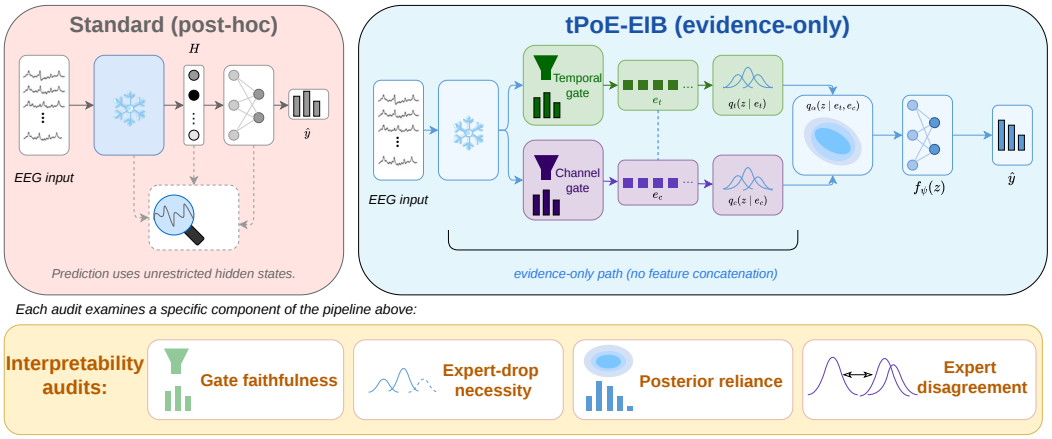


Figure 4: Conceptual contrast between a standard post-hoc downstream head and tPoE-EIB. A standard head (left) predicts from unrestricted hidden states H and explains the decision afterward. tPoE-EIB (right) routes prediction through temporal and channel gates, evidence summaries e_t, e_c, Gaussian experts $q_t(z \mid e_t)$ and $q_c(z \mid e_c)$, and a tempered product-of-experts posterior $q_\alpha(z \mid e_t, e_c)$. The classifier f_ψ observes only the latent, so no concatenated evidence or unrestricted hidden state reaches the classifier.

Table 5: Compact glossary of core symbols and audit metrics.

Symbol	Domain / Scale	Description
x, $H = B_\theta(x)$	$\mathbb{R}^{C\times S\times P}$, $\mathbb{R}^{C\times S\times D}$	Input EEG window and frozen backbone patch-grid representation
$g_t(x)$, $g_c(x)$	$[0,1]^S$, $[0,1]^C$	Learned marginal temporal and channel gates
$E = (e_t, e_c)$	$\mathbb{R}^D \times \mathbb{R}^D$	Gated temporal and channel evidence summaries (Eq. 5)
$q_t(z \mid e_t)$, $q_c(z \mid e_c)$	$\mathcal{N}(\mu_a, \mathrm{diag}(\sigma_a^2))$	Axis-specific Gaussian evidence experts over latent $z \in \mathbb{R}^{d_z}$ (Eq. 8)
α_t, α_c	$[0,1]$	Tempered PoE fusion exponents (default: learnable; $\alpha = 1$: standard PoE)
$q_\alpha(z \mid e_t, e_c)$	$\mathcal{N}(\mu_\alpha, \mathrm{diag}(\sigma_\alpha^2))$	Closed-form tempered PoE posterior (Proposition 3)
$\beta_{\text{eff}}(t)$	$[0, \beta]$	Warmed-up KL rate coefficient $\beta \min(1, t/T_{\text{warm}})$ in Eq. (4)
gap_t, gap_c	$[-1,1]$	Insertion–deletion AUC faithfulness gap under zero-masking
Δ_t, Δ_c	BAcc drop	Expert-drop necessity when setting $\alpha_t = 0$ or $\alpha_c = 0$ at evaluation
ρ_t, ρ_c, ρ_0	$[0,1]$, $\sum \rho = 1$	Normalized posterior precision reliance on temporal, channel, and prior
D_{SKL}, D_{SKL}/d_z	$[0, \infty)$	Raw and per-dimension symmetric KL disagreement between q_t and q_c
LI, $\lvert\text{LI}\rvert$	$[-1,1]$, $[0,1]$	Signed lateralization index and unsigned magnitude $\lvert(L-R)/(L+R)\rvert$

B Theoretical properties (extended)

This appendix expands the theoretical statements referenced in the main text. We record two further properties: a rate-limited path bound that follows from the variational bound in Proposition 1, and a constant-gate failure mode (Lemma 1) that motivates the audits in Sec. 4. We also state a variance-conservatism corollary that complements the calibration argument for tempering given by Proposition 2 in the main text.

B.1 Rate-limited evidence path

Corollary 1 (Rate-limited evidence path)**.** *Along the forward graph $E \to Z \to \hat{Y}$, the data-processing inequality gives*

$$I(\hat{Y}; E) \le I(Z; E) \le \mathbb{E}_{p(E)} D_{\mathrm{KL}}\big(q_\alpha(z \mid E) \,\|\, p(z)\big).$$

Because $X \to H \to E$ is deterministic under the frozen backbone and deterministic gates, the same stochastic path bounds the dependence of the prediction on the input: $I(\hat{Y}; X) \le \mathbb{E}_{p(E)} D_{\mathrm{KL}}(q_\alpha(z \mid E) \| p(z))$.

Remark 2. This information-theoretic statement applies to the stochastic training path $Z \sim q_\alpha(z \mid E)$ in Eq. (14). At evaluation, we use the posterior mean μ_α as a variance-reduced predictor. This preserves the architectural restriction that the classifier sees only the tPoE latent parameterization, while the mutual-information bound itself is specific to the stochastic path (on an auxiliary evaluation subset, predictions from $Z = \mu_\alpha$ and a single stochastic draw $Z \sim q_\alpha(z \mid E)$ differed by at most 0.6 BAcc points).

B.2 Stability is not selectivity

Sparsity and consistency penalties on the gates do not by themselves ensure sample-dependent evidence selection: a constant gate can minimize both.

Lemma 1 (Stability is not selectivity)**.** *Let $g_t(x) = \bar{g}_t$ and $g_c(x) = \bar{g}_c$ be constant for all x. Then the consistency loss in Eq.* (7) *is zero for any augmentation operator* Aug*. The sparsity loss in Eq.* (6) *equals $\frac{1}{2}\big(\|\bar{g}_t\|_1/S + \|\bar{g}_c\|_1/C\big)$, which depends only on the chosen constants and is small whenever those constants are sparse.*

The proof is in Appendix C.1. Lemma 1 formalizes the constant-gate failure: such a gate is a degenerate rationale in the sense of Yu et al. [11]. The bottleneck does not make this shortcut impossible, but Corollary 1 ensures that any information passing through a collapsed selector is still rate-limited, and Sec. 4 turns that constraint into measurable evidence-quality tests.

B.3 Variance conservatism of tempered PoE

Proposition 2 (in the main text) gives the calibration argument for tempering. A complementary observation is that tempering is also conservative with respect to posterior variance.

Corollary 2 (Tempering is conservative relative to standard PoE)**.** *If $0 \le \alpha_t, \alpha_c \le 1$, then the tPoE posterior variance is element-wise no smaller than the standard PoE variance obtained with $\alpha_t = \alpha_c = 1$. This follows directly from Eqs.* (11)*.*

C Additional Proofs

C.1 Proof of Lemma 1 (stability is not selectivity)

For every x, $g_t(\mathrm{Aug}(x)) = \bar{g}_t = g_t(x)$ and similarly for g_c, so the squared differences in Eq. (7) vanish. The sparsity claim is immediate from Eq. (6) since both averages equal $\|\bar{g}_t\|_1/S$ and $\|\bar{g}_c\|_1/C$. □

C.2 Proof of Proposition 3 (closed-form tempered PoE posterior)

Let $\Sigma_t = \mathrm{diag}(\sigma_t^2)$ and $\Sigma_c = \mathrm{diag}(\sigma_c^2)$ denote the diagonal covariance matrices of the two experts in Eq. (8). Expanding $\log q_\alpha(z)$ from Eq. (9) gives the sum of three quadratic forms in z: the standard normal prior term $-\frac{1}{2}\|z\|^2$ and the two tempered expert log-densities $-\frac{\alpha_t}{2}(z-\mu_t)^\top \Sigma_t^{-1}(z-\mu_t)$ and $-\frac{\alpha_c}{2}(z-\mu_c)^\top \Sigma_c^{-1}(z-\mu_c)$. Collecting quadratic and linear terms in z yields a Gaussian with precision matrix $I + \alpha_t\Sigma_t^{-1} + \alpha_c\Sigma_c^{-1}$ and natural parameter $\alpha_t\Sigma_t^{-1}\mu_t + \alpha_c\Sigma_c^{-1}\mu_c$. Diagonal covariance gives Eqs. (11)–(12). □

C.3 Proof of Proposition 1 (variational lower bound)

We bound the two terms of $\mathcal{J}_{\mathrm{EIB}} = I(Y;Z) - \beta I(E;Z)$ separately.

Label term (lower bound on $I(Y;Z)$). By definition, $I(Y;Z) = H(Y) - H(Y \mid Z) = H(Y) + \mathbb{E}_{p(y,z)}[\log p(y \mid z)]$. For any decoder r_ψ, $\mathbb{E}_{p(y,z)}[\log p(y \mid z)] = \mathbb{E}_{p(y,z)}[\log r_\psi(y \mid z)] + \mathbb{E}_{p(z)} D_{\mathrm{KL}}(p(y \mid z) \| r_\psi(y \mid z))$, and the KL term is non-negative. Therefore

$$I(Y;Z) \geq H(Y) + \mathbb{E}_{p(x,y)} \mathbb{E}_{q_\alpha(z|E)}[\log r_\psi(y \mid z)]. \tag{15}$$

Rate term (upper bound on $I(E;Z)$). Let $\bar{q}(z) = \mathbb{E}_{p(E)}[q_\alpha(z \mid E)]$ be the aggregated posterior. Using the identity $\mathbb{E}_{p(E)} D_{\mathrm{KL}}(q_\alpha(z \mid E) \| p(z)) = I(E;Z) + D_{\mathrm{KL}}(\bar{q}(z) \| p(z))$ [26] and the non-negativity of $D_{\mathrm{KL}}(\bar{q} \| p)$,

$$I(E;Z) \leq \mathbb{E}_{p(E)} D_{\mathrm{KL}}(q_\alpha(z \mid E) \| p(z)). \tag{16}$$

Substituting the lower bound on $I(Y;Z)$ from Eq. (15) and the upper bound on $I(E;Z)$ from Eq. (16) into $\mathcal{J}_{\mathrm{EIB}} = I(Y;Z) - \beta I(E;Z)$ yields Eq. (3). □

C.4 Proof of Proposition 2 (tempering corrects overconfidence under correlated experts)

The joint distribution $p(z, \varepsilon_t, \varepsilon_c)$ is Gaussian. Conditioning on (y_t, y_c) and completing the square in z gives the true posterior precision $\tau^* = 1 + \mathbf{1}^\top \Sigma_\varepsilon^{-1} \mathbf{1}$, where $\Sigma_\varepsilon = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ and $\mathbf{1} = (1,1)^\top$. Direct inversion yields $\tau^* = 1 + 2/(\sigma^2(1+\rho))$. For tempered PoE, the diagonal formula of Proposition 3, specialized to $d_z = 1$, $\sigma_t^2 = \sigma_c^2 = \sigma^2$, and $\alpha_t = \alpha_c = \alpha$, gives $\tau_\alpha = 1 + 2\alpha/\sigma^2$. Equating $\tau_\alpha = \tau^*$ yields $\alpha^* = 1/(1+\rho)$. Monotonicity of τ_α in α gives the calibration claims. □

C.5 Nuisance leakage consumes bottleneck rate

Suppose evidence decomposes as $E = (S, N)$, where S is label-relevant signal, N is nuisance variation, and $Y \perp N \mid S$. Then

$$I(E;Z) = I(S,N;Z) = I(S;Z) + I(N;Z \mid S). \tag{17}$$

The term $I(N;Z \mid S)$ is non-negative and consumes bottleneck rate. Under the evidence-IB objective $I(Y;Z) - \beta I(E;Z)$, nuisance information that reaches Z is penalized but cannot increase the label information once S is known. This does not prove that optimization will always discard nuisance, but it motivates selecting irrelevant components before compression.

C.6 Diagonal Gaussian KL used for expert disagreement

For two diagonal Gaussians $q_A = \mathcal{N}(\mu_A, \mathrm{diag}(\sigma_A^2))$ and $q_B = \mathcal{N}(\mu_B, \mathrm{diag}(\sigma_B^2))$,

$$D_{\mathrm{KL}}(q_A \| q_B) = \frac{1}{2} \sum_j \left[\log \frac{\sigma_{B,j}^2}{\sigma_{A,j}^2} + \frac{\sigma_{A,j}^2 + (\mu_{A,j} - \mu_{B,j})^2}{\sigma_{B,j}^2} - 1 \right]. \tag{18}$$

The symmetric expert-conflict diagnostic is

$$D_{\mathrm{SKL}}(q_t, q_c) = \frac{1}{2} \left[D_{\mathrm{KL}}(q_t \| q_c) + D_{\mathrm{KL}}(q_c \| q_t) \right]. \tag{19}$$

D Implementation Details

The head exposes a small number of interpretable knobs. The bottleneck coefficient β controls the evidence information budget and is warmed up via $\beta_{\mathrm{eff}}(t) = \beta \min(1, t/T_{\mathrm{warm}})$ over T_{warm} epochs. The sparsity and consistency coefficients η and γ control gate mass and stability. The temperatures $\alpha_t, \alpha_c \in [0,1]$ discount correlated evidence axes, and the latent dimension d_z sets the dimensionality of the decision path. Temporal and channel gates are two-layer MLPs with hidden width 64 and GELU activations. Each Gaussian expert uses a linear–GELU–linear projection to produce μ and

$\log \sigma^2$, with log-variance clamped to $[-10, 10]$ for numerical stability. Unless otherwise stated, $d_z = 64$.

We evaluate two tempered configurations: a fixed variant with $\alpha_t = \alpha_c = 0.5$ and a learnable variant in which α_t, α_c are constrained to $[0, 1]$ by sigmoid reparameterization initialized at 0.5. Standard non-tempered PoE is recovered by setting $\alpha_t = \alpha_c = 1.0$ and is used as an ablation.

At evaluation time, all stochastic bottlenecks are made deterministic by setting $z = \mu_\alpha$. This makes insertion–deletion curves, posterior reliance, and expert-drop diagnostics reproducible. Expert-drop changes only posterior aggregation weights; it does not mask the EEG input.

Table 6: Default head and optimization hyperparameters used unless otherwise stated.

Parameter	Value
Latent dimension d_z	64
Bottleneck coefficient β	10^{-4}
Warmup T_{warm}	5 epochs
Sparsity coefficient η	10^{-3}
Consistency coefficient γ	10^{-2}
Optimizer	AdamW
Learning rate	10^{-3}
Weight decay	10^{-3}
Maximum epochs	50
Early-stopping patience	15 epochs

E Dataset Details

Table 7 reports the raw dataset statistics and split conventions used in Sec. 5. In addition to TUH-derived public benchmarks, we use private resting-state clinical EEG cohorts to test disease-oriented diagnosis under the same frozen-backbone adaptation setting. These in-house labels reflect routine clinical assessment informed by EEG review and chart-level context rather than waveform annotation alone. The depression and cerebrovascular cohorts are binary condition-vs-control datasets built from de-identified 5 s windows; AD staging is a four-way AD/MCI/SCD/HC cohort. Together, the in-house cohorts provide a denser channel axis (58–129 EEG channels post-preprocessing) than the public benchmarks (16–22) and therefore a more demanding setting for channel-faithfulness analysis.

Preprocessing. We follow the CBraMod [4] preprocessing protocol, which is also the convention used by CodeBrain [5]: all recordings are uniformly resampled to 200 Hz before windowing so that the temporal axis is aligned across heterogeneous public and in-house cohorts (the "f_s" column of Table 7 reports the original acquisition rates). Channel selection is dataset-specific. (i) The cerebrovascular cohort is acquired with a 140-channel EGI net; we retain the 129 EEG-bearing electrodes and drop the 11 non-EEG sensors (reference / face / EOG / ground leads) before windowing. (ii) The depression cohort mixes EGI (129-channel HydroCel) and BrainVision (64-channel) acquisitions; we first align the two systems on their common 10–20-positioned electrodes, then retain the 63 EEG channels shared by both montages and exclude the BrainVision ECG lead. (iii) The AD-staging cohort is acquired with EGI on a 58-electrode grid and used as-is. The TUH public benchmarks (TUAB, TUEV, TUSZ) follow the standard CBraMod bipolar montage and label-window construction unchanged. The "Ch." column in Table 7 therefore lists the maximum acquisition channel count of each cohort; the channel grid actually fed to the model after preprocessing is 129 for cerebrovascular, 63 for depression, and matches the table value for the remaining tasks.

Data sources and ethics. TUAB, TUEV, and TUSZ are public TUH EEG benchmarks. The private clinical EEG cohorts used in this study are covered by institutional ethics approvals (approval Nos. 81501155, ZYLX201607, and KD2022165). All in-house data are private and outside any pretraining corpus. Only de-identified windowed EEG segments and derived record metadata enter the modeling pipeline. Splits are constructed before windowing so that no subject contributes segments to more than one split. De-identified in-house data may be made available for research use upon reasonable request, subject to coauthor approval and institutional data-governance constraints. Demographic variables (age, sex) and medication status are not recorded in these private cohorts, so retrospective confounding probes for those covariates are not available.

Table 7: Raw dataset statistics. Splits are reported in number of windows; the "Ch." column gives the maximum acquisition channel count and "f_s" the original sampling rate (Hz). All recordings are uniformly resampled to 200 Hz before windowing; channel selection (cerebrovascular $140 \rightarrow 129$, depression $64 \rightarrow 63$ after EGI/BrainVision alignment) is described above.

Dataset	Task	$\lvert\mathcal{Y}\rvert$	Subj.	Train/Val/Test	Ch.	f_s
TUAB	Abnormality	2	2,329	297k/75k/37k	16	250/256
TUEV	Event type	6	294	176k/—/29k	16	250
TUSZ	Seizure	2	675	542k/—/92k	22	200
AD staging	Cognitive decline	4	108	11.7k/2.5k/2.5k	58	250
Depression	MDD vs. control	2	159	4.6k/1.0k/0.6k	64	500/1000
Cerebrovascular	Stroke / lesion	2	83	7.4k/1.5k/1.8k	140	500/5000

F Evaluation Protocol Details

This appendix expands the audit protocols summarized in Section 4.

Insertion–deletion gap (extended). For each axis $a \in \{t, c\}$, we rank temporal segments or channels by descending gate value. Insertion starts from a fully zero-masked input and progressively restores high-ranked segments or channels; deletion starts from the original input and progressively replaces the same ranked units with zeros. The full model is re-run after each perturbation, while the ranking is fixed from the clean input. Mask fractions are uniformly spaced on $\{0, 0.1, \ldots, 1.0\}$. For each fraction, we record the predicted target-class probability, average over evaluation samples, and integrate to obtain $\mathrm{AUC}_{\mathrm{ins}}^{(a)}$ and $\mathrm{AUC}_{\mathrm{del}}^{(a)}$. The gap is $\mathrm{gap}_a = \mathrm{AUC}_{\mathrm{ins}}^{(a)} - \mathrm{AUC}_{\mathrm{del}}^{(a)}$. To connect selection with posterior integration, we additionally stratify samples by posterior reliance: if high-ρ_t samples show larger temporal gaps than low-ρ_t samples, temporal reliance aligns with temporal gate faithfulness.

Expert necessity (extended). Setting both temperatures to zero gives the prior-only noise floor $q_\alpha = p(z)$. We use balanced accuracy as the predictive metric M throughout; Δ values reported in tables are averaged over three seeds.

Gate causality (extended). The override sweep is evaluated on the same held-out test set as the predictive metric. We use 8 sweep values $v \in \{0.0, 0.3, 0.5, 0.6, 0.7, 0.8, 0.9, 0.999\}$, with finer resolution near the typical learned-gate regime. The resulting $M(v)$ curve is reported for CVD and Depression in Fig. 6.

Symmetric KL between experts (diagonal-Gaussian form). For two diagonal-covariance Gaussians $q_t = \mathcal{N}(\mu_t, \mathrm{diag}(\sigma_t^2))$ and $q_c = \mathcal{N}(\mu_c, \mathrm{diag}(\sigma_c^2))$,

$$D_{\mathrm{KL}}(q_t \| q_c) = \frac{1}{2} \sum_j \left[\log \frac{\sigma_{c,j}^2}{\sigma_{t,j}^2} + \frac{\sigma_{t,j}^2 + (\mu_{t,j} - \mu_{c,j})^2}{\sigma_{c,j}^2} - 1 \right],$$

and $D_{\mathrm{SKL}} = \frac{1}{2}[D_{\mathrm{KL}}(q_t \| q_c) + D_{\mathrm{KL}}(q_c \| q_t)]$. We report sample-level D_{SKL}, its per-dimension value D_{SKL}/d_z ($d_z = 64$), and its conditional means under correct/incorrect prediction subsets.

G Reporting Details

The evaluation records both summary and per-sample quantities. Summary reports include predictive metrics, calibration, terminal KL, evidence gaps with reliance stratification, expert-drop deltas, posterior reliance means, and expert disagreement. Per-sample reports include labels, predictions, confidence, gap_t, gap_c, ρ_t, ρ_c, ρ_0, $D_{\mathrm{SKL}}(q_t, q_c)$, and posterior variance. These quantities separate accuracy, gate faithfulness, posterior reliance, and temporal–channel conflict, enabling audits of both evidence selection and posterior integration.

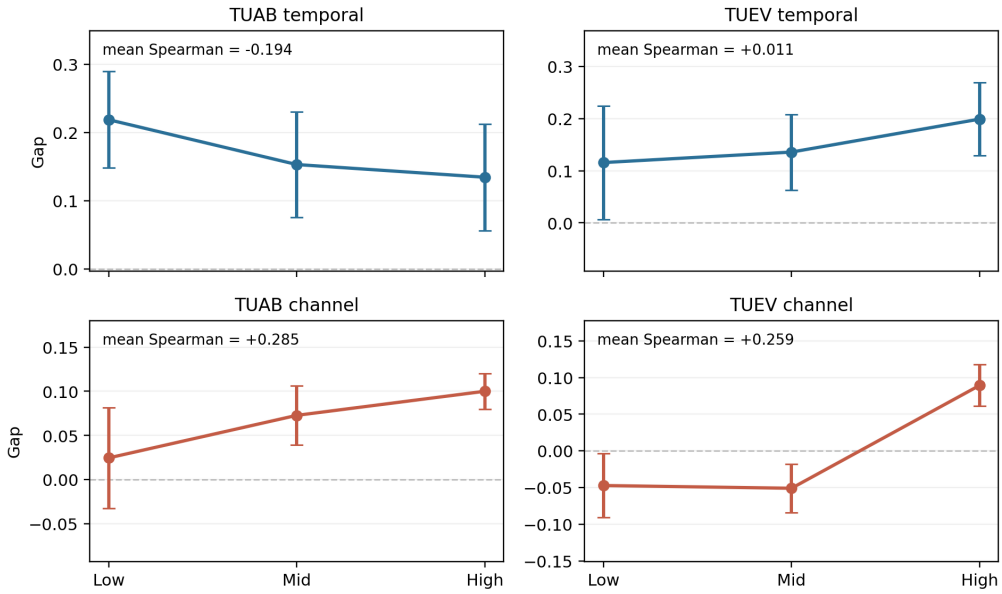


Figure 5: Gate-faithfulness gap stratified by posterior-reliance tertile for the learnable-α tPoE-EIB variant on TUAB and TUEV. Each panel reports the mean gap in low/mid/high reliance groups, averaged over three seeds. Higher ρ_c aligns with larger gap_c on both datasets, whereas temporal alignment is task-dependent.

H Additional Experimental Results

H.1 Reliance-stratified gap

The clearest alignment is on the channel axis: in both TUAB and TUEV, higher ρ_c corresponds to larger gap_c, so samples that receive more posterior precision from the channel expert tend to show more faithful channel perturbation rankings. Temporal alignment is task-dependent: TUEV shows an average increase in gap_t from low to high ρ_t, whereas TUAB does not. On TUAB, the temporal-reliance–gap correlation is weakly negative ($r \approx -0.19$), indicating that on this dataset temporal reliance reflects model confidence rather than perturbation faithfulness.

H.2 Full calibration table

Table 8: Full calibration diagnostics across all four heads, public benchmarks. Terminal KL is the final training KL for bottlenecked heads.

Dataset	Method	ECE ↓	NLL ↓	Terminal KL
TUAB	Linear probe	0.024	0.449	0.0
	Axis-Concat	**0.018**	**0.425**	0.0
	PoE-EIB ($\alpha = 1$)	0.030	0.429	48.1
	tPoE-EIB (learnable α)	0.029	0.427	43.9
TUEV	Linear probe	0.126	0.895	0.0
	Axis-Concat	0.139	1.220	0.0
	PoE-EIB ($\alpha = 1$)	0.138	1.114	30.0
	tPoE-EIB (learnable α)	**0.117**	**0.882**	28.2
TUSZ	Linear probe	0.051	0.299	0.0
	Axis-Concat	0.026	0.272	0.0
	PoE-EIB ($\alpha = 1$)	0.021	0.255	39.2
	tPoE-EIB (learnable α)	**0.018**	**0.227**	39.2

Table 9: Balanced accuracy (%) on TUAB with a frozen CBraMod backbone (full four-head sweep). Mean ± std over 3 seeds. Companion to Sec. 5.2; CBraMod TUEV and TUSZ results (Linear and tPoE-EIB) are reported in Table 1 and Table 10.

Dataset	Linear probe	Axis-Concat	PoE-EIB ($\alpha = 1$)	tPoE-EIB (learnable α)
TUAB	68.2±0.6	72.0±0.2	71.8±0.6	**72.5±0.2**

Table 10: Cross-backbone transfer on a frozen LaBraM encoder and CBraMod TUSZ endpoint (balanced accuracy, %, mean ± std over 3 seeds). These rows freeze the pretrained encoder and train the matched Linear or tPoE-EIB downstream head.

Frozen backbone	Task	Linear probe	tPoE-EIB	Δ BAcc
LaBraM	TUAB	79.6±0.2	79.7±0.7	+0.1
LaBraM	TUEV	46.9±0.3	46.6±1.1	−0.3
LaBraM	TUSZ	73.9±0.5	78.6±0.3	+4.7
CBraMod	TUSZ	57.6±1.2	60.3±1.7	+2.7

H.3 Cross-backbone generalization (CBraMod and LaBraM)

On TUAB, CBraMod's lower absolute performance compresses the gap among heads, but the relative ordering is preserved: tPoE-EIB (72.5%) > Axis-Concat (72.0%) > PoE-EIB (71.8%) ≫ Linear (68.2%). On the frozen LaBraM encoder (Table 10), tPoE-EIB matches the Linear head on TUAB (+0.1) and TUEV (−0.3) and improves TUSZ by +4.7 pts.

H.4 Temperature ablation

Table 11 compares fixed $\alpha = 0.5$, fixed $\alpha = 1.0$, and a learnable $\alpha \in [0, 1]$ via sigmoid reparameterization. The learnable variant attains the highest balanced accuracy on all three tasks and the lowest ECE on TUSZ and TUEV; standard PoE ($\alpha = 1.0$) gives the lowest ECE on TUAB. Tempering therefore trades calibration for accuracy on event-centric and balanced tasks, and recovers both on the most imbalanced setting (TUSZ and TUEV).

Table 11: Temperature ablation. Each cell reports balanced accuracy (%) / ECE.

Variant	TUAB	TUEV	TUSZ
$\alpha = 0.5$ (fixed)	80.2 / 0.031	46.8 / 0.135	75.7 / 0.021
$\alpha = 1.0$ (standard PoE)	80.3 / **0.028**	47.1 / 0.128	75.9 / 0.021
Learnable α	**80.7** / 0.031	**48.8 / 0.117**	**76.4 / 0.018**

H.5 Loss ablation

We remove each regularizer in $\mathcal{L}_{\text{total}}$ in turn (Table 12). The pattern is channel-count-dependent. On TUAB, single-regularizer ablations change BAcc by at most 0.5 pts and leave gap_c positive but small ($+0.051$ to $+0.061$), so the regularizers reshape the evidence interface only modestly when the channel count is small. On the dense-channel Depression cohort, the regularizers are essential for channel compression: removing any single regularizer flips gap_c sharply negative ($+0.104$ with the full loss vs. -0.113, -0.158, -0.190 when removing $\mathcal{L}_{\text{sparse}}$, $\mathcal{L}_{\text{cons}}$, $\mathcal{L}_{\text{KL}}$ respectively), with BAcc largely preserved. On TUEV (6-class), single-regularizer ablations cost 2–3 pts of BAcc, and removing $\mathcal{L}_{\text{KL}}$ flips gap_c from -0.003 to $+0.037$. Terminal KL on TUEV is reported on a per-sample basis throughout this ablation, matching the per-example coding-rate interpretation that motivates the bottleneck; since every row of the table uses this same convention, the sign and ranking of within-table differences are invariant to the normalization choice.

H.6 Gate shaping diagnostics and nuisance probes

Table 13 summarizes TUAB diagnostics from a matched bottleneck versus no-bottleneck analysis. These are mechanism checks rather than headline benchmarks: they test whether the bottleneck makes gates more selective, less collapsed, and less subject-identifiable. The pattern is axis-dependent.

Table 12: Loss ablation on TUAB, TUEV, and Depression.

Dataset	Variant	Bal. Acc. ↑	ECE ↓	Terminal KL	gap_t	gap_c
TUAB	Full	80.7	0.029	43.9	$+0.161$	$+0.061$
	w/o $\mathcal{L}_{\mathrm{sparse}}$	80.2	0.028	45.6	$+0.153$	$+0.059$
	w/o $\mathcal{L}_{\mathrm{cons}}$	80.6	0.022	43.9	$+0.148$	$+0.051$
	w/o $\mathcal{L}_{\mathrm{KL}}$ (axcat)	80.6	0.030	–	$+0.154$	$+0.056$
TUEV	Full	48.8	0.117	106.2	$+0.150$	-0.003
	w/o $\mathcal{L}_{\mathrm{sparse}}$	45.4	0.232	109.8	0.131	-0.012
	w/o $\mathcal{L}_{\mathrm{cons}}$	46.5	0.217	105.8	0.057	-0.114
	w/o $\mathcal{L}_{\mathrm{KL}}$ (axcat)	46.8	0.139	–	0.051	0.037
Depression	Full	58.8	0.152	66.4	0.148	0.104
	w/o $\mathcal{L}_{\mathrm{sparse}}$	59.2	0.122	68.5	0.151	-0.113
	w/o $\mathcal{L}_{\mathrm{cons}}$	60.1	0.152	66.3	0.244	-0.158
	w/o $\mathcal{L}_{\mathrm{KL}}$ (axcat)	59.4	0.182	–	0.240	-0.190

Temporal support becomes less flat, channel support becomes less concentrated, and linear subject-identity decodability is attenuated from the gates (30.9%) toward the latent (21.1%) and logits (1.1%, vs. 0.4% chance).

Table 13: Supporting TUAB gate-shaping and nuisance diagnostics from a matched bottleneck versus no-bottleneck analysis. Chance subject-ID accuracy is 0.4%.

Diagnostic	No bottleneck	Bottlenecked VIB	Interpretation
Temporal gate range	0.137	0.316	less flat, more selective temporal support
Coverage above 0.5	0.002	0.207	high-confidence temporal support emerges
Channel Gini	0.472	0.358	less concentrated channel weighting
Effective channels	10.8	12.7	broader channel support, less collapse
Subject-ID from channel gate	43.1%	30.9%	nuisance attenuated, not removed
Subject-ID from latent μ	—	21.1%	residual nuisance survives compression
Subject-ID from logits	0.9%	1.1%	final decision stays near subject chance

H.7 Audit summary across all six tasks

Table 14 extends Table 3 with all audit values across the six tasks.

Table 14: Audit summary across all six tasks (best-validation checkpoint, learned-gate ranking).

Dataset	gap_t	gap_c	Δ_t	Δ_c	ρ_t	ρ_c	ρ_0	D_{SKL}	Approx. D_{SKL}/d_z
TUAB	$+0.161$	$+0.061$	0.226	0.301	0.14	0.24	0.62	445k	6,953
TUEV	$+0.150$	-0.003	0.145	0.224	0.18	0.23	0.59	68k	1,063
TUSZ	$+0.098$	-0.010	0.130	0.269	0.18	0.30	0.52	42k	656
AD staging	$+0.340$	-0.172	0.028	0.026	0.43	0.55	0.03	41k	641
Depression	$+0.148$	$+0.104$	0.169	0.153	0.37	0.39	0.24	2k	31.3
Cerebrovascular	$+0.196$	-0.040	0.037	0.018	0.34	0.38	0.28	2k	31.3

H.8 Distinguishing axis regimes via gate-causality

On the CVD dataset, the temporal gate converges to a near-constant high value (mean 0.998, std $\sim 8{\times}10^{-5}$) while the channel gate retains sample-dependent variation; insertion–deletion is hard to interpret because rankings are nearly tied. The gate-causality sweep (Fig. 6) resolves the ambiguity: CVD shows a flat response (reducing v from 0.999 to 0 costs only 7.3 pts, no threshold), whereas Depression shows a sharp cliff between $v = 0.5$ and $v = 0.6$ (44.1 pts drop). CVD temporal evidence is therefore weakly causal at the gate stage; Depression requires near-open temporal gating. A nearly constant gate can be irrelevant or uniformly necessary; the sweep separates these cases.

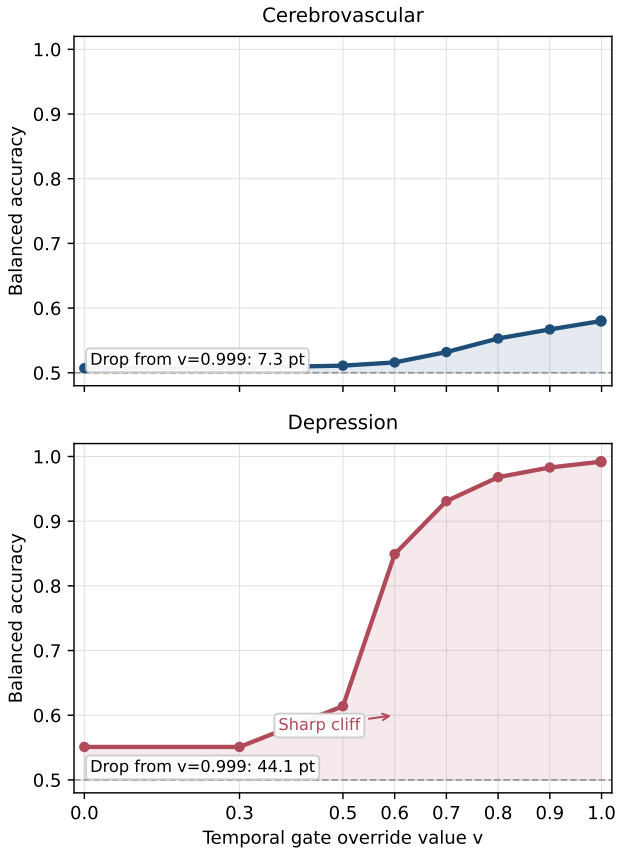


Figure 6: Gate-causality sweep (CVD vs. Depression). Balanced accuracy under temporal gate override v: CVD has a flat response; Depression shows a sharp cliff between $v = 0.5$ and $v = 0.6$.

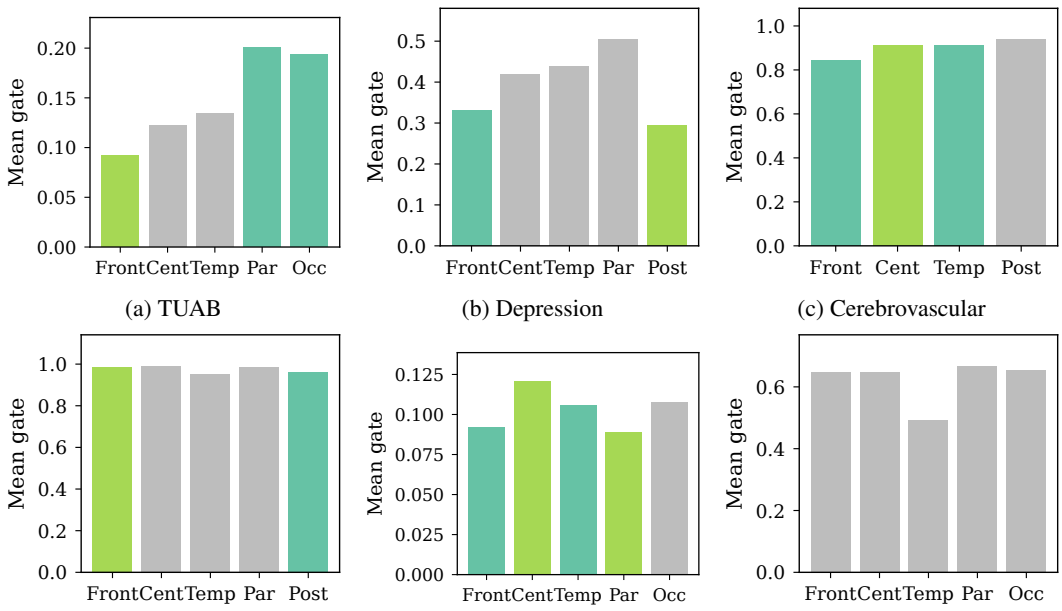
H.9 Per-task region-level gate distributions

This appendix expands Fig. 3 with the full set of region-level visualizations supporting Table 4. Unless stated otherwise, plots are produced from a single seed (s=42) and the held-out test split. The evaluation pipeline groups channels into canonical 10–20 region sets (frontal: AF/Fp/F; central: FC/C/CP; temporal: T/FT/TP; parietal: P; posterior/occipital: PO/O); for the bipolar TUAB and TUSZ montages the same grouping is applied to derivation pairs by majority vote.

Per-task region-averaged gate. Fig. 7 shows the mean channel gate aggregated by region for each task. Highlighted bars mark hypothesized target / control regions per Table 4; gray bars are non-hypothesized regions. TUAB shows a clean parietal/occipital peak with the lowest weight on frontal channels; TUSZ shows a central/occipital-dominant profile that contradicts the focal-onset (temporal/frontal) prior, consistent with the rejected hypothesis in Table 4; TUEV shows a flat profile across central/frontal/parietal/occipital with a temporal dip, consistent with mixed-event heterogeneity; CVD and Depression both show a posterior peak in the per-region average; AD shows a saturated profile near 1.0 with no resolvable contrast.

Correct-only, class-conditioned profiles (used in Fig. 3). Fig. 8 restricts the analysis to samples the model classified correctly and then splits by ground-truth class. This view removes both the class-imbalance confound (the dominant class also contributes most correct predictions) and the prediction-error confound (selector behavior on misclassified samples is hard to interpret). The resulting differences are direct evidence that the selector encodes class-specific anatomical information.

Per-class regional profiles (all samples). Fig. 9 stratifies the regional profile by ground-truth class on *all* test samples (i.e., without the correct-only filter used in Fig. 8). TUAB shows the abnormal class consistently above the normal class at every region, with the largest gap at parietal/occipital (0.24–



(a) TUAB (b) Depression (c) Cerebrovascular

(d) AD staging (e) TUSZ (f) TUEV

Figure 7: Region-averaged mean channel gate per task (single seed s=42). Highlighted bars (teal / lime-green) mark hypothesized target and control regions per Table 4; gray bars are non-hypothesized regions. TUEV uses a per-event-class hypothesis (no single directional target), so all bars are gray. The CVD panel uses a 4-region grouping without Par/Occ. Note the saturated near-1.0 range on AD, indicating a non-discriminative selector at the 4-way granularity, and the central/occipital pattern on TUSZ that contradicts the focal-onset prior.

0.25 vs. 0.15–0.16), mirroring the correct-only view in Fig. 8a; CVD shows a clear posterior shift in the disease class (0.96) relative to normal controls (0.86) and a complementary central/temporal reduction; Depression shows broader elevation at central, parietal, and temporal sites for the depressive class, with parietal carrying the largest class separation (0.58 vs. 0.43); AD class profiles are visibly compressed (all five regions within $\sim$ 0.05 of 1.0 across AD/MCI/SCD/NC); TUSZ separates seizure from non-seizure markedly at every region (seizure $\approx$ 0.18–0.28 vs. non-seizure $\approx$ 0.08–0.11); TUEV separates GPED/PLED (broad, high) from ARTF (broad, low) and EYEM (frontal-peaked).

Correct vs. incorrect breakdown. Fig. 10 compares the regional gate profile of correctly classified samples against incorrectly classified ones, irrespective of class. On TUAB, CVD, and Depression, the correctly predicted subset shows higher gate at the clinically expected target region. On TUSZ the pattern reverses: incorrectly classified samples have markedly higher gate everywhere, suggesting the selector over-recruits when the input is ambiguous. On the imbalanced CVD and Depression splits the contrast partly reflects the fact that the dominant disease class is also the better-classified one, so this view should be read together with Fig. 8. AD shows nearly identical profiles between correct and incorrect, mirroring the saturated 4-way gate even though balanced accuracy is best-in-class.

Lateralization index. Fig. 11 shows the per-sample signed lateralization index histogram (plotted on the horizontal axis as $\mathrm{LI} = (R - L)/(R + L)$, whereas the statistical effect-size analysis in Table 4 evaluates the unsigned magnitude $|\mathrm{LI}| = |(L - R)/(L + R)|$). On TUAB, normal and abnormal class distributions are broad and largely overlapping, consistent with PDR being a bilateral posterior phenomenon rather than a lateralized one. The CVD class shows a substantially broader distribution than normal controls (Mann–Whitney U test on $|\mathrm{LI}|$: $p < 0.001$, n_{CVD}=1,384, n_{normal}=413), consistent with focal lesion-side slowing rather than a fixed hemispheric bias. On AD the histogram is concentrated near zero on a $\pm$0.015 axis, indicating that the selector does not recover a lateralized pattern even within the AD class. TUSZ and TUEV are included for completeness.

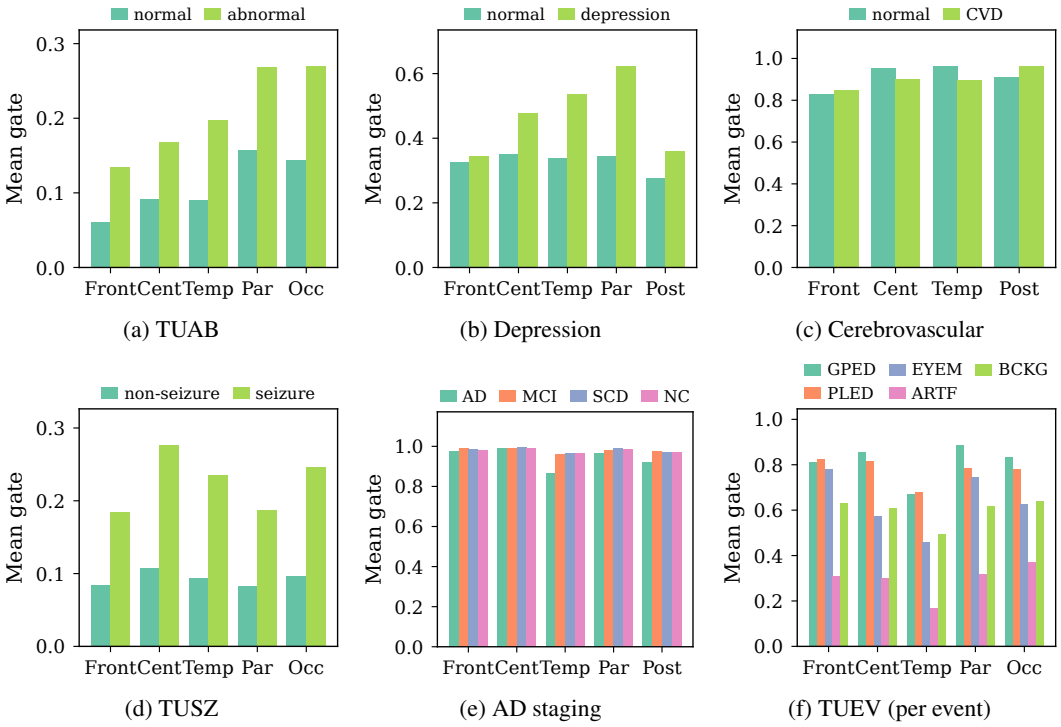


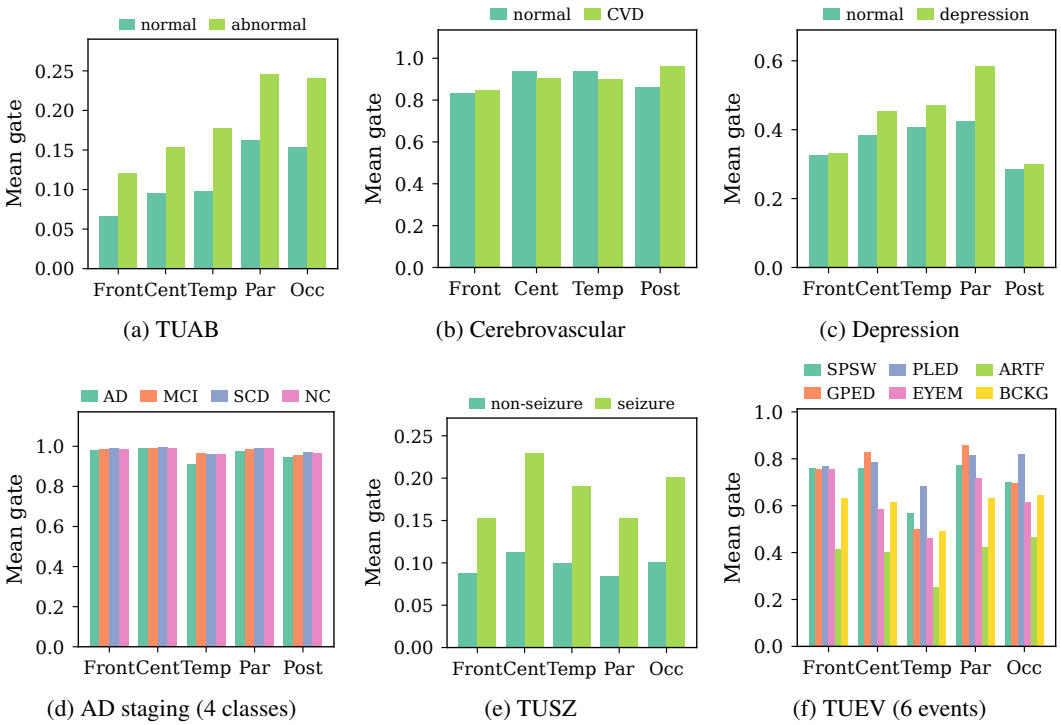
Figure 8: Region-averaged channel gate on correctly classified samples, split by ground-truth class. (a) TUAB: abnormal-class gate $\approx 2\times$ normal-class at every region with the largest absolute gap at occipital/parietal. (b) Depression: parietal carries the largest class separation (0.62 vs. 0.34). (c) CVD: disease class shifts mass posteriorly (posterior 0.96 vs. 0.91 on normal) and reduces central/temporal recruitment. (d) TUSZ: seizure class recruits gate ≈ 2–$3\times$ broader than non-seizure, with central/occipital dominance rather than the focal temporal/frontal predicted. (e) AD staging: AD/MCI/SCD/NC profiles are near-identical due to gate saturation. (f) TUEV: EYEM peaks at frontal (eye-movement origin); GPED/PLED at central; ARTF lowest overall.

H.10 AD vs. NC binary reformulation

The four-way AD/MCI/SCD/NC staging task in Sec. 5.5 is the hardest in-house setting: although tPoE-EIB attains the highest balanced accuracy among heads ($38.5\pm4.1\%$, chance 25%), the per-region channel gate is saturated near 1.0 and the regional prior test is rejected. To test whether this reflects a fundamental inability of the bottleneck to capture AD-relevant evidence in this cohort or simply task granularity, we re-train tPoE-EIB on a binary AD vs. NC reformulation of the same cohort (same 58-channel grid, same frozen CodeBrain backbone, same default hyperparameters; subject-disjoint splits restricted to AD and NC subjects, with n_{AD}=751 and n_{NC}=784 test windows). We then probe the learned channel-gate space with a logistic regression (LR) on the 58-dimensional channel gate $g_c(x)$ and characterize differential gate behaviour across three seeds.

Linear separability of the channel-gate space. Table 15 reports balanced accuracy of a logistic regression trained on $g_c(x)$ produced by the tPoE-EIB head, separately for each training seed. The LR reaches 88.8–97.9% balanced accuracy, indicating that the channel-gate space is approximately linearly separable for AD vs. NC even though the bottlenecked classifier rate-limits the information that can flow from the gates to the prediction. This complements rather than replaces the four-way staging result: at finer label granularity the bottlenecked classifier loses discriminative resolution; at the binary level the gate space already encodes class-discriminative spatial information.

Differential channel-gate behaviour. Mann–Whitney U tests per channel identify channels whose mean gate differs between AD and NC. Across seeds, NC consistently shows higher gate activation than AD on most differentially gated channels: 3-seed mean differences are negative for the top 15 differentiated channels, with absolute gaps of 0.13–0.23. The differentially gated channels span parietal (P2, P5, P6, P7), occipital (O2), temporal–parietal (TP7, TP8), central (C1, C2, C6), and



(a) TUAB (b) Cerebrovascular (c) Depression

(d) AD staging (4 classes) (e) TUSZ (f) TUEV (6 events)

Figure 9: Per-class regional gate profiles on all test samples (no correct-only filter). Each bar gives the mean channel gate within the region, conditioned on ground-truth class.

Table 15: Logistic regression on the 58-dimensional channel gate $g_c(x)$ produced by tPoE-EIB on the binary AD vs. NC reformulation, evaluated per training seed. Top channels are listed by absolute LR weight; sign in parentheses indicates the direction of the AD vs. NC contrast (negative = NC favoured by larger gate value).

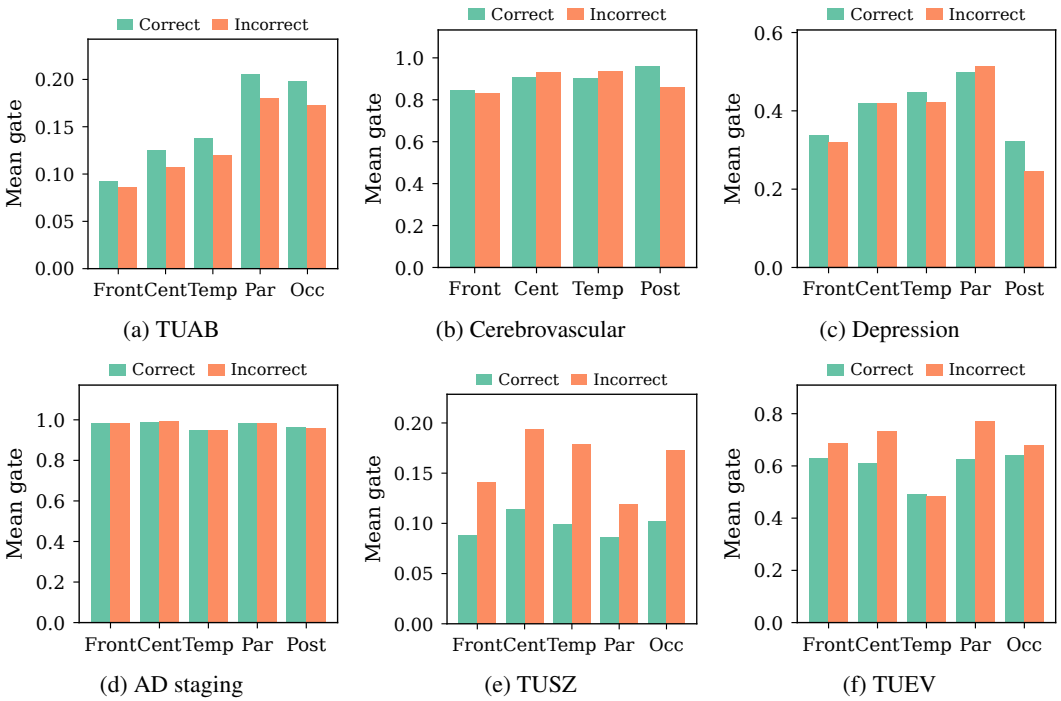
Seed	LR BAcc	Top weighted channels
s42	0.948	POz$(-)$, CP2$(-)$, PO4$(+)$, P8$(+)$, TP8$(-)$
s2025	0.979	T8$(-)$, O2$(-)$, TP8$(-)$, AF3$(-)$, F3$(-)$
s3407	0.888	CP2$(-)$, O1$(+)$, TP7$(+)$, AF4$(-)$, F3$(-)$

frontal (F3, AF3, AF4, F8, FC5) sites rather than concentrating at a single locus. The pattern is therefore inconsistent with a clean "posterior $>$ frontal" rule but consistent with AD as a diffuse rather than focal process.

Fronto-parietal correlation reversal. The most pronounced AD$-$NC differences in pairwise correlations of channel-gate values occur between frontal and parietal channels. Across seeds, the F3–P3, F8–P3, F4–P3, FC5–P3, F8–P6, Fz–P6, FC5–P6, and F7–P7 pairs all show gate-value correlations that flip sign from positive in NC to negative in AD, with $\Delta\text{corr} = \text{corr}_{\text{AD}} - \text{corr}_{\text{NC}} \in [-1.13, -0.84]$ on average. This fronto-parietal decoupling at the gate level is not built into the architecture; it emerges from the data and is broadly consistent with diffuse fronto-parietal disruption reported in the AD EEG literature [41, 42].

Temporal-gate behaviour. Temporal-gate statistics are seed-dependent. In seed s42 the AD class shows a higher temporal-gate mean (0.877 vs. 0.826, $p < 10^{-22}$); in seed s2025 the direction reverses (0.970 vs. 0.976, $p \approx 10^{-3}$); in seed s3407 the temporal gate saturates near 1.0 in both classes with no significant difference. Temporal-gate behaviour therefore does not provide a stable AD vs. NC marker in this cohort, in contrast to the channel-gate pattern.

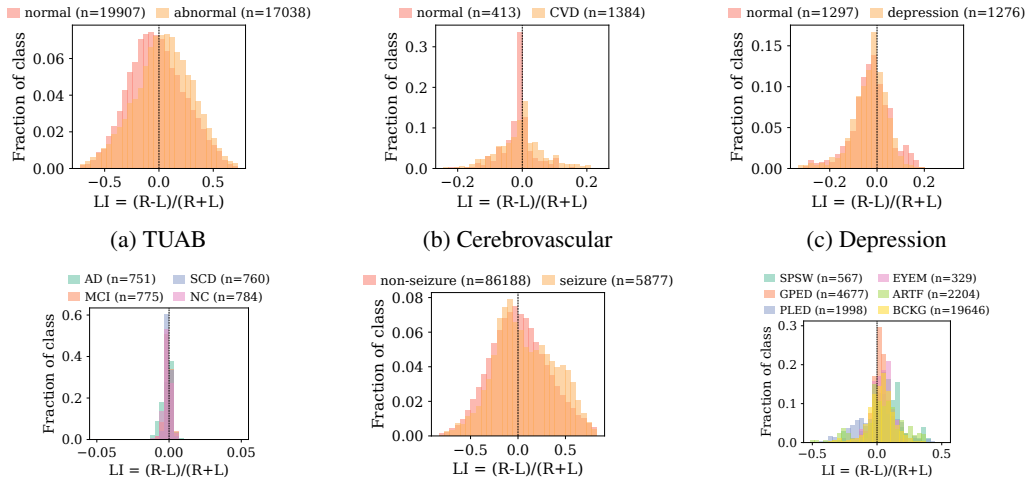
Implication. The four-way staging saturation is driven by label granularity rather than an inability of the bottleneck to capture AD-relevant features in this cohort. On the binary reformulation, the



(a) TUAB (b) Cerebrovascular (c) Depression

(d) AD staging (e) TUSZ (f) TUEV

Figure 10: Region-averaged channel gate, split by prediction outcome. AD shows no separation under the saturated 4-way gate; TUSZ shows the opposite pattern from TUAB/CVD/Depression, with incorrect samples carrying higher gate.



(a) TUAB (b) Cerebrovascular (c) Depression

(d) AD staging (e) TUSZ (f) TUEV

Figure 11: Per-sample signed lateralization index LI histograms (horizontal axis shows signed $(R-L)/(R+L)$; Table 4 tests unsigned $|\mathrm{LI}|$). TUAB normal and abnormal distributions overlap, consistent with bilateral PDR; CVD shows clear broadening in the disease class.

gate space encodes AD vs. NC in a distributed fronto-parietal–occipital pattern with measurable correlation reorganization between frontal and parietal sites. Because this distributed pattern differs from the pre-registered "posterior $>$ frontal" anchor, the Q4 alignment test for AD remains negative while the binary gate representation itself is highly discriminative.

I Per-sample TUEV gate overlays

Fig. 12 shows nine correctly-classified TUEV samples with the trained gates overlaid on the raw signal. Each panel plots the four channels with the highest channel-gate value, annotated with their channel-gate weight; gray bars indicate the temporal-gate selection and the orange shaded region marks the ground-truth event window. The panels illustrate how the temporal gate concentrates around the event annotation while the channel gate selects axes that are visually consistent with the event class (e.g., frontal channels for ocular EYEM events, central/parietal sites for periodic-discharge classes).

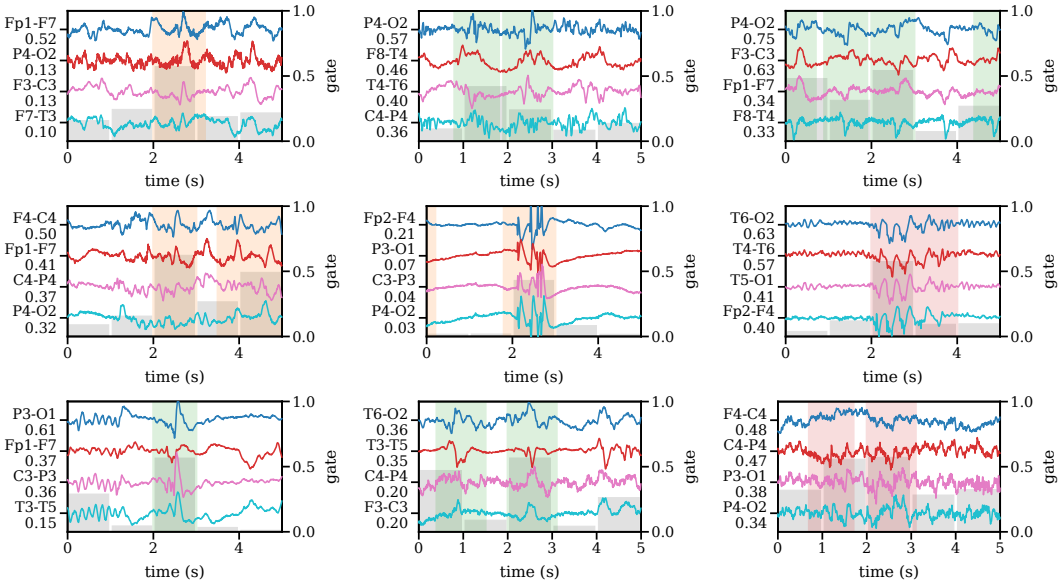


Figure 12: Per-sample TUEV gate overlays on nine correctly-classified pathological events (seed s42). Each panel shows the four highest channel-gate channels with gate weights in parentheses; gray bars mark the temporal-gate selection and the orange band marks the ground-truth event annotation. The selected support concentrates on the event window across diverse event classes and montages.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction (Sec. 3 and Sec. 5) state four contributions—an evidence-only EEG adaptation constraint, a tempered PoE evidence bottleneck, the closed-form theory plus audit protocol, and empirical validation across six EEG diagnosis settings. Each is delivered: Sec. 3 formalizes the constraint and derives the variational EIB bound and closed-form tempered posterior; Sec. 4 defines the audit metrics; Sec. 5 reports the six-task evaluation across two backbones.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The Conclusion (Sec. 6) explicitly notes that the evidence interface is intentionally factorized (marginal temporal/channel rather than joint time–channel patterns), that the bottleneck does not certify clinical correctness, and that group-level audits have blind spots when gate values are nearly uniform or highly patient-specific. We further discuss task-granularity limits (AD 4-way staging) in Sec. 5.5 and Appendix H.10.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Propositions 1 (variational EIB lower bound), 2 (tempering corrects overconfidence under correlated experts), 3 (closed-form tempered PoE posterior), Lemma 1 (stability is not selectivity), and Corollary 1 (rate-limited evidence path) all state assumptions explicitly, and full proofs are provided in Appendix B and Appendix C.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Sec. 5.1 reports the four-head comparison protocol, optimizer, learning rate, latent dimension, KL/sparsity/consistency coefficients, epochs, patience, and seeds. Appendix D gives the full hyperparameter table and architectural details for the gates and experts. Appendix E gives dataset preprocessing, splits, and channel montages. Code is released at the official repository (`https://github.com/W-Yinghao/tPoe_EIB`).

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code is released at `https://github.com/W-Yinghao/tPoe_EIB` with instructions to reproduce all reported results. Of the six datasets, the three public TUH benchmarks (TUAB, TUEV, TUSZ) are available through standard application to the Temple University Hospital data portal. The three in-house clinical cohorts (AD/MCI/SCD/HC, Depression, Cerebrovascular) are not publicly available due to the underlying clinical data agreements; researchers may request access by contacting the corresponding authors and will be granted controlled access after institutional review.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Sec. 5.1 states the train/test protocol (subject-disjoint splits before windowing, 3 seeds, AdamW with lr 10^{-3}, 50 epochs, patience 15) and all main hyperparameters. Default hyperparameters (Table 6) and the temperature/sparsity/consistency coefficient choices are documented in Appendix D. Per-dataset preprocessing, channel montages, and split sizes are in Appendix E.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All headline metrics in Tables 1, 2, and 3 report 3-seed mean$\pm$std (seeds 42, 2025, 3407). Faithfulness gaps in Sec. 5.4 carry standard deviations across seeds. Clinical alignment in Table 4 reports 3-seed Cohen's d effect sizes with one-sided Wilcoxon decisions per seed; for the lateralization-index analysis, we additionally report Mann–Whitney U test p-values (Appendix H.9).

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: Across the project we ran experiments on a heterogeneous set of GPUs (NVIDIA H100, A100, L40S, V100, and P100). Because the head is small (two-layer MLP gates and Gaussian expert projections, d_z=64) and operates on top of a frozen backbone, the head itself trains on a single consumer GPU within hours per dataset, and the total compute cost is not materially different from standard fine-tuning of EEG foundation backbones; we therefore did not view a per-experiment compute budget as central to the methodology and omitted it from the main paper. If reviewers find a detailed breakdown valuable, we will report per-task GPU-hour estimates in the camera-ready version.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses de-identified EEG recordings under approved data agreements, and the in-house clinical experiments were reviewed and approved by the relevant ethics committees (see also IRB question below). No human-subject experiments were conducted by the authors as part of this submission; we worked with already-collected data under existing approvals.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The intended positive impact is auditable EEG diagnosis: by routing predictions through a rate-limited, axis-specific evidence interface, the method makes it possible for clinicians to inspect which time segments and channels supported a decision. Potential negative impacts include over-reliance on automated diagnosis when audit signals are interpreted as clinical sufficiency—the paper explicitly cautions (Sec. 3.1, Conclusion) that the bottleneck does not certify clinical correctness, and the audits in Sec. 4 are intervention-based diagnostics, not clinical guarantees.

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release pre-trained generative models, scraped data, or other assets with elevated misuse risk. The released code is an EEG diagnosis adaptation head; it operates on top of frozen public EEG backbones and does not generate content. Clinical EEG datasets are released through controlled application processes (Q5 above) rather than open download.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All EEG backbones and baseline methods (CodeBrain, CBraMod, LaBraM, BIOT, BENDR, ContraWR, EEGNet, EEG Conformer) are cited in Sec. 5 and the Related Work section, and the public TUH datasets (TUAB, TUEV, TUSZ) are cited where introduced. We use the public model weights and dataset licenses under their respective public-research terms.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The released code at `https://github.com/W-Yinghao/tPoe_EIB` contains the tPoE-EIB head implementation, training scripts, and audit utilities (insertion–deletion, expert-drop, gate-causality, posterior-reliance, expert-disagreement) with a README documenting environment setup and reproduction commands. We do not release new datasets.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or new research with human subjects. All data are pre-collected EEG recordings used under existing data agreements and prior ethics-board approvals.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [Yes]

Justification: TUAB, TUEV, and TUSZ are public TUH EEG benchmarks. The private clinical EEG cohorts used in this study are covered by institutional ethics approvals (approval Nos. 81501155, ZYLX201607, and KD2022165), as reported in Appendix E. Only de-identified windowed EEG segments and derived record metadata enter the modeling pipeline.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?

Answer: [N/A]

Justification: LLMs are not a component of the core method or experiments. We used LLMs only for language polishing of the manuscript draft; the methodology, theoretical results, experiments, and analysis were entirely conducted by the authors.