EXACT TEMPORAL READOUTS EXPOSE AND TEST A TWO-BOUNDARY DECISION PATTERN IN IMAGINED-SPEECH EEG

Anonymous authors
Paper under double-blind review

ABSTRACT

Temporal aggregation can determine an imagined-speech EEG prediction, yet its role is usually opaque. We introduce an exact temporal readout framework built around a bidirectional exponential-memory mixture (BEM): each trial logit is a normalized, analytically recoverable kernel over frame-level class evidence anchored at both trial boundaries. On BCI Competition 2020 Track 3, BEM raises development accuracy from $77.07 \pm 1.01\%$ to $85.73 \pm 1.57\%$ and gains 7.11 participant-paired percentage points on a training-excluded historical audit. Its exact kernels place 89.67% of their mass in the outer fifths. Distilling this pattern into endpoint-fifths pooling yields no detected difference from BEM; applying the rule to the *same frozen mean-trained evidence tensors* gains 4.13 points on development and 2.98 on audit, isolating aggregation from encoder retraining. Boundary-masking and filtering controls further support reliance on broad two-edge structure. In a separate NEMAR on007591 calibration-to-online stress test, models were locally frozen before EEG or decoder-outcome inspection, although event labels and official quality-control metadata were known. The declared raw-accuracy gate passes: all three participants favor BEM, with a 4.67-point mean gain. However, balanced accuracy falls for all three, and a disclosed post-outcome same-evidence endpoint transfer loses 2.67 points. Exact readouts therefore turn a temporal preference into a distillable, intervenable, and externally falsifiable hypothesis, while the mixed stress test rules out a general two-boundary law.

1 INTRODUCTION

Imagined-speech EEG classifiers must decide from weak neural measurements without an observed speech event that anchors time. A trial boundary marks an experimental window, not necessarily the onset of a stable neural event. Class evidence may therefore be diffuse, participant dependent, or concentrated near either edge. In this low-data setting, the final operation that turns a temporal evidence path into one decision is part of the model's inductive bias, yet it is rarely examined directly.

Many EEG decoding systems instead emphasize the representation learner: spatial–temporal convolutions, attention, metric learning, and state-space models seek more discriminative latent features (Lee et al., 2021a; Li et al., 2021; Song et al., 2023; Gui et al., 2024). Once frame representations exist, mean pooling assumes equal relevance at every time step, while flexible attention or recurrence can be difficult to identify from a few trials and does not expose a simple temporal law. This leaves a practical question unanswered: *which temporal assumption is a trained classifier actually using, and does that assumption itself affect the decision?*

We answer that question with an interpretable probe. The bidirectional exponential-memory mixture (BEM) places a probability measure over stable exponential decays evaluated from both trial boundaries. Each class combines these summaries with a uniform atom, producing an exact nonnegative kernel whose weights sum to one. Unlike a post-hoc saliency map, this kernel is the evidence-to-logit operator itself. Its directional branches are completely monotone Hausdorff-moment sequences, and the complete operator is normalized, stable, and computable in linear time. More importantly for this study, the parameterization turns a learned temporal preference into a concrete rule that can be distilled, intervened on, and tested with the encoder held fixed.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

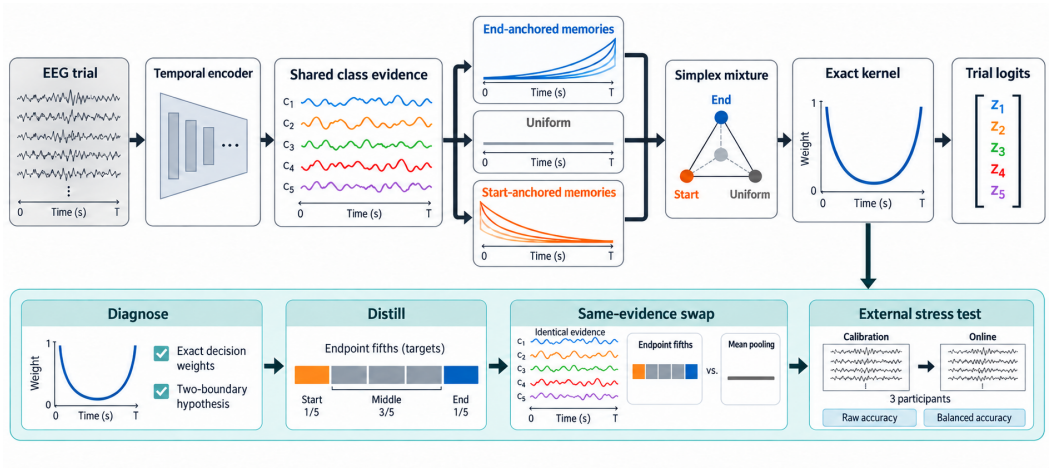


Figure 1: From exact readout to a testable temporal hypothesis. Stable memories anchored at both boundaries and a uniform component form an exact normalized kernel. Its broad shape is distilled into endpoint fifths, compared with mean pooling on identical evidence, and stress-tested externally.

On the 15-participant BCI Competition 2020 benchmark, the probe reveals a strong U-shaped kernel: 89.67% of its mass lies in the first and last fifths of the encoded sequence. This observation leads to a fixed endpoint-fifths rule for which no difference from BEM is detected. The central isolation experiment then freezes the mean-trained encoder and evidence head and changes only the aggregation applied to the resulting evidence tensor. Endpoint pooling gains $4.13\,pp$ (95% CI $[1.78, 6.49]\,pp$) on development and $2.98\,pp$ ($[1.29, 4.76]\,pp$) on the audit, with 13 of 15 participants improving on each. The conclusion is structural: two-boundary aggregation changes predictions on these archived evidence paths. BEM is valuable because its exact kernel exposed this testable hypothesis; no advantage over the distilled endpoint rule is detected.

Retrospective structure is not external evidence. After inspecting public event tables and official bad-channel metadata, but before opening any EEG samples or decoder outcomes from the independent NEMAR on007591 release, we locally froze the preprocessing, compact encoder, two readouts, three seeds, online outcome, and success criterion. The frozen raw-accuracy gate passes: BEM improves all three participants and gains $4.67\,pp$ on average. Yet the participant-level interval includes zero and balanced accuracy decreases for all three. This metric disagreement is scientifically useful: it prevents a nominally positive gate from being mistaken for robust external confirmation.

We organize the experiments by evidentiary role rather than treating every comparison as another benchmark row (table 1). Learned kernels generate the hypothesis; fixed rules and identical-evidence swaps isolate its aggregation component; edge/center interventions and filtering controls test within-protocol consistency; and a pre-outcome-frozen online-acquisition stress test plus encoder and corpus experiments locate the boundary of the result.

This paper makes three contributions:

- It turns an opaque pooling choice into an evidence ladder: exact-kernel discovery, fixed-rule distillation, and a same-evidence intervention that isolates a positive aggregation effect from encoder retraining.
- It introduces an exact normalized readout whose kernel is the decision operator itself, with convex-hull and fixed-kernel Lipschitz guarantees, directional complete monotonicity, and $O(TAC)$ evaluation.
- It stress-tests the structured-readout advantage on an online-acquisition split under a locally hash-frozen, label-aware protocol. The raw-accuracy gate passes but balanced accuracy reverses, while encoder and corpus controls further bound the claim. We report this mixed outcome rather than relabel it as a win.

2 Related Work

Imagined-speech EEG decoding. Imagined-speech EEG combines low signal-to-noise ratio, limited training data, substantial between-participant variation, and heterogeneous evaluation protocols (Panachakel & Ramakrishnan, 2021; Lopez-Bernal et al., 2022). Classical pipelines use covariance geometry or engineered spectral features (Nguyen et al., 2018; Bakhshali et al., 2020); recent systems learn spatial–temporal representations with dilated convolution, metric learning, or attention (Lee et al., 2021a; Li et al., 2021; Lee et al., 2022). Compact EEGNet is a compact architecture shown to perform competitively with limited training data (Lawhern et al., 2018), while EEG Conformer and bidirectional state-space models introduce more flexible sequence encoders (Song et al., 2023; Gui et al., 2024). These approaches primarily change the feature extractor or training objective. Our question begins after the encoder has produced class evidence and isolates the law that aggregates it. Recent imagined-speech work has also explored task-specific Transformers, multi-corpus transfer, and subject-invariant augmentation (Şahin & Özdemir, 2025; Duhair et al., 2026; Dörterler et al., 2026). Because their datasets and validation regimes differ, we use them to locate the problem rather than make numerical rank claims. A recent benchmark of two general-purpose EEG foundation models likewise reports no consistent advantage over task-specific baselines across two speech corpora (Khanday et al., 2026). Closest domain precedents localize imagined-speech evidence over time using integrated gradients or Grad-CAM plus masking (Sato et al., 2024; Balasubramanian et al., 2025). We instead distill and replace the exact implemented evidence-to-logit operator while holding its input tensor identical.

Temporal pooling and evidence accumulation. Uniform mean and maximum pooling are parameter free but impose strong timing assumptions. Attention learns normalized frame weights, and deep adaptive temporal pooling applies input-dependent weights directly to segment scores (Song et al., 2018). Learnable dynamic temporal pooling instead segments latent sequences by DTW alignment to learned prototypes (Lee et al., 2021b). Recurrent readouts offer still greater freedom, echoing classical leaky evidence accumulators (Usher & McClelland, 2001). BEM is closest in placement to score pooling, but differs in its hypothesis class: it learns a positive measure over shared decay atoms, acts separately on each class-evidence path, and combines two fixed temporal orientations. Our comparisons place attention, free temporal weights, recurrence, and a diagonal state-space model at the same class-evidence interface with an otherwise identical architecture, avoiding an encoder-level confound.

Exponential memory and state-space sequence models. Exponentially weighted averages are a long-standing mechanism for stable memory (Roberts, 1959). Modern sequence models embed multiple decay scales inside gated attention or linear state dynamics: MEGA uses a multidimensional damped moving average (Ma et al., 2023); S4 and its diagonal variants parameterize efficient state-space layers (Gu et al., 2022b;a; Gupta et al., 2022); and the Linear Recurrent Unit uses stable diagonal recurrences (Orvieto et al., 2023). ETSMLP gives the closest explicit predecessor to our two-sided exponential construction: it describes forward and backward exponential-smoothing kernels as a token-mixing module (Chu & Lin, 2024). Those modules transform latent representations and can use signed or complex modes. BEM instead uses real poles and a positive normalized measure at the final decision layer, so its full evidence-to-logit kernel is explicit. We therefore do not claim bidirectional exponential smoothing itself as novel; the contribution is its decision-layer measure and the discovery–distillation–intervention protocol that the exact kernel enables.

Hausdorff moment structure. A sequence is completely monotone exactly when it is the moment sequence of a finite positive measure on $[0, 1]$ (Hausdorff, 1921; Shohat & Tamarkin, 1943; Schilling et al., 2012). Positive mixtures of stable exponentials therefore give a discrete, interpretable memory cone. We use this structure for each directional branch. Nonnegative mixtures of decaying exponentials have also enforced complete monotonicity for learned copula generators (Ling et al., 2020); their object is a continuous copula generator rather than a discrete evidence-to-decision kernel. Time reversal generally breaks complete monotonicity, so the bidirectional sum is characterized instead by nonnegativity and normalization. This distinction separates our construction from claims about a globally monotone U-shaped kernel.

3 BIDIRECTIONAL EXPONENTIAL-MEMORY MIXTURES

3.1 FROM FRAME EVIDENCE TO A TRIAL DECISION

Let $\mathbf{x}_{1:T} \in \mathbb{R}^{D\times T}$ be an EEG trial and let an encoder f_θ produce features $\mathbf{h}_t \in \mathbb{R}^H$. A shared evidence head maps each feature to class evidence $\mathbf{e}_t = g_\theta(\mathbf{h}_t) \in \mathbb{R}^C$. The remaining problem is to turn the path $e_{1:T,c}$ into one logit for class c. Mean pooling fixes the temporal kernel to $1/T$; an unconstrained readout must instead estimate an arbitrary aggregation law. We restrict this law to a probability mixture of stable exponential memories evaluated from both trial boundaries.

For a learnable pole $\beta_a = \sigma(\rho_a) \in (0, 1)$, define end- and start-anchored atoms, in the same order used by the implementation,

$$k_a^{\mathrm{E}}(t) = \frac{\beta_a^{T-t}}{Z_a}, \qquad k_a^{\mathrm{S}}(t) = \frac{\beta_a^{t-1}}{Z_a}, \qquad Z_a = \sum_{r=0}^{T-1} \beta_a^r. \tag{1}$$

Small poles concentrate mass at a boundary, whereas poles near one approach a long memory. The same A poles are shared across classes and directions. Their summaries are

$$s_{a,c}^d = \sum_{t=1}^{T} k_a^d(t) e_{t,c}, \quad d \in \{\mathrm{E}, \mathrm{S}\}, \qquad s_c^{\mathrm{U}} = \frac{1}{T}\sum_{t=1}^{T} e_{t,c}. \tag{2}$$

3.2 A CLASS-SPECIFIC MEASURE OVER MEMORIES

Each class learns logits $q_c \in \mathbb{R}^{2A+1}$. A softmax converts them to a simplex measure

$$\boldsymbol{\pi}_c = \mathrm{softmax}(\mathbf{q}_c), \qquad \sum_{d,a} \pi_{a,c}^d + \pi_c^{\mathrm{U}} = 1, \quad \boldsymbol{\pi}_c \geq 0, \tag{3}$$

where the sum over (d, a) covers the end- and start-anchored banks. The trial logit is

$$z_c = \gamma \left(\sum_{a=1}^{A} \pi_{a,c}^{\mathrm{E}} s_{a,c}^{\mathrm{E}} + \sum_{a=1}^{A} \pi_{a,c}^{\mathrm{S}} s_{a,c}^{\mathrm{S}} + \pi_c^{\mathrm{U}} s_c^{\mathrm{U}} \right), \qquad \gamma = \exp(\eta) > 0. \tag{4}$$

Equivalently, equation (4) applies the exact class-specific kernel

$$w_c(t) = \sum_{a=1}^{A} \pi_{a,c}^{\mathrm{E}} k_a^{\mathrm{E}}(t) + \sum_{a=1}^{A} \pi_{a,c}^{\mathrm{S}} k_a^{\mathrm{S}}(t) + \frac{\pi_c^{\mathrm{U}}}{T}, \qquad z_c = \gamma \sum_{t=1}^{T} w_c(t) e_{t,c}. \tag{5}$$

Thus the readout changes only how a pre-existing class-evidence path is aggregated. Mean pooling and BEM both learn the scale γ; with $A = 12$ and $C = 5$, the poles and simplex logits therefore add $A + (2A + 1)C = 137$ parameters over mean pooling. The shared encoder and evidence head are unchanged.

3.3 STRUCTURE AND COMPUTATION

Proposition 1 (Normalized exact kernel)**.** *For every class c, $w_c(t) \geq 0$ and $\sum_{t=1}^{T} w_c(t) = 1$. Consequently, z_c/γ lies in the convex hull of $\{e_{t,c}\}_{t=1}^{T}$.*

Proof sketch. Every atom in equation (1) and the uniform component is a normalized nonnegative vector. Equation (3) takes a convex combination of these vectors. Substitution into equation (5) gives the convex-hull statement. A complete proof is given in appendix A. □

Proposition 2 (Directional complete monotonicity)**.** *Fix $T \geq 1$ and use boundary lag $r = T - t$ for the end branch and $r = t - 1$ for the start branch. Extend either branch from its first T values by $K_c(r) = \sum_{a=1}^{A} \alpha_{a,c}\beta_a^r/Z_a$ for $r \geq 0$, $\alpha_{a,c} \geq 0$, and $0 < \beta_a < 1$. With the forward difference $\Delta K(r) = K(r + 1) - K(r)$, $(-1)^m \Delta^m K_c(r) \geq 0$ for all integers $m, r \geq 0$; hence K_c is a Hausdorff-moment sequence and the implemented branch is its first T terms.*

Proof sketch. Repeated forward differences give $(-1)^m \Delta^m K_c(r) = \sum_a (\alpha_{a,c}/Z_a)\beta_a^r(1-\beta_a)^m \geq 0$. In a common end-lag coordinate, the complete kernel is an end-lag branch plus the finite reversal of a generally different start-lag branch, plus the uniform atom. It therefore need not itself be completely monotone; the appendix gives an exact counterexample within the model family. □

Proposition 3 (Non-expansion before scaling)**.** *Holding the learned parameters, and hence w_c, fixed, for two evidence paths e_c and e'_c,*

$$\left| \frac{z_c(e_c)}{\gamma} - \frac{z_c(e'_c)}{\gamma} \right| \leq \|e_c - e'_c\|_\infty. \tag{6}$$

The scaled logit is therefore γ-Lipschitz under the same norm.

Computing all atom summaries directly costs $O(BTAC)$ operations for a batch of size B and requires no quadratic temporal interaction. A pair of forward recurrences also evaluates both directions in $O(TAC)$ time and $O(AC)$ auxiliary state per trial, excluding encoder activations. These properties constrain the evidence-to-decision map; they do not constrain the encoder that produces the evidence.

4 Experimental Protocol

4.1 Datasets and split roles

BCI Competition 2020 Track 3. The primary benchmark contains 64-channel EEG sampled at 256 Hz from 15 participants imagining five communication commands (BCI Competition Committee, 2020; Jeong et al., 2022). In NEMAR's version-1.0.0 BIDS conversion, the official training, validation, and test sets are named run-00, run-01, and run-02, respectively, and the run-02 event labels were obtained from `Track3_Answer Sheet_Test.xlsx` (Lee et al., 2026). Each participant contributes 300, 50, and 50 trials to these partitions, respectively, giving 4,500/750/750 pooled trials. We train on run 00 and choose checkpoints on run 01. Run 02 is excluded from training and selection, but its labels were available in the project cache and its results were inspected during earlier development; we therefore call it a historical audit, not a pristine blind test. The official competition test labels are not exposed in the original OSF deposit. All three runs contain the same participants, so this is pooled cross-run evaluation rather than subject-independent transfer.

Each cached trial contains 795 samples and is filtered by a 50-Hz IIR notch, then common-average referenced. Channelwise means and standard deviations are estimated on run 00 only and applied to all splits. Because zero-phase filtering is performed on finite trials, we separately test a causal notch and no-notch/common-average-reference preprocessing, as well as frozen-checkpoint edge perturbations.

External scope tests. We test the six-participant NEMAR ASU short–long subset and the 21-participant, 16-class FEIS corpus (Nguyen et al., 2018; 2026; Wellington & Clayton, 2019; Clayton et al., 2020). Participant-wise nested cross-validation keeps epoch selection and normalization inside each outer training fold; complete dataset counts, split construction, and aggregation rules are in appendix B.

Pre-outcome-frozen online-acquisition stress test. NEMAR on007591 provides covert-speech calibration and online acquisitions for three participants, 128 EEG channels, and five color words (Sato et al., 2026). Public event tables and official signal-derived bad-channel metadata were known when the analysis was designed. Before opening an EDF signal sample or decoder outcome, we locally hash-froze preprocessing, a grouped calibration split, the compact encoder, mean and eight-atom BEM readouts, seeds 7/42/2027, and event-level online evaluation. The primary gate requires all three seed-mean accuracy differences to be positive and their participant mean to be at least $3\,pp$; balanced accuracy and macro-F1 are secondary. This is a locally time-ordered analysis freeze, not a public preregistration or new prospective acquisition. Exact preprocessing, event aggregation, archive identity, and the metadata erratum are documented in appendix B. Since $n = 3$, even three positive signs give one-sided exact $p = 0.125$.

After opening those outcomes, disclosed post-hoc diagnostics add three matched readouts and exact deterministic replays before swapping rules on identical mean-model evidence; they test mechanism but cannot amend the frozen gate.

4.2 Encoder and matched readout architectures

The primary encoder applies a width-11 temporal convolution ($64 \to 256$), average pooling by two, a pointwise expansion to 320 channels, and a width-9 dilated convolution back to 64 channels followed by average pooling by four. A $64 \to 64 \to 5$ pointwise head produces the 99-step class-evidence path. BEM uses $A = 12$ poles. Its 453,007 trainable parameters exceed mean pooling by 137, precisely the pole and measure parameters in section 3.

All primary baselines consume only the class-evidence tensor produced by an identical encoder and evidence-head architecture; each complete model is trained end-to-end from its fixed seed. We compare uniform mean, temporal maximum, fixed first/last-fifth endpoint pooling, classwise attention, class-specific free temporal weights, a bidirectional evidence-GRU, and a causal stable diagonal evidence-SSM. Structural controls include a fixed symmetric U-shaped kernel, fixed versus learned poles, and start-plus-uniform and end-plus-uniform mixtures. A compact EEGNet-style encoder tests whether the readout effect survives an encoder transplant. The appendix gives exact layers and parameter counts.

The endpoint control is diagnostic rather than an untouched confirmatory baseline: its one-fifth width was chosen after inspection of the archived BEM kernels, but was not selected by training or evaluating an endpoint-width sweep. We therefore use it to test what structure explains the archived gain, not to claim an independently prespecified benchmark win. The frozen edge/center perturbation family was motivated by the same kernel inspection and is likewise post-hoc. To separate aggregation from end-to-end encoder differences, an additional post-hoc diagnostic loads the archived BEM-, mean-, and endpoint-trained checkpoints, freezes each encoder and evidence head, and applies the trained, uniform-mean, and endpoint-fifths rules to the identical evidence tensor produced in a single forward pass. It fits no parameters and performs no additional checkpoint selection.

4.3 Optimization and inference

For the primary full-encoder BCI2020 comparisons, every matched model trains for 350 epochs with AdamW, batch size 256, learning rate 3×10^{-4} cosine-decayed to 3×10^{-5}, weight decay 2×10^{-4}, dropout 0.25, and label smoothing 0.15. Checkpoints are evaluated every ten epochs and selected first by run-01 accuracy, then macro-F1. We use fixed seeds 7, 42, and 2027 with deterministic CuDNN execution. External-corpus models use class-weighted cross-entropy, batch size 32, at most 250 epochs, and the fold-nested selection rule described above. The on007591 models use the same optimizer schedule but select from the single frozen calibration-development partition and evaluate the online acquisition exactly once per run.

We average reported accuracy and macro-F1 over seeds; where uncertainty is shown, it is the sample standard deviation across seeds. For participant-paired inference, we first average accuracy over seeds within each participant, leaving participants as the independent units. We use 100,000 deterministic participant bootstrap resamples for 95% intervals; paired t and Wilcoxon tests provide parametric and rank-based checks. Exact McNemar tests compare trial correctness separately within each seed. The reported focal Holm family contains the eight BEM comparisons against mean, maximum, attention, free temporal weights, endpoint fifths, evidence-GRU, evidence-SSM, and fixed U, separately for each split and test. Each post-hoc edge-sensitivity family contains the three mask widths for one model and split. Participant-bootstrap intervals are conditional on the three archived checkpoint pairs, pooled training, checkpoint selection, and, where applicable, the already chosen post-hoc rule. They quantify heterogeneity among the 15 included participants rather than uncertainty from every stage of training and rule selection. Raw predictions, targets, participant identifiers, and histories are archived for every reported run, together with BCI2020 checkpoints. External nested-CV records additionally retain the outer and inner split indices, every fold's selected epoch, and predictions; fold-specific model states are not retained.

5 Results

5.1 Exact kernels expose a two-boundary pattern

BEM serves first as the hypothesis-generating probe. Relative to mean pooling, it raises development accuracy from 77.07% to 85.73% and the historical audit from 72.40% to 79.51%. After averaging

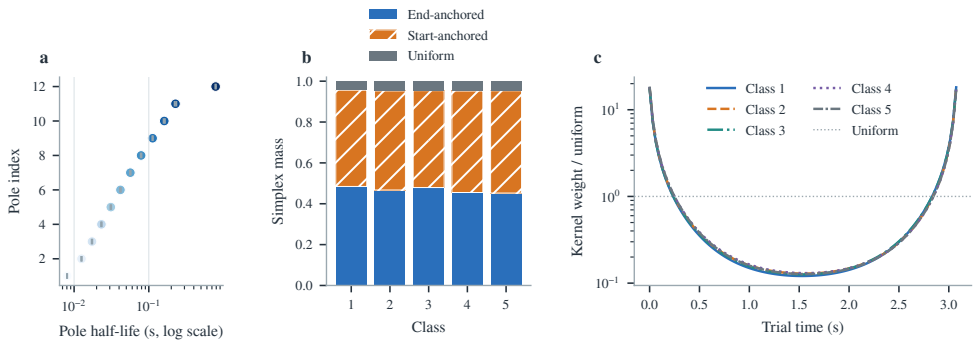


Figure 2: The exact BEM kernel generates the two-boundary hypothesis. (a) Learned pole half-lives; bars show ± sample SD across seeds. (b) Class-specific simplex mass assigned to the two directional banks and uniform atom. (c) Effective class kernels relative to uniform weighting. The U shape is part of the evidence-to-logit operator itself.

seeds within participant, the gains are $8.67\,pp$ (95% CI $[6.13, 11.16]\,pp$) and $7.11\,pp$ ($[4.62, 9.69]\,pp$); 14 of 15 participants improve on each split.

Expanding each archived checkpoint with equation (5) reveals where the probe obtains this gain. On average, 89.67% of kernel mass falls in the first and last 19 of 99 encoded steps, compared with 38.38% under a uniform kernel. End- and start-anchored banks receive nearly balanced mass, while the uniform atom receives less than 5% (figure 2). These are exact decision weights rather than a saliency approximation.

The directional controls test whether either edge alone explains the pattern. End-only and start-only models reach at most 68.93% development and 66.67% audit accuracy; adding a uniform component improves them but remains well below the bidirectional readouts (table 2). Thus both directions are required in these matched implementations; neither boundary alone reproduces the primary performance.

5.2 Distillation and frozen-evidence swaps isolate aggregation

The kernel suggests a simpler surrogate: uniformly average the first and last fifths of the evidence path. When trained end-to-end, this endpoint rule reaches $85.02 \pm 1.08\%$ development and $80.44 \pm 1.21\%$ audit accuracy (table 3). Its participant-paired gains over mean pooling are $7.96\,pp$ (95% CI $[5.60, 10.13]\,pp$) and $8.04\,pp$ ($[6.27, 9.87]\,pp$). BEM-minus-endpoint is $+0.71\,pp$ ($[-0.62, 2.27]\,pp$) on development and $-0.93\,pp$ ($[-2.53, 0.62]\,pp$) on audit. No advantage of learning the exponential mixture over this distilled rule is detected.

Matched retraining compares complete systems and can mix aggregation with encoder/readout co-adaptation. We therefore apply mean and endpoint pooling to the identical evidence tensor from each archived checkpoint. On mean-trained evidence, endpoint-minus-mean is $+4.13\,pp$ (95% CI $[1.78, 6.49]\,pp$) on development and $+2.98\,pp$ ($[1.29, 4.76]\,pp$) on audit, with 13 of 15 participants positive on both (figure 4a). On BEM-trained evidence, the corresponding gains are $+6.04\,pp$ ($[3.96, 8.04]\,pp$) and $+4.98\,pp$ ($[2.93, 7.20]\,pp$). This swap isolates a positive aggregation effect conditional on already learned evidence paths; it does not attribute the entire end-to-end difference to pooling.

A fixed symmetric U kernel also captures most of the improvement, and fixing all 12 poles at initialization changes accuracy by only $+0.09\,pp$ ($[-0.71, 0.89]\,pp$) and $-0.36\,pp$ ($[-1.29, 0.49]\,pp$). Together, these controls locate the supported result in broad two-boundary shape rather than adaptive pole placement.

5.3 Interventions triangulate boundary-cue reliance

Frozen-checkpoint interventions test the prediction made by the exact kernel. For BEM, masking both raw-signal edges by 5/10/20% rather than masking an equal total width at the center causes an addi-

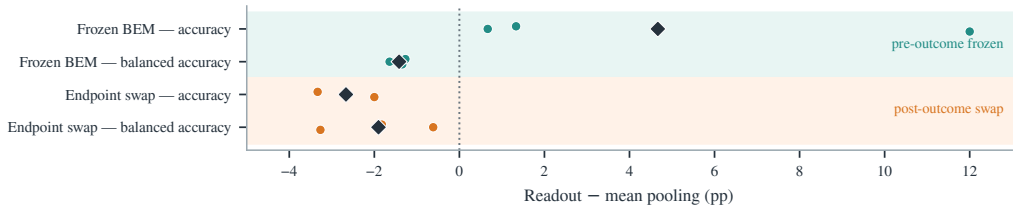


Figure 3: The external stress test separates two non-comparable analyses. Dots are participants and diamonds are means ($n = 3$): upper rows are the pre-outcome-frozen learned-BEM comparison; lower rows are disclosed post-outcome endpoint swaps on identical evidence.

tional development loss of $16.09/34.67/52.44\,pp$ and audit loss of $13.87/34.49/48.13\,pp$. All six participant-mean edge-versus-center effects are positive. Relative to the same contrast for mean pooling, BEM's excess sensitivity is $14.13/30.84/46.62\,pp$ on development and $12.04/31.42/42.31\,pp$ on audit; all participant-bootstrap intervals exclude zero. Because the mask family was motivated by the observed kernels, these interventions triangulate the archived models' reliance rather than provide an independent confirmation; the complete perturbation matrix is in figure 7; the participant-level dose response is shown in figure 4b.

Retraining with different filtering choices probes a distinct alternative. With common-average reference but no notch, the BEM-minus-mean gains are $8.40\,pp$ (95% CI $[6.27, 10.44]\,pp$) and $7.29\,pp$ ($[5.42, 9.33]\,pp$). With a causal 50-Hz notch they are $8.27\,pp$ ($[5.38, 11.33]\,pp$) and $8.36\,pp$ ($[5.64, 11.56]\,pp$). Endpoint pooling remains similarly strong in both settings. The pattern is therefore not explained solely by the primary zero-phase notch implementation.

5.4 A PRE-OUTCOME-FROZEN STRESS TEST GIVES A METRIC-DEPENDENT RESULT

All 18 on007591 runs complete the frozen 3 participants $\times$ 2 readouts $\times$ 3 seeds design. Recomputing every metric from the stored targets and predictions reproduces the run records exactly. The primary raw-accuracy gate passes: participant differences are $+12.00$, $+0.67$, and $+1.33\,pp$, so all three are positive and the participant-mean gain is $+4.67\,pp$ (table 4). With $n = 3$, however, the one-sided exact sign-test value is $p = 0.125$, the one-sided paired t value is $p = 0.166$, and the 95% t interval is $[-11.13, 20.46]\,pp$.

The online event labels are imbalanced, and the secondary metrics prevent a simple replication claim. Balanced accuracy changes from 25.60% to 24.18% and decreases for every participant; macro-F1 changes from 18.97% to 21.23%. BEM is also below the per-participant majority accuracy for Participants 1 and 3. The pre-outcome-frozen result is therefore mixed: it satisfies the declared accuracy rule but does not confirm a class-balanced two-boundary advantage (figure 3).

Post-outcome controls sharpen that interpretation (table 5). Fixed-pole BEM retains a $4.00\,pp$ raw gain, so adaptive poles are unnecessary. Yet all nine mean-model replays exactly match the frozen histories and predictions, and changing only their evidence to endpoint pooling loses $2.67\,pp$ accuracy and $1.90\,pp$ balanced accuracy, with all three participant differences negative. The BCI-derived endpoint law fails direct transport; the retraining contrast is not a portable rule.

5.5 ENCODER AND CORPUS TESTS DELIMIT THE SCOPE

The boundary advantage is stable across filtering choices but does not generalize cleanly across the tested alternative encoder and corpora (figure 5 and table 6). With the compact encoder, all three readouts remain near 38–40% and their participant-paired differences include zero. On the six-participant ASU subset, mean pooling reaches 73.28%, compared with 60.48% for endpoint pooling and 56.61% for BEM; all participants favor mean. FEIS provides no usable readout test because none of the systems demonstrably learns the task (full results in table 7). The two-boundary advantage is therefore not robustly observed outside the primary encoder–dataset pipeline in the settings tested.

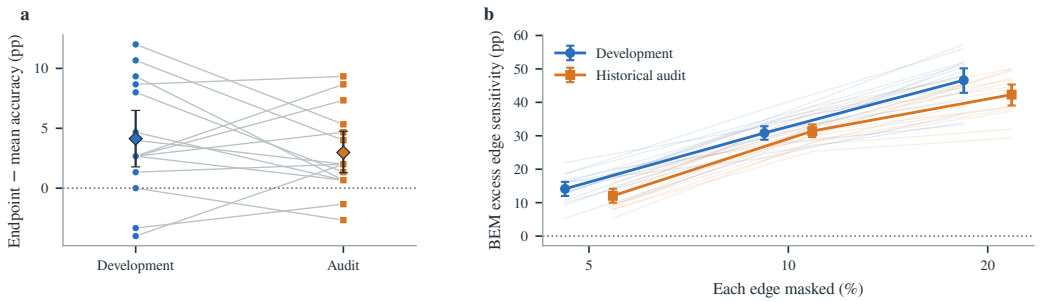


Figure 4: Participant-unit interventions on BCI2020. (a) The post-hoc endpoint-fifths rule is applied to identical frozen mean-model evidence; lines connect splits, and diamonds/bars are participant means and 95% bootstrap CIs ($n = 15$), conditional on the selected rule. (b) BEM's excess edge-versus-center sensitivity over mean increases with mask width; pale lines are participants and bold lines/intervals are means/95% CIs ($n = 15$).

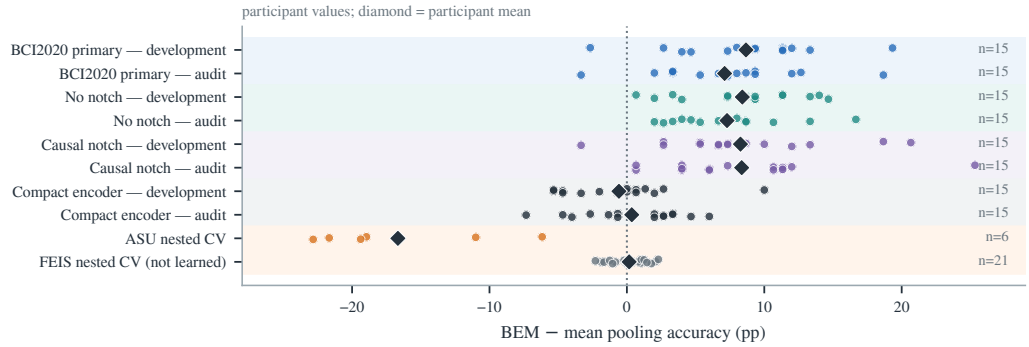


Figure 5: Scope across encoders, preprocessing choices, and corpora. Points are participant-unit BEM-minus-mean accuracy after seed averaging; diamonds are participant means. Primary, no-notch, and causal-notch effects remain positive; the compact encoder centers near zero, ASU reverses, and FEIS is gray because no method demonstrably learned that task.

6 Discussion

What is identified. An exact U-shaped kernel, weak one-sided readouts, a positive same-evidence endpoint swap, and greater edge- than center-mask sensitivity isolate an aggregation contribution in the primary BCI2020 pipeline. Fixed controls locate it in broad boundary shape, not adaptive poles. The contribution is the chain from exact kernel to distillation, intervention, and external stress test; bidirectional exponential smoothing itself is not new.

Evidentiary boundary. On on007591, raw accuracy agrees directionally, but balanced accuracy and the post-outcome endpoint transport disagree. Decision weights show classifier reliance, not neural localization; run 02 was historically inspected; and the endpoint and masking rules are post-hoc. The compact transplant is null, ASU reverses, FEIS is not learned, and on007591 has only three participants. We therefore claim a forensic BCI2020 finding plus a mixed stress test, not a physiological or transferable law; a larger balanced replication is needed.

7 Conclusion

An exact readout exposed a two-boundary decision pattern in BCI2020; distillation, same-evidence swaps, and perturbations isolate its aggregation component. A locally frozen external stress test passes its raw-accuracy gate but reverses under balanced accuracy, and endpoint transport is negative. Exact kernels can thus expose and falsify a temporal assumption without turning a pipeline-specific pattern into a general law.

REFERENCES

Mohamad Amin Bakhshali, Morteza Khademi, Abbas Ebrahimi-Moghadam, and Sahar Moghimi. EEG signal classification of imagined speech based on riemannian distance of correntropy spectral density. *Biomedical Signal Processing and Control*, 59:101899, 2020. doi: 10.1016/j.bspc.2020.101899.

Arun Balasubramanian, Kartik Pandey, Gautam Veer, and Debasis Samanta. Explainable artificial intelligence-based identification of the localized events in imagined speech electroencephalogram. *Computers & Electrical Engineering*, 127:110608, 2025. doi: 10.1016/j.compeleceng.2025.110608.

BCI Competition Committee. 2020 international BCI competition, 2020. URL https://osf.io/pq7vb/.

Jiqun Chu and Zuoquan Lin. Incorporating exponential smoothing into MLP: A simple but effective sequence model. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 326–337, 2024. doi: 10.18653/v1/2024.findings-naacl.23. URL https://aclanthology.org/2024.findings-naacl.23/.

Jonathan Clayton, Scott Wellington, Cassia Valentini-Botinhao, and Oliver Watts. Decoding imagined, heard, and spoken speech: Classification and regression of EEG using a 14-channel dry-contact mobile headset. In *Proc. Interspeech*, pp. 4886–4890, 2020. doi: 10.21437/Interspeech.2020-2745.

Safa Dörterler, Emrullah Şahin, and Durmuş Özdemir. A dynamic subject-invariant fragment mixing strategy to suppress subject variability in EEG imagined speech classification. *Neuroscience*, 608: 18–45, 2026. doi: 10.1016/j.neuroscience.2026.05.004.

Hatem T. M. Duhair, Masrullizam bin mat Ibrahim, Jamil Abedalrahim Jamil Alsayaydeh, Rex Bacarra, and Ahmed Hussein Ahmed. Progressive transfer learning unifies multi-corpus EEG for robust and scalable imagined-speech decoding. *Biomedical Signal Processing and Control*, 116: 109475, 2026. doi: 10.1016/j.bspc.2026.109475.

Albert Gu, Karan Goel, Ankit Gupta, and Christopher Ré. On the parameterization and initialization of diagonal state space models. In *Advances in Neural Information Processing Systems*, volume 35, 2022a.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *Proc. Int. Conf. Learning Representations (ICLR)*, 2022b.

Yiyu Gui, MingZhi Chen, Yuqi Su, Guibo Luo, and Yuchao Yang. EEGMamba: Bidirectional state space model with mixture of experts for EEG multi-task classification, 2024. URL https://arxiv.org/abs/2407.20254. Preprint.

Ankit Gupta, Albert Gu, and Jonathan Berant. Diagonal state spaces are as effective as structured state spaces. In *Advances in Neural Information Processing Systems*, volume 35, pp. 22982–22994, 2022. doi: 10.52202/068431-1670.

Felix Hausdorff. Summationsmethoden und momentfolgen. i. *Mathematische Zeitschrift*, 9(1–2): 74–109, 1921. doi: 10.1007/BF01378337.

Ji-Hoon Jeong, Jeong-Hyun Cho, Young-Eun Lee, Seo-Hyun Lee, Gi-Hwan Shin, Young-Seok Kweon, José del R. Millán, Klaus-Robert Müller, and Seong-Whan Lee. 2020 international brain–computer interface competition: A review. *Frontiers in Human Neuroscience*, 16:898300, 2022. doi: 10.3389/fnhum.2022.898300.

Owais Mujtaba Khanday, Mohamed Baha Ben Ticha, Sanae Belfrouh, Marc Ouellet, and Jose A. Gonzalez-Lopez. Do EEG foundation models transfer to speech? a benchmark on overt and imagined speech decoding, 2026. URL https://arxiv.org/abs/2607.27268. Preprint.

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

Dae-Hyeok Lee, Sung-Jin Kim, and Keon-Woo Lee. Decoding high-level imagined speech using attention-based deep neural networks. In *2022 10th International Winter Conference on Brain-Computer Interface (BCI)*, pp. 1–4, 2022. doi: 10.1109/BCI53720.2022.9734310.

Dong-Yeon Lee, Minji Lee, and Seong-Whan Lee. Decoding imagined speech based on deep metric learning for intuitive BCI communication. *IEEE Trans. Neural Syst. Rehabil. Eng.*, 29:1363–1374, 2021a. doi: 10.1109/TNSRE.2021.3096874.

Dongha Lee, Seonghyeon Lee, and Hwanjo Yu. Learnable dynamic temporal pooling for time series classification. In *Proc. AAAI Conf. Artificial Intelligence*, volume 35, pp. 8288–8296, 2021b. doi: 10.1609/aaai.v35i9.17008.

Seong-Whan Lee, Klaus-Robert Müller, and José del R. Millán. 2020 BCI competition, track 3, 2026. URL https://nemar.org/dataset/nm000113?v=v1.0.0.

Fu Li, Weibing Chao, Yang Li, Boxun Fu, Youshuo Ji, Hao Wu, and Guangming Shi. Decoding imagined speech from EEG signals using hybrid-scale spatial-temporal dilated convolution network. *Journal of Neural Engineering*, 18(4):0460c4, 2021. doi: 10.1088/1741-2552/ac13c0.

Chun Kai Ling, Fei Fang, and J. Zico Kolter. Deep archimedean copulas. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1535–1545, 2020. URL https://arxiv.org/abs/2012.03137.

Diego Lopez-Bernal, David Balderas, Pedro Ponce, and Arturo Molina. A state-of-the-art review of EEG-based imagined speech decoding. *Frontiers in Human Neuroscience*, 16:867281, 2022. doi: 10.3389/fnhum.2022.867281.

Xuezhe Ma, Chunting Zhou, Xiang Kong, Junxian He, Liangke Gui, Graham Neubig, Jonathan May, and Luke Zettlemoyer. MEGA: Moving average equipped gated attention. In *Proc. Int. Conf. Learning Representations (ICLR)*, 2023.

Chuong H. Nguyen, George K. Karavas, and Panagiotis Artemiadis. Inferring imagined speech using EEG signals: A new approach using riemannian manifold features. *Journal of Neural Engineering*, 15(1):016002, 2018. doi: 10.1088/1741-2552/aa8235.

Chuong H. Nguyen, George K. Karavas, and Panagiotis Artemiadis. Imagined speech EEG dataset—short and long words (Nguyen et al. 2017), 2026. URL https://www.nemar.org/dataset/nm000224.

Antonio Orvieto, Samuel L. Smith, Albert Gu, Anushan Fernando, Caglar Gulcehre, Razvan Pascanu, and Soham De. Resurrecting recurrent neural networks for long sequences. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 26670–26698, 2023. URL https://proceedings.mlr.press/v202/orvieto23a.html.

Jerrin Thomas Panachakel and Angarai Ganesan Ramakrishnan. Decoding covert speech from EEG—a comprehensive review. *Frontiers in Neuroscience*, 15:642251, 2021. doi: 10.3389/fnins.2021.642251.

S. W. Roberts. Control chart tests based on geometric moving averages. *Technometrics*, 1(3):239–250, 1959. doi: 10.1080/00401706.1959.10489860.

Emrullah Şahin and Durmuş Özdemir. ThinkSTra: A transformer-driven architecture for decoding imagined speech from EEG with spatial–temporal dynamics. *Medical & Biological Engineering & Computing*, 2025. doi: 10.1007/s11517-025-03478-9.

Motoshige Sato, Yasuo Kabe, Sensho Nobe, Akito Yoshida, Masakazu Inoue, Mayumi Shimizu, Kenichi Tomeoka, and Shuntaro Sasai. Delineating neural contributions to electroencephalogram-based speech decoding. *bioRxiv*, 2024. doi: 10.1101/2024.05.09.591996. URL https://doi.org/10.1101/2024.05.09.591996. Preprint.

Motoshige Sato, Yasuo Kabe, Sensho Nobe, Akito Yoshida, Masakazu Inoue, Mayumi Shimizu, Kenichi Tomeoka, and Shuntaro Sasai. Delineating neural contributions to EEG-based speech decoding. NEMAR dataset, version 1.0.0, 2026. URL https://nemar.org/dataset/on007591.

René L. Schilling, Renming Song, and Zoran Vondraček. *Bernstein Functions: Theory and Applications*. De Gruyter, 2nd edition, 2012. doi: 10.1515/9783110269338.

James A. Shohat and Jacob D. Tamarkin. *The Problem of Moments*, volume 1 of *Mathematical Surveys*. American Mathematical Society, 1943.

Sibo Song, Ngai-Man Cheung, Vijay Chandrasekhar, and Bappaditya Mandal. Deep adaptive temporal pooling for activity recognition. In *Proceedings of the 26th ACM International Conference on Multimedia*, pp. 1829–1837, 2018. doi: 10.1145/3240508.3240713.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. EEG Conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Trans. Neural Syst. Rehabil. Eng.*, 31: 710–719, 2023. doi: 10.1109/TNSRE.2022.3230250.

Marius Usher and James L. McClelland. The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, 108(3):550–592, 2001. doi: 10.1037/0033-295X.108.3.550.

Scott Wellington and Jonathan Clayton. Fourteen-channel EEG with imagined speech (FEIS) dataset, 2019. URL https://doi.org/10.5281/zenodo.3554128.

AI USE STATEMENT

Generative AI tools assisted research ideation, methodology and experiment design, implementation, result interpretation, literature discovery, and drafting and editing of the manuscript. The authors inspected the generated code, executed unit and end-to-end tests, verified numerical claims against saved predictions and result files, and checked bibliographic metadata against primary sources. The authors take responsibility for the final content, including all AI-assisted text, claims, code, and artifacts.

ETHICS STATEMENT

All experiments use previously released, de-identified EEG datasets and do not recruit new participants. EEG-based imagined-speech systems raise privacy and agency concerns if presented as decoding unconstrained thoughts. The tasks studied here classify a small set of prompted laboratory labels and do not support such an interpretation. Participant-level variability is reported rather than hidden by pooled accuracy. The models are research prototypes, not clinical or assistive devices.

REPRODUCIBILITY STATEMENT

Section 4 specifies split roles, preprocessing, optimization, seeds, metrics, and inference units. The appendix lists complete architectures and hyperparameters, and appendix A supplies full proofs. The anonymous supplement contains source code, environment metadata, dataset provenance and hashes, per-run configurations, BCI2020 checkpoints, predictions, participant and fold indices, and deterministic scripts that regenerate every table and figure.

A PROOFS

A.1 PROOF OF PROPOSITION 1

For every a and $d \in \{\mathrm{E}, \mathrm{S}\}$, $k_a^d(t) \geq 0$ and

$$\sum_{t=1}^{T} k_a^d(t) = \frac{\sum_{r=0}^{T-1} \beta_a^r}{Z_a} = 1. \tag{7}$$

The uniform component is also nonnegative and sums to one. Because $\boldsymbol{\pi}_c$ lies on the simplex, equation (5) gives

$$\sum_{t=1}^{T} w_c(t) = \sum_a \pi^{\mathrm{E}}_{a,c} + \sum_a \pi^{\mathrm{S}}_{a,c} + \pi^{\mathrm{U}}_c = 1, \tag{8}$$

$$w_c(t) \geq 0. \tag{9}$$

It follows that $z_c/\gamma = \sum_t w_c(t) e_{t,c}$ is a convex combination of the scalar frame evidences and therefore lies between their minimum and maximum.

A.2 Proof of proposition 2

Let $\Delta K(r) = K(r+1) - K(r)$. For one exponential atom,

$$\Delta^m \beta^r = \beta^r (\beta - 1)^m. \tag{10}$$

Linearity of the difference operator gives

$$(-1)^m \Delta^m K_c(r) = \sum_{a=1}^{A} \frac{\alpha_{a,c}}{Z_a} \beta_a^r (1 - \beta_a)^m \geq 0 \tag{11}$$

for every $m, r \geq 0$. Equivalently, define the finite positive measure $\nu_c = \sum_a (\alpha_{a,c}/Z_a)\delta_{\beta_a}$ on $[0, 1]$. Then $K_c(r) = \int_{[0,1]} \beta^r \, \mathrm{d}\nu_c(\beta)$, which is a Hausdorff moment representation. The complete kernel is not generally one branch plus its own reversal: in end-lag coordinates it is $K_c^{\mathrm{E}}(r) + K_c^{\mathrm{S}}(T - 1 - r) + \pi_c^{\mathrm{U}}/T$, with generally different coefficients in the two branches.

An exact in-family example also shows why complete monotonicity does not extend to that sum. Take $T = 3$, $A = 1$, $\beta_1 = 1/2$, and equal simplex masses $\pi^{\mathrm{E}} = \pi^{\mathrm{S}} = \pi^{\mathrm{U}} = 1/3$, which finite equal softmax logits realize exactly. In end-lag order, the effective kernel is

$$w = \frac{1}{63}(22, 19, 22). \tag{12}$$

Thus $-\Delta w(1) = -(w(2) - w(1)) = -1/21 < 0$. The proposition therefore applies to each directional branch in its own boundary-lag coordinate, not to the complete bidirectional kernel.

A.3 Proof of proposition 3

By proposition 1, w_c is nonnegative and sums to one. Hence

$$\left| \frac{z_c(e_c) - z_c(e'_c)}{\gamma} \right| = \left| \sum_{t=1}^{T} w_c(t)(e_{t,c} - e'_{t,c}) \right| \tag{13}$$

$$\leq \sum_{t=1}^{T} w_c(t) |e_{t,c} - e'_{t,c}| \tag{14}$$

$$\leq \|e_c - e'_c\|_\infty \sum_{t=1}^{T} w_c(t) = \|e_c - e'_c\|_\infty. \tag{15}$$

Multiplying both sides by γ yields the scaled bound.

A.4 Recurrent evaluation

The unnormalized end-anchored summary for pole β_a obeys

$$u_{a,c}(t) = \beta_a u_{a,c}(t-1) + e_{t,c}, \qquad u_{a,c}(0) = 0. \tag{16}$$

At $t = T$, division by Z_a gives the end-anchored summary in equation (2). The start-anchored summary can be accumulated in the same forward scan without storing the reversed path:

$$v_{a,c}(t) = v_{a,c}(t-1) + p_a(t) e_{t,c}, \qquad p_a(1) = 1, \quad p_a(t+1) = \beta_a p_a(t), \quad v_{a,c}(0) = 0. \tag{17}$$

Then $v_{a,c}(T) = \sum_{t=1}^{T} \beta_a^{t-1} e_{t,c}$, so division by Z_a gives the start-anchored summary. All A states and powers for each class can be updated in parallel, giving $O(TAC)$ time and $O(AC)$ auxiliary state for one trial, excluding evidence and encoder storage.

B Reproducibility Details

B.1 Primary encoder

The full encoder uses Conv1d$(64, 256, 11)$ with padding 5, batch normalization, GELU, dropout 0.25, and average pooling by 2; Conv1d$(256, 320, 1)$ with batch normalization, GELU, and dropout $0.25/3$; and Conv1d$(320, 64, 9)$ with dilation 2 and padding 8, batch normalization, GELU, dropout 0.25, and average pooling by 4. The evidence head is Linear$(64, 64)$–GELU–Linear$(64, 5)$. BEM initializes its 12 poles uniformly from 0.05 to 0.98 before the logit transform, initializes all measure logits to zero, and initializes the learned logit scale to $\exp(1.5)$. The fixed-pole control retains those 12 initial values throughout training. The fixed symmetric U control uses a pole of 0.95 in both directions with equal mass, while endpoint fifths assigns mass $1/2$ uniformly to each of the first and last $\lfloor T/5 \rfloor$ encoded steps. Direction-plus-uniform controls retain the corresponding directional atom bank and uniform atom but mask the opposite bank.

B.2 Matched readout definitions

All readouts receive a $T \times C$ class-evidence tensor and share a learned positive logit scale. Mean and maximum pool classwise over time. Attention applies a zero-initialized Linear(C, C) map at each step, takes a classwise softmax over time, and weights the original evidence. Free temporal pooling learns 100 zero-initialized logits per class, linearly interpolates them to T, and applies a temporal softmax. The evidence-GRU is a one-layer bidirectional GRU with input dimension C and 16 hidden units per direction; its two terminal states are concatenated and classified by Linear$(32, C)$. The evidence-SSM has C stable diagonal states. Sigmoid-parameterized decays are initialized uniformly from 0.1 to 0.99; a Linear(C, C)–tanh drive is summed with weights $(1 - \beta)\beta^r$ and classified by Linear(C, C). Attention, free weights, GRU, and SSM add 30, 500, 2,373, and 65 parameters, respectively, over mean pooling when $C = 5$. Every complete model is trained end-to-end under the common schedule; no baseline-specific hyperparameter sweep was performed, so conclusions refer to these implementations.

The primary notch uses SciPy's second-order IIR notch at 50 Hz with quality factor 35 and sampling rate 256 Hz. The primary protocol applies forward–backward filtering independently to each epoch; the causal control uses the same coefficients in one forward pass, and the no-notch control skips both operations. All three variants then apply common-average reference.

B.3 Compact encoder and external-corpus training

The compact encoder follows the EEGNet design: a 16-channel width-63 temporal convolution; a grouped spatial convolution over all input channels producing 32 channels; average pooling by 4; and a depthwise width-15 plus pointwise separable temporal block followed by average pooling by 8. Both blocks use batch normalization, ELU, and dropout 0.35. The evidence head is Linear$(32, 32)$–ELU–Linear$(32, C)$. The BCI2020 encoder transplant retains the primary readout's 12 atoms; the external-corpus experiments use eight.

The ASU scope experiment uses the six-participant NEMAR short–long subset, not the complete 15-participant release. The cache contains 1,120 whole *cooperate*-versus-*in* trials, 60 EEG channels, and 1,280 samples per trial. The FEIS cache contains 3,312 usable 1,280-sample trials from 21 participants and 16 phoneme classes; Participant 12 contributes 112 rather than 160 trials. Each participant is evaluated by five-fold stratified outer cross-validation. An inner 20% split selects the epoch by balanced accuracy, then the selected epoch is used to retrain on the complete outer training fold before one outer-test evaluation. Participant metrics are recomputed from the complete out-of-fold prediction vector; no overlapping windows are made.

Each external outer fold uses an inner stratified 20% validation split. Models train for at most 250 epochs with AdamW, learning rate 5×10^{-4}, cosine decay to 5×10^{-5}, weight decay 10^{-2}, batch size 32, gradient-norm clipping at 2, and class-weighted cross-entropy. Validation is performed every ten epochs; ties are resolved by accuracy and then the earlier epoch. Channel mean and standard deviation are estimated from the appropriate training partition only.

Table 1: Evidence ladder. Each stage answers a distinct question, preventing matched retraining, frozen-evidence diagnostics, and scope tests from being given the same interpretation.

Stage	Evidence	Supported inference
Discovery	Exact BEM kernels	Candidate two-boundary decision pattern in the archived primary models
Distillation and isolation	Endpoint control; identical-evidence swap	Boundary aggregation contributes on these learned evidence paths
Within-protocol checks	One-sided readouts; edge/center masks; filter retraining	Both directional branches are needed in the tested matched implementations, and the effect is not solely a zero-phase-notch artifact
External stress test	Locally frozen on007591 calibration-to-online evaluation	Raw accuracy is directionally positive, but balanced accuracy rejects a robust replication claim
Boundary tests	Compact encoder; ASU; FEIS	The current evidence does not support encoder- or corpus-level generality

For on007591, each participant has one covert calibration run with 100 events and one online run with 50 events. Every 6.25-s event contains five nonoverlapping 1.25-s repetitions. We drop the participant-wise union of run-level bad EEG channels, filter the continuous recording from 1–45 Hz with a zero-phase FIR filter, apply common-average reference, and resample from 256 to 128 Hz before extraction. Channel mean and standard deviation use only calibration-training repetitions. The same compact model and schedule train for exactly 250 epochs. The calibration split is grouped by original event and fixed with seed 20260926; selection is lexicographic by validation balanced accuracy, validation accuracy, and earliest epoch. Online data are evaluated only after loading the selected state. The run record stores the bad-channel union, channel order, calibration group IDs, every validation checkpoint, online repetition and event targets/predictions, and elapsed time. A separate validator recomputes accuracy, balanced accuracy, and macro-F1 from every stored prediction and verifies five repetitions per online event. The bad-channel union includes official signal-derived online-run QC metadata; this is transductive preprocessing, although it uses neither labels nor model outcomes. Selected checkpoints and repetition logits were not retained, so the stored event predictions can be metric-checked but their five-logit aggregation cannot be independently replayed. The downloaded v1.0.0 archive has an archived SHA-256 beginning `bcf3203c` and ending `c32903`.

The post-outcome auditability run retrains each of the nine mean configurations deterministically and requires exact agreement with the frozen selected epoch, 25-point validation history, repetition predictions, and event predictions. Only after all four checks pass does it archive the selected checkpoint and mean/endpoint/fixed-U repetition and event logits. A second post-outcome set retrains endpoint, fixed-U, and fixed-pole BEM controls using the unchanged split and optimization schedule. The supplement labels all 36 runs as post-hoc and includes a hash-complete run manifest.

B.4 Compute and software

The external-run records report Python 3.13.5, PyTorch 2.13.0+cu130, and CUDA 13.0; the primary checkpoints were produced on NVIDIA RTX 3090 GPUs. The executable entry points seed all stochastic sources, disable CuDNN benchmarking, and request deterministic CuDNN execution. A timestamped configuration freeze records SHA-256 hashes of the model, data, training, analysis, validation, and protocol sources before the external-corpus runs. A distinct local pre-EDF freeze and a documented pre-outcome metadata erratum cover on007591; the erratum corrects only the number of matching online files and a split-description typo. A provenance note corrects two impossible manually typed times and records the filesystem chronology; this local record is not a public registration or third-party timestamp. A separate freeze records the post-hoc identical-evidence diagnostic, all nine input checkpoint hashes, its script, cached data, and output.

C Additional Diagnostics

C.1 Evidence ladder

C.2 Complete BCI2020 control matrix

Table 2: Complete BCI2020 structural, preprocessing, and encoder control matrix. Accuracy is mean ± sample SD over the same three fixed seeds.

Configuration	Parameters	Development	Audit
Primary filter/encoder			
Mean pooling	452,870	77.07 ± 1.01	72.40 ± 0.40
Endpoint fifths	452,870	85.02 ± 1.08	80.44 ± 1.21
Fixed symmetric U	452,870	82.31 ± 0.95	78.00 ± 0.67
BEM	453,007	85.73 ± 1.57	79.51 ± 1.92
Structural ablations			
End-anchored only	453,007	67.91 ± 0.15	63.60 ± 0.83
Start-anchored only	453,007	68.93 ± 1.15	66.67 ± 1.57
End-anchored + uniform	453,007	71.47 ± 0.53	66.89 ± 0.63
Start-anchored + uniform	453,007	72.53 ± 1.70	70.67 ± 1.67
Fixed poles + learned measure	452,995	85.64 ± 0.96	79.87 ± 1.64
No notch			
No notch + mean	452,870	77.24 ± 1.46	73.02 ± 0.38
No notch + endpoint fifths	452,870	85.78 ± 1.00	80.93 ± 1.04
No notch + BEM	453,007	85.64 ± 1.44	80.31 ± 1.79
Causal notch			
Causal notch + mean	452,870	77.11 ± 0.60	72.36 ± 1.00
Causal notch + endpoint fifths	452,870	85.91 ± 0.66	80.84 ± 0.86
Causal notch + BEM	453,007	85.38 ± 1.59	80.71 ± 2.35
Compact encoder			
Compact encoder + mean	5,942	39.78 ± 1.27	37.96 ± 0.89
Compact encoder + endpoint fifths	5,942	39.82 ± 1.55	38.67 ± 0.69
Compact encoder + BEM	6,079	39.20 ± 1.97	38.31 ± 0.56

C.3 Matched readout comparison

Table 3: BCI2020 accuracy (%), mean ± sample SD across the three fixed seeds. All learned baselines use the same class-evidence interface and are trained end-to-end separately. Shading groups the two readouts that emphasize both boundaries; bold marks the best point estimate in each split.

Readout	Parameters	Development	Audit
Mean pooling	452,870	77.07 ± 1.01	72.40 ± 0.40
Maximum pooling	452,870	77.64 ± 1.39	70.98 ± 2.48
Class-evidence attention	452,900	78.00 ± 0.96	72.67 ± 1.15
Free classwise weights	453,370	78.04 ± 1.24	72.40 ± 0.74
Evidence-GRU	455,243	72.27 ± 2.68	67.47 ± 1.97
Evidence-SSM	452,935	60.36 ± 0.38	58.31 ± 2.14
Fixed symmetric U	452,870	82.31 ± 0.95	78.00 ± 0.67
Endpoint fifths	452,870	85.02 ± 1.08	$\mathbf{80.44 \pm 1.21}$
BEM	453,007	$\mathbf{85.73 \pm 1.57}$	79.51 ± 1.92

Table 4: Locally pre-outcome-frozen calibration-to-online stress test on NEMAR on007591. Entries average three seeds within participant. BA denotes balanced accuracy; the majority column is the online largest-class prior.

Participant	Mean	BEM	ΔAcc.	ΔBA	Majority
1	14.00	26.00	+12.00	−1.34	34.00
2	32.67	33.33	+0.67	−1.64	30.00
3	24.67	26.00	+1.33	−1.27	28.00
Mean	23.78	28.44	+4.67	−1.42	30.67

Table 5: Post-outcome on007591 mechanism controls. Each entry averages seeds within participant and then participants. Panel (a) retrains a matched model; panel (b) deterministically replays the nine frozen mean models and changes only the rule on the identical evidence tensor. None belongs to the locally frozen gate.

Analysis	Readout	Accuracy	Balanced accuracy	Macro-F1
(a) Matched retraining				
	Mean	23.78	25.60	18.97
	Learned BEM	28.44	24.18	21.23
	Endpoint fifths	28.00	23.77	21.87
	Fixed symmetric U	24.89	24.81	18.78
	BEM, fixed poles	27.78	25.41	20.31
(b) Mean-source identical evidence				
	Mean	23.78	25.60	18.97
	Endpoint fifths	21.11	23.70	17.49
	Fixed symmetric U	23.56	25.43	18.78

Table 6: Scope of the two-boundary effect. Entries are mean accuracy (%) over three seeds; BCI2020 cells show development/audit. Full uncertainty and macro-F1 are in tables 2 and 7.

Setting	Mean	Endpoint	BEM	Interpretation
BCI2020 primary	77.07/72.40	85.02/80.44	85.73/79.51	two-boundary gain
BCI2020 no notch	77.24/73.02	85.78/80.93	85.64/80.31	gain retained
BCI2020 causal notch	77.11/72.36	85.91/80.84	85.38/80.71	gain retained
BCI2020 compact encoder	39.78/37.96	39.82/38.67	39.20/38.31	no gain detected
on007591 online	23.78	—	28.44	accuracy ↑, balanced accuracy ↓
ASU	73.28	60.48	56.61	mean favored

C.4 Pre-outcome-frozen online-acquisition result

C.5 Scope summary

C.6 External-corpus replication matrix

Table 7 reports the complete participant-level summary for the two external corpora discussed in the main text.

C.7 Identical-evidence readout swaps

Table 8 reports every source-checkpoint family from the post-hoc diagnostic. The endpoint advantage over mean remains positive when the encoder and frame-level evidence are held literally fixed, including on evidence learned by the mean-pooling systems. Because each evidence path was optimized with its source readout, these counterfactual swaps diagnose the aggregation operation but are not a replacement for matched retraining.

Table 7: Nested within-participant cross-validation on external corpora. Accuracy reports the participant mean and sample SD across three seeds; Macro-F1 is computed from pooled out-of-fold predictions within each seed and then averaged across seeds. Outer folds never select epochs.

Corpus	Readout	Participants	Accuracy	Macro-F1
ASU	Mean	6	73.28 ± 0.75	72.97
ASU	Endpoint fifths	6	60.48 ± 1.24	60.31
ASU	BEM	6	56.61 ± 0.88	56.47
FEIS	Mean	21	5.56 ± 0.43	5.45
FEIS	Endpoint fifths	21	5.77 ± 0.38	5.70
FEIS	BEM	21	5.73 ± 0.18	5.67

Table 8: Post-hoc readout swaps on identical frozen evidence tensors. Each row averages the three archived source checkpoints after aggregating seeds within participant. Values are percentage-point differences with 95% participant-bootstrap intervals; no readout is refit. The intervals condition on the already chosen endpoint rule and exclude rule-selection uncertainty.

Evidence source	Contrast	Development	Audit	Positive n (dev.; audit)
BEM-trained	Endpoint − mean	$6.04\ [3.96, 8.04]$	$4.98\ [2.93, 7.20]$	14/15; 13/15
Mean-trained	Endpoint − mean	$4.13\ [1.78, 6.49]$	$2.98\ [1.29, 4.76]$	13/15; 13/15
Endpoint-trained	Endpoint − mean	$5.60\ [3.95, 7.20]$	$5.51\ [4.31, 6.76]$	14/15; 15/15
BEM-trained	BEM − endpoint	$1.29\ [0.22, 2.53]$	$-0.40\ [-1.51, 0.67]$	9/15; 6/15

C.8 Focal multiplicity family

Table 9 enumerates the eight BEM comparisons included in the Holm correction. Bootstrap intervals in the main text and figure 6 remain the primary effect-size summaries; paired t and seedwise exact McNemar results are retained in the released machine-readable analysis artifact as sensitivity checks.

Table 9: Participant-unit Wilcoxon inference for the focal eight-comparison family. Δ is BEM minus the listed readout in percentage points; $p_{\mathrm{raw}}/p_{\mathrm{Holm}}$ gives the raw and within-split adjusted two-sided values.

Readout	Δ dev.	$p_{\mathrm{raw}}/p_{\mathrm{Holm}}$	Δ audit	$p_{\mathrm{raw}}/p_{\mathrm{Holm}}$
Mean	8.67	0.000803/0.00520	7.11	0.001782/0.00713
Maximum	8.09	0.001225/0.00520	8.53	0.000653/0.00457
Attention	7.73	0.000801/0.00520	6.84	0.002140/0.00713
Free classwise weights	7.69	0.000801/0.00520	7.11	0.001326/0.00663
Endpoint fifths	0.71	0.729699/0.72970	−0.93	0.277088/0.27709
Evidence-GRU	13.47	0.000061/0.00049	12.04	0.000061/0.00049
Evidence-SSM	25.38	0.000650/0.00520	21.20	0.000653/0.00457
Fixed symmetric U	3.42	0.003071/0.00614	1.51	0.125798/0.25160

C.9 Participant-level effects

C.10 Exact learned-kernel decomposition

Figure 2 expands all 15 class-specific kernels from the three archived primary BEM checkpoints (three seeds and five classes). The learned poles span short through trial-scale half-lives. Simplex mass is nearly balanced between the end- and start-anchored banks: averaged over seeds, the respective classwise ranges are 45.23–48.47% and 47.13–50.07%. The uniform component receives only 4.40–4.82%. Despite this directional balance, the class kernels are not exactly symmetric because their two simplex measures are learned independently.

The complete kernel visualization is promoted to figure 2 in the main paper.

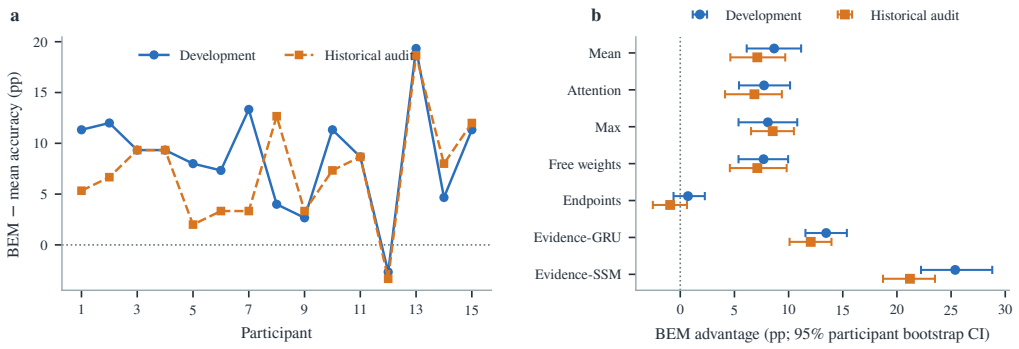


Figure 6: Participant-unit BCI2020 comparisons after averaging the three seeds. (a) BEM minus mean pooling for every participant. (b) Mean differences and 95% participant-bootstrap intervals against selected matched readouts. The endpoint interval crosses zero, motivating the structural rather than method-superiority interpretation.

C.11 Additional frozen-checkpoint perturbations

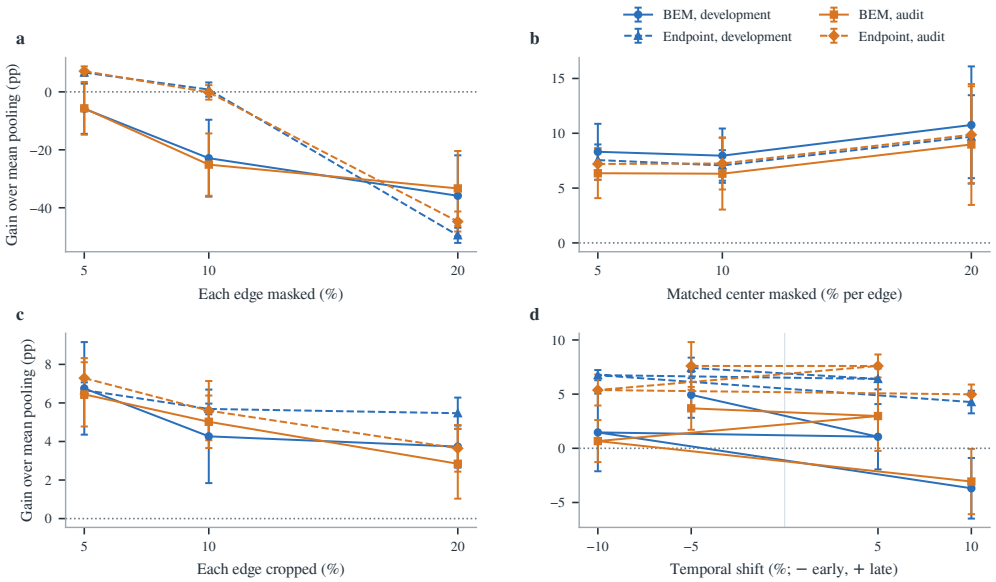


Figure 7: Complete frozen-checkpoint perturbation matrix. The main paper focuses on the hypothesis-directed edge-versus-center comparison; panels (c)–(d) retain the crop and temporal-shift sensitivity analyses.

C.12 Readout-only computational cost

The primary BEM model adds 137 trainable parameters over mean pooling. For a single-thread CPU microbenchmark on an AMD EPYC 7302, with batch size 256, 99 evidence steps, five classes, and 12 atoms, the median operator-only times are 0.0522 ms per batch for mean pooling, 0.1769 ms for endpoint fifths, and 0.6164 ms for BEM. These correspond to 0.204, 0.691, and 2.408 microseconds per trial. Interquartile ranges are 0.0044, 0.0069, and 0.0300 ms per batch, respectively. The benchmark excludes the encoder and host–device transfer and therefore characterizes only the evidence-to-logit operator; it is not an end-to-end throughput claim.