WHEN DOES EEG DECODING SURVIVE ELECTROMAGNETIC INTERFERENCE?

Anonymous authors
Paper under double-blind review

ABSTRACT

EEG decoders are trained in electromagnetically clean laboratories but deployed in hospitals, cockpits, and wearable settings. Structured electromagnetic interference collides with neural frequency bands, driving decoders to chance. Removing those frequencies can also erase activity that the task needs. We ask when decoding survives this shift. We introduce the Interference–Neural Confusion benchmark with four physically motivated interference families and a compact sensing–gating–calibration front end. It locates occupied frequencies in a fixed short-time Fourier representation, suppresses those bins, and withholds predictions when sensing is unreliable. For a fixed mask, the output is unchanged by interference whose transform support the mask covers; an operator-norm bound quantifies residual leakage. In subject-disjoint experiments on STEW, narrowband interference at -10 dB drops mean accuracy from 0.858 to 0.504, while gating restores 0.747, with gains on every held-out subject. For left- versus right-hand imagery in BCI Competition IV 2a, accuracy remains 0.556; band ablation points to the task information lost by gating. Prediction-set coverage on STEW rises from 0.478 to 0.854 after gating (clean: 0.915), but broadband sweeps remain largely undetected. These results connect recovery to detectable interference and the preservation of task-relevant rhythms, identifying when spectral gating can support reliable EEG decoding.

1 INTRODUCTION

EEG systems are commonly evaluated with clean recordings, although practical deployments place electrodes near power supplies, radios, motors, and other electronic equipment. The resulting electromagnetic interference (EMI) can be structured: mains harmonics can form narrow lines, communication signals can appear as modulated bursts, switching events can form pulses, and motor drives can produce broadband sweeps. This structure distinguishes EMI from the independent noise models used by many robustness benchmarks, but it also creates a difficult failure mode: an interference component can occupy the same band as task-relevant neural activity.

We study EMI as an explicit distribution-shift axis for EEG decoding. To our knowledge, this is the first controlled, multi-family study of complex electromagnetic interference in EEG decoding. We call the benchmark Interference–Neural Confusion (INC). Our experiments point to a concrete spectral criterion: recovery needs a visible interference signature and a frequency band that the task can spare. Existing distribution-shift benchmarks (Koh et al., 2021; Gulrajani & Lopez-Paz, 2021) and recent EEG representation and domain-generalization methods (Yao et al., 2026; Zhong et al., 2025; Wang et al., 2024a; Song et al., 2025; Duan et al., 2026) study subject, site, or dataset variation; INC adds electromagnetic structure as an independent shift.

We test this hypothesis with Sense–Gate–Calibrate (SGC). The sensor estimates spectral occupancy and an in-band signal-to-interference ratio (SIR), then flags windows for which the estimate is unreliable. A fixed short-time Fourier transform (STFT) gate removes the occupied bins. Linearity and data-independent overlap-add normalization make the resulting representation **exactly invariant** to additive interference whose STFT support the mask covers. A randomized adaptive prediction-set (APS) layer (Romano et al., 2020) supplies prediction sets on clean calibration data; unreliable windows are withheld rather than assigned an overconfident label. SGC is not a learned denoiser. It

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.

changes the representation while leaving the decoder intact, allowing the experiments to separate invariance from the loss of task information.

A cross-family split makes this criterion testable: learned interference augmentation sees only F1, while the fixed sensor and gate face all four families. On STEW, plug-in gating rescues a frozen model under detectable narrowband EMI. On the left/right-hand imagery subset of BCI Competition IV 2a (BCIC-LR), band ablation links the limited recovery to the information removed by the gate. Broadband sweeps expose the sensor's blind spot. Our contributions are:

- the Interference–Neural Confusion benchmark, a controlled testbed for complex electromagnetic interference in EEG decoding with four physically motivated families, four SIR levels, and subject-disjoint evaluation on two public tasks;
- a representation-level intervention for learned EEG decoders that combines spectral sensing, gating, and gate-consistent training with an exact support-conditioned invariance result, a leakage bound, and abstention-aware calibration;
- a paired positive/negative evaluation of the observability hypothesis: a learned decoder recovers from detectable interference on STEW, while gating removes task-relevant information on BCIC-LR.

2 Related work

Cross-subject EEG decoding already poses a distribution-shift problem. Compact convolutional decoders such as EEGNet (Lawhern et al., 2018) and broader representation studies (Schirrmeister et al., 2017) provide strong clean baselines, while STEW (Lim et al., 2018) and BCI Competition IV 2a (Brunner et al., 2008) support widely used workload and motor-imagery tasks. Recent EEG domain-generalization methods use multi-source alignment, reconstruction, or latent-distribution exploration (Zhong et al., 2025; Wang et al., 2024a; Song et al., 2025); new neurophysiology benchmarks emphasize harmonized metadata and evaluation across recording sites (Duan et al., 2026). We study these decoding problems under an additional, physically specified environment shift.

Robust representation learning often learns invariance from training perturbations or source domains. Fishr (Rame et al., 2022) matches gradient variability across domains, PGD training (Madry et al., 2018) targets norm-bounded perturbations, and augmentation methods such as AugMix (Hendrycks et al., 2020) broaden the input distribution. Test-time adaptation, including TENT (Wang et al., 2021), updates a model with unlabeled target samples; surveys examine its assumptions under distribution shift (Wang et al., 2024b). Our baselines include interference mixing, band-stop augmentation, and TENT. The cross-family split tests whether a fixed spectral representation transfers to interference families absent from training.

Our guarantee complements general certified robustness. Randomized smoothing (Cohen et al., 2019) bounds probabilistic or norm-bounded perturbations, whereas our exact statement conditions on a physically interpretable STFT support. Split conformal prediction (Vovk et al., 2005; Romano et al., 2020; Zhang & Zhou, 2025) lets us report coverage alongside prediction-set size. Recent ICLR studies show that calibration depends on the score beyond softmax confidence (Akhavan Attar et al., 2026) and should be reported separately from accuracy (Adler et al., 2026); we follow that practice.

3 The INC benchmark

3.1 Interference model

The simulator generates unit-RMS waveforms and adds them to each channel with independent common-mode gains sampled from $U(0.5, 1.5)$. For a window x and unit-RMS waveform i, we form

$$y_c(t) = x_c(t) + g_c\, a\, i(t), \qquad a = \sqrt{P_x/P_i}\, 10^{-\mathrm{SIR}/20}, \quad g_c \sim U(0.5, 1.5), \tag{1}$$

where P_x and P_i are powers over the demeaned window and SIR is the nominal signal-to-interference ratio in decibels. We normalize each waveform to unit RMS before scaling. Random channel gains slightly change the realized average-channel SIR, which we recompute for every sample. This benchmark SIR differs from the sensor's separate 1.5–45 Hz excess-power estimate. The conditions are:

- F1: a slowly drifting continuous-wave component in 8–25 Hz, with mains-frequency components and harmonics added when they fall below the Nyquist guard band;
- F2: quadrature phase-shift keying or binary frequency-shift keying (QPSK/2FSK) bursts with windowed-sinc pulse shaping and random carriers in 8–30 Hz;
- F3: radar-like carrier bursts with nominal 0.05–0.3 s duration, constrained to contain at least four carrier cycles, and separated by random gaps;
- F4: a 5 Hz to $0.4f_s$ chirp with in-band $1/f^{0.5}$ noise.

We evaluate all families at SIR $\{10, 0, -10, -20\}$ dB, with a clean F0 control.

3.2 Tasks and evaluation

STEW contains recordings from 48 subjects with 14 channels sampled at 128 Hz. We classify low versus high workload using two-second windows with 50% overlap. BCI Competition IV 2a contains four motor-imagery classes—left hand, right hand, feet, and tongue—recorded from 9 subjects with 22 EEG channels at 250 Hz. To keep the primary theoretical comparison in the same binary label space as STEW, we first evaluate the left- versus right-hand subtask, denoted BCIC-LR, with 144 trials per subject. We additionally evaluate the complete four-class task as a multiclass stress test (Section 6.3 and Appendix F). We band-pass filter at 4–38 Hz, resample to 128 Hz, and analyze 3.5-second windows starting approximately 0.5 seconds after the imagery cue (Appendix G). The resulting inputs have shape $C \times T$, with $T = 256$ for STEW and $T = 448$ for BCIC.

For STEW, we sample ten of the 48 subjects with seed 42 and evaluate ten leave-one-subject-out folds; each fold trains on the other 47 subjects. BCIC-LR uses all nine leave-one-subject-out folds. Each held-out subject contributes to every interference condition. Learned methods that use interference examples see F1 only; the fixed sensor and gate require no interference training and also face F2–F4. We report fold means and standard deviations, paired differences across subjects, accuracy, coverage, prediction-set size, abstention, and sensing statistics. Overlapping windows are not independent replicates, and folds share most training subjects; intervals and signed-rank tests therefore serve as descriptive summaries across subjects.

Unless stated otherwise, the sensor uses $\alpha_{\mathrm{dB}} = 5$, a median noise floor ($q_{\mathrm{floor}} = 0.5$), a four-bin merge gap, 0.4 Hz minimum component width, and 1.5–45 Hz fitting limits. Runtime sensing uses $n_{\mathrm{seg}} = T/2$; the gate uses a Hann window with $n_{\mathrm{fft}} = T/2$, hop $T/8$, and two-bin dilation. Training uses Adam (learning rate 10^{-3}, batch size 64, 30 epochs); mixing and gate-consistent training inject F1 with probability 0.5 at SIR $\{10, 0, -10\}$ dB. We fix these settings without using test labels.

4 Sense–Gate–Calibrate

SGC sits between a trained decoder and each incoming EEG window. The sensor first asks which frequencies are occupied, the gate removes only those bins, and calibration turns the remaining uncertainty into a prediction set or an abstention. The first two steps use a fixed spectral representation; the last uses a clean calibration set. This ordering keeps representation invariance separate from the question of how much task information the gate removes.

4.1 Sensing

For each window, we estimate a Welch power spectrum, fit a two-pass log–log background trend using bins with nonzero power, and apply a constant-false-alarm-rate (CFAR) threshold (Rohling, 1983). We dilate occupied components before gating to cover spectral sidelobes. The sensor also estimates in-band SIR from excess power, classifies each component from harmonic spacing and bandwidth, and sets $r(x) = 1$ when that estimate is reliable; broad or highly occupied spectra set $r(x) = 0$. We evaluate detection, false occupancy, missed interference, and SIR error without using task labels. A missed broadband component can still yield $r(x) = 1$ and bypass abstention.

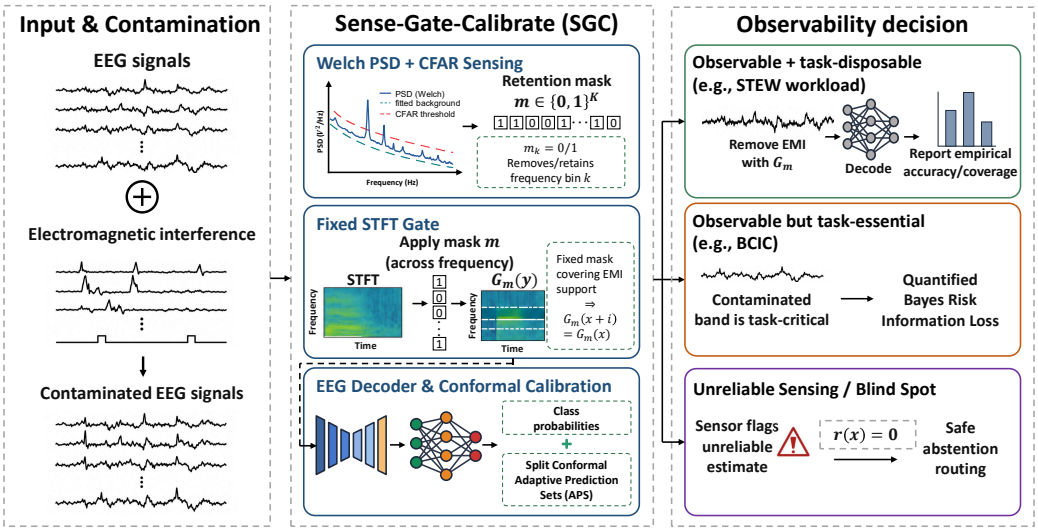


Figure 1: Sense–Gate–Calibrate (SGC) and its three operating cases. The sensor identifies occupied frequencies with a Welch spectrum and CFAR threshold, and the fixed short-time Fourier gate removes the detected bins. The right-hand panels distinguish recovery when the retained frequencies support the task, information loss when the removed band is task-relevant, and abstention when the sensor flags an unreliable estimate. The BCIC example refers to left/right-hand imagery. Missed interference can bypass this flag; the abstention branch applies only to detected unreliability. Coverage is evaluated across labeled test windows.

4.2 Gating

Let STFT be a center-free STFT with Hann window w and hop H. A retention mask $m \in [0,1]^F$ removes bin k when $m_k = 0$. The gate applies this mask before overlap-add synthesis:

$$G_m(x) = S(m \odot \mathrm{STFT}(x)). \tag{2}$$

The synthesis operator S uses the normalization $n_t = \sum_k w^2[t - kH]$. We use fixed filters, reflect padding, and interior cropping for short windows. The operator is linear in x and differentiable without learned parameters. The sensor supplies m, although a conservative mask can also come from a known interference spectrum. The theory below assumes that m stays fixed after sensing; a signal-dependent mask can change between x and $x + i$. Real-window experiments measure that sensing step separately.

4.3 Calibration and Abstention

We fit a split conformal APS threshold on clean calibration windows. For class probabilities $p(x)$, the randomized APS score is

$$s(x,y) = \sum_{j:p_j(x)>p_y(x)} p_j(x) + U\,p_y(x), \qquad U \sim \mathrm{Unif}(0,1), \tag{3}$$

and the prediction set contains classes with score at most the calibration quantile. At test time, we apply the same threshold to the decoder output for $G_m(x)$. When $r(x) = 0$, the system abstains instead of treating a low-confidence set as a valid prediction. The formal coverage statement belongs to the exchangeable calibration distribution; under EMI, we measure how much coverage returns.

5 Theory

5.1 Exact Invariance and Leakage

Theorem 1 (Exact invariance). *For a fixed retention mask m, if $m_k = 0$ wherever an additive interference i has a nonzero STFT coefficient, then $G_m(x + i) = G_m(x)$ in exact arithmetic.*

Proof. Linearity gives $\mathrm{STFT}(x + i) = \mathrm{STFT}(x) + \mathrm{STFT}(i)$. The support assumption implies $m \odot \mathrm{STFT}(i) = 0$, and data-independent overlap-add normalization preserves linearity. Therefore $G_m(x + i) - G_m(x) = S(0) = 0$. □

Corollary 1 (Hann-bin support)**.** *Under an ideal periodic Hann window, exact bin alignment, interior frames, and exact arithmetic, a sinusoid has support only at its central bin and its two immediate neighbors. A three-bin (or wider) removal mask therefore removes it exactly. The leakage bound covers the finite, reflect-padded implementation.*

Theorem 2 (Leakage bound)**.** *Let $C_S = \|S\|_{F\to 2}$ be the operator norm of the finite-dimensional overlap-add synthesis map induced by the chosen window, hop, normalization, and FFT convention. For any fixed mask, signal, and interference,*

$$\|G_m(x + i) - G_m(x)\|_2 \le C_S \|m \odot \mathrm{STFT}(i)\|_F. \tag{4}$$

For an L-Lipschitz encoder, the same bound multiplied by L controls representation drift.

Proof. The difference equals $S(m \odot \mathrm{STFT}(i))$. The definition of the synthesis operator norm bounds this quantity by C_S times the coefficient Frobenius norm. Lipschitz continuity gives the encoder statement. We leave C_S explicit because FFT normalization and boundary handling determine its value. □

5.2 Task information in the removed band

Write $X = X_B + X_{B^\complement}$ for a frequency-band projection. If the mask removes B and I is supported on B, Theorem 1 leaves the decoder with a deterministic transform of $X_{B^\complement}$. Let $R^*(Y \mid Z) = 1 - \mathbb{E}[\max_y p(Y = y \mid Z)]$ denote the Bayes error when only Z is observed. Thus, for any decoder f,

$$\Pr[f(G_m(X + I)) \neq Y] \ge R^*(Y \mid X_{B^\complement}). \tag{5}$$

The gap $R^*(Y \mid X_{B^\complement}) - R^*(Y \mid X)$ is nonnegative by data processing. In neural terms, once the decoder loses X_B, it must classify from the rhythms that remain; if B carried class-specific activity, some errors are unavoidable. The gap is zero when $Y \perp\!\!\!\perp X_B \mid X_{B^\complement}$, and it is positive only when the removed band lowers the Bayes risk. We estimate this task dependence with clean band ablations.

For ideal bin-centered components and an information-preserving retained transform, a Bayes-optimal decoder can attain the lower bound. Whether a real sensor finds the relevant support is an empirical question addressed by the sensing experiments.

6 Experiments

6.1 Models and baselines

EEGNet-lite (about 3k parameters) is the default decoder. We compare empirical risk minimization (ERM), F1 mixing, blind band-stop augmentation, test-time entropy minimization (TENT), SimpleCNN, frozen ERM with a sensing gate, and a model trained on gated inputs (the complete SGC procedure). All methods use the same subject splits and clean labels. "ERM + gate" changes only the test-time representation; SGC training also exposes the decoder to gated inputs and can abstain when $r(x) = 0$. Tables report fold means and standard deviations.

6.2 Sensing characterization

Across STEW and BCIC-LR windows with simulated interference, F1–F3 detection ranges from 78–100% across SIR levels (the lower end occurs for weak F2 bursts). On STEW, the in-band SIR estimator is less stable than detection for these families (2–22 dB mean absolute error depending on family and window length), so we use it as a reliability cue. The BCIC-LR sensing analysis measures detection and reliability. F4 detection is 0–20% on STEW and 0–56% on BCIC-LR, while $r(x) = 1$ remains common; this broadband blind spot therefore bypasses abstention. A controlled 40 Hz tone on synthetic band-limited signals yields a SIR mean absolute error of 0.38 dB. The difference highlights the difficulty of estimating interference power against a neural background.

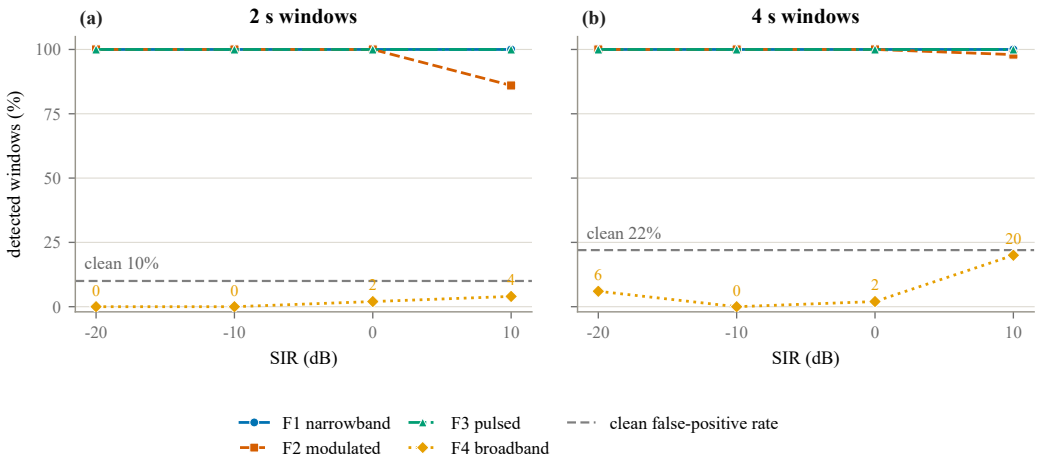


Figure 2: Detection rate on 50 fixed STEW windows for each condition. Each panel shows the four SIR levels; line and marker styles identify interference families, and the gray dashed line is the clean false-positive rate. F1–F3 remain detectable, whereas F4 detection ranges from 0–20%.

6.3 Accuracy under EMI

On STEW, narrowband F1 at -10 dB collapses ERM from 0.858 to 0.504. Applying the sensor and gate to the frozen model recovers $\mathbf{0.747}$ (paired improvement $+\mathbf{0.243}$, all 10 folds positive; paired Wilcoxon signed-rank $p = 0.002$; fold-level 95% t interval $[0.170, 0.316]$). Gate-consistent training followed by the same gate reaches $\mathbf{0.799}$, close to its clean reference of 0.820. In the additional baseline comparison, blind band-stop augmentation reaches 0.828 clean accuracy but only 0.505 under F1, suggesting that unconditioned frequency removal alone is insufficient. Mixing reaches 0.768 under F1 and 0.658 under F4, while TENT provides a smaller gain (Appendix C).

BCIC-LR shows a different outcome. ERM is 0.677 clean and 0.556 under F1 at -10 dB; **the gate does not improve it** (paired Wilcoxon $p = 0.82$; fold-level 95% t interval for the paired difference $[-0.032, 0.032]$). Removing 8–30 Hz from clean inputs reduces accuracy from 0.671 ± 0.151 to 0.505 ± 0.057, whereas removing 4–8 or 30–38 Hz leaves it at 0.667 and 0.669. The contaminated band carries the decision-relevant rhythm. At SIR -20, ungated ERM is near chance on both datasets, but gated narrowband conditions remain above chance on STEW, where the component is detectable; broadband F4 remains near chance. Appendix B gives the complete matrices and the conditions that call for abstention.

Full four-class stress test. We next remove the binary restriction and evaluate all four BCIC classes with a four-output EEGNet under the same nine subject-disjoint folds. Across three seeds, artifact-free ERM improves from 0.308 to 0.326 on F1 at -10 dB (gate minus raw $+0.018$; 18 of 27 paired folds improve). The gate-consistent model (SGC) improves from 0.276 to 0.311 ($+0.035$; 20/27 folds), while F4 remains unchanged (ERM 0.294 raw versus 0.293 gated). The complete four-class task therefore supports the same boundary seen in the binary comparison: a visible narrowband component can be partly removed, but a broadband blind spot and removal of task-relevant content set a hard limit.

6.4 Cross-family transfer

The fixed front end requires no interference training. In an independent full-family STEW stress test, applying it to a frozen ERM model at -10 dB gives 0.753 on F1, 0.666 on previously unseen F2, 0.604 on previously unseen F3, and 0.500 on F4, compared with ungated accuracies of 0.509, 0.503, 0.528, and 0.500. The gain fades from narrowband to pulsed interference and disappears for broadband F4, tracking the detection pattern. On this benchmark, F1–F3 are often visible enough for the gate to help, while F4 usually calls for an independent occupancy check or abstention. Together, these results tie recovery to the sensor's ability to see the interference.

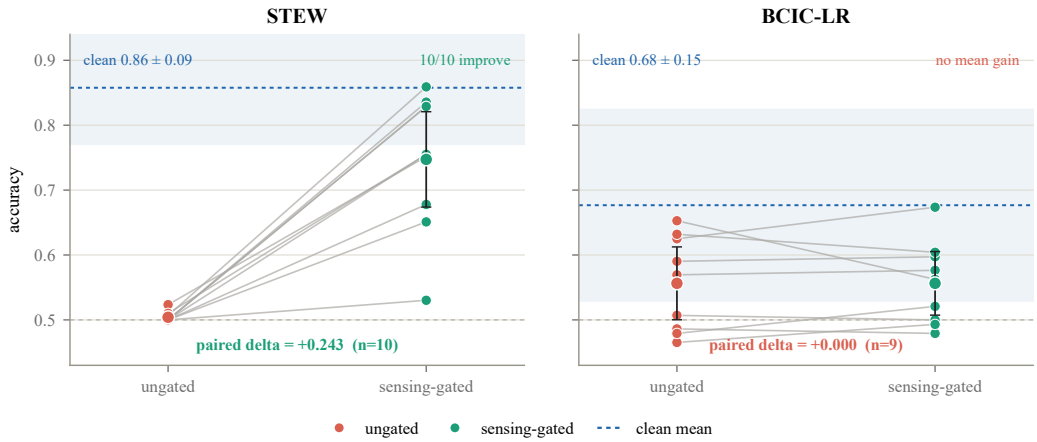


Figure 3: Fold-level paired accuracy for a frozen ERM decoder under F1 at SIR -10 dB. Lines connect each held-out subject before and after the gate; large markers show means with 95% t intervals, while the shaded band and dashed line show the clean reference mean $\pm$ subject standard deviation. The gate raises STEW accuracy by $+0.243$ (all ten folds improve) but has zero mean gain on BCIC-LR, exposing the task-relevance boundary.

Table 1: Accuracy (mean $\pm$ fold standard deviation) on STEW and the left/right-hand subset of BCI Competition IV 2a (BCIC-LR). Bold marks the largest mean within each row. The SGC column uses the gate-consistent model on contaminated rows and its ungated evaluation on clean inputs. The STEW clean-gate entry comes from a separate training run on the same folds; BCIC-LR uses the main models (Appendix A).

Dataset	condition	ERM raw	ERM + gate	SGC train + gate	ERM-mix raw
STEW	clean	$\mathbf{0.858 \pm 0.087}$	0.788 ± 0.086	0.820 ± 0.088	0.832 ± 0.108
	F1, -10 dB	0.504 ± 0.008	0.747 ± 0.103	$\mathbf{0.799 \pm 0.124}$	0.768 ± 0.140
	F4, -10 dB	0.500 ± 0.000	0.500 ± 0.000	0.516 ± 0.028	$\mathbf{0.658 \pm 0.108}$
BCIC-LR	clean	$\mathbf{0.677 \pm 0.147}$	0.559 ± 0.044	0.566 ± 0.086	0.664 ± 0.131
	F1, -10 dB	$\mathbf{0.556 \pm 0.073}$	$\mathbf{0.556 \pm 0.064}$	0.529 ± 0.036	0.549 ± 0.039
	F4, -10 dB	0.550 ± 0.064	$\mathbf{0.573 \pm 0.074}$	0.525 ± 0.036	0.563 ± 0.074

6.5 Conformal coverage under interference

For the main conformal analysis, we randomly divide each target subject's windows into equal calibration and evaluation halves. Only the calibration labels determine the threshold. On STEW, coverage is 0.915 ± 0.021 on clean windows, 0.478 ± 0.039 for F1 -10 dB without gating, and 0.854 ± 0.049 after gating; mean set size changes from 1.25 to 0.94 to 1.28. BCIC-LR gives coverage 0.892 ± 0.047, 0.853 ± 0.091, and 0.843 ± 0.064, with set sizes 1.55, 1.68, and 1.63, respectively. The gate restores most of the lost STEW coverage with larger prediction sets. Coverage changes little on BCIC-LR, consistent with its limited accuracy recovery. Appendix E examines calibration using source subjects only.

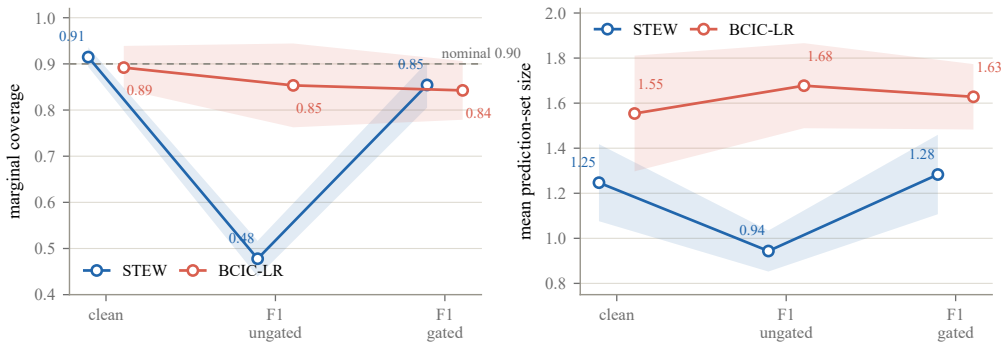


Figure 4: Conformal coverage and prediction-set size under F1 interference. Left: marginal coverage; right: mean set size. Lines show fold means and shaded bands show $\pm$ one fold standard deviation; colors distinguish STEW and BCIC-LR. The dashed line is the nominal $1 - \alpha = 0.9$ target. The main calibration split uses target-subject windows (see text); the plotted recovery is an empirical shift diagnostic.

7 Discussion and Limitations

The invariance result assumes that the mask covers the interference support. For missed components, Theorem 2 relates output drift to the residual spectral energy. Narrowband lines are detectable in our experiments, while broadband energy can blend into the whitening trend and appear reliable to the sensor. The threshold sweep shows stable F1 detection over the tested settings, with variable clean false positives and F4 detection. Masking a neural alpha or motor-imagery component can also lower clean accuracy. Clean gating, mask dilation, and the BCIC-LR band ablation quantify this cost.

Our simulator uses additive common-mode coupling and idealized waveform families. Hardware notch filters, electrode impedance, muscle artifacts, and non-additive coupling can change the observed spectrum. The hand-imagery subset was selected during preliminary experiments; the BCIC-LR results concern that binary task, while the full four-class stress test in Section 6.3 confirms the same narrowband-recovery and broadband-limit pattern. The study examines controlled cross-subject stress. We use one stochastic training run per fold for the main matrix and provide a full 48-subject STEW replication (0.504 to 0.718 when pooled over three training runs). The source-subject calibration analysis examines sensitivity to calibration choice. Measured EMI, spatially varying coupling waveforms, and an external sensing check remain important next steps.

Conclusion. In the settings studied here, EEG decoding survives electromagnetic interference when the interference is visible and the occupied band is expendable. When either condition fails, a reliable decoder should withhold its prediction rather than erase task information or trust an unseen spectrum.

AI use statement

Generative AI tools were used to assist with preliminary brainstorming of the research hypothesis, routine code implementation and debugging, data tabulation/result formatting, and language polishing of the manuscript. All experimental designs, scientific interpretations of the results, and final conclusions were conceived, verified, and directed by the human authors. The authors take full responsibility for the integrity, claims, and final content of this work.

Ethics statement

The experiments use public, de-identified EEG datasets and do not collect new human data. We follow the dataset licenses and do not make clinical or safety-critical claims. The simulator is intended for robustness evaluation; it should not be used to infer an individual's health status or to design an unvalidated biomedical device.

Reproducibility statement

Section 3 specifies the tasks, subject splits, interference families, and evaluation settings; Sections 4–5 define the method and prove its properties. The appendices report the full result matrices, ablations, sensitivity analyses, and calibration comparison. The accompanying anonymous code attachment includes data-processing instructions, experiment configurations, fold-level results, and scripts for Figures 2–4, together with numerical checks of the spectral gate. The public EEG datasets are available from their original providers.

References

Coen Adler, Yuxin Chang, Felix Draxler, Samar Abdi, and Padhraic Smyth. Beyond accuracy: Are time series foundation models well-calibrated? In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.16060.

Navid Akhavan Attar, Hesam Asadollahzadeh, Ling Luo, and Uwe Aickelin. Softmax is not enough (for adaptive conformal classification). In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2602.19498.

Clemens Brunner, Robert Leeb, Gernot Muller-Putz, Alois Schloegl, and Gert Pfurtscheller. BCI competition 2008–graz data set a. In *Graz BCI Competition*, 2008.

Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, 2019.

Chenda Duan, Yipeng Zhang, Sotaro Kanai, Yuanyi Ding, Atsuro Daida, Pengyue Yu, Tiancheng Zheng, Naoto Kuroda, Shaun A. Hussain, Eishi Asano, Hiroki Nariai, and Vwani Roychowdhury. Omni-iEEG: A large-scale, comprehensive iEEG dataset and benchmark for epilepsy research. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2602.16072.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations*, 2020.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard L. Phillips, Sara Beery, et al. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, 2021.

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), 2018.

Wilson Lim, Olga Sourina, and Libin Wang. STEW: Simultaneous task EEG workload data set. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(11):2216–2225, 2018.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. In *International Conference on Machine Learning*, 2022.

Hermann Rohling. Radar CFAR thresholding in clutter and multiple target situations. *IEEE Transactions on Aerospace and Electronic Systems*, AES-19(4):608–621, 1983.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*, 2020.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas D. J. Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, and Tonio Ball. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, 38(11):5391–5420, 2017.

Hao Song, Qingshan She, Feng Fang, Su Liu, Yun Chen, and Yingchun Zhang. Domain generalization through latent distribution exploration for motor imagery EEG classification. *Neurocomputing*, 614:128889, 2025. doi: 10.1016/j.neucom.2024.128889.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. TENT: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.

Yiming Wang, Bin Zhang, and Yujiao Tang. DMMR: Cross-subject domain generalization for EEG-based emotion recognition via denoising mixed mutual reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 628–636, 2024a. doi: 10.1609/aaai.v38i1.27819.

Zixin Wang, Yadan Luo, Liang Zheng, Zhuoxiao Chen, Sen Wang, and Zi Huang. In search of lost online test-time adaptation: A survey. *International Journal of Computer Vision*, 133(3):1106–1139, 2024b. doi: 10.1007/s11263-024-02213-5.

Junyi Yao, Parham Eftekhar, Yao Wang, Gene Cheung, Xujin Chris Liu, and Wei Hu. Lightweight transformer for EEG classification via balanced signed graph algorithm unrolling. In *International Conference on Learning Representations*, 2026. URL `https://arxiv.org/abs/2510.03027`.

Ruiyao Zhang and Ping Zhou. Uncertainty quantification based on conformal prediction for industrial time series with distribution shift. *IEEE Transactions on Industrial Informatics*, 21(5):3676–3685, 2025. doi: 10.1109/TII.2025.3529920.

Xiao-Cong Zhong, Qisong Wang, Dan Liu, Zhihuang Chen, Jing-Xiao Liao, Jinwei Sun, Yudong Zhang, and Feng-Lei Fan. EEG-DG: A multi-source domain generalization framework for motor imagery EEG classification. *IEEE Journal of Biomedical and Health Informatics*, 29(4):2484–2495, 2025. doi: 10.1109/JBHI.2024.3431230.

A Clean plug-in gate check

Applying the gate to clean EEG measures the cost of changing the decoder's input representation. The STEW clean-gating evaluation retrains ERM on the same ten folds with the same settings; the BCIC-LR evaluation uses the models from the nine-fold main experiment. On STEW, the main clean ERM reference is 0.858 ± 0.087; the separately retrained models reach 0.788 ± 0.086 with the gate. This comparison includes training variability as well as gating. The corresponding BCIC-LR gated value is 0.559 ± 0.044.

B Complete result matrices

The full-family matrices below evaluate independently retrained ERM models, frozen across test conditions, with separate interference draws. Their F1/F4 entries can therefore differ from Table 1; the paired STEW recovery and BCIC-LR no-recovery pattern remains the same. Bold marks the higher accuracy within each raw/gated pair; exact ties are left unbolded.

Table 2: STEW full-family stress test, ten prespecified LOSO folds. Accuracy is mean ± fold standard deviation for a frozen ERM decoder.

family	SIR	raw	sensing-gated	Δ
F1 narrowband	10	0.707 ± 0.133	0.772 ± 0.111	$+0.065$
	0	0.574 ± 0.086	0.772 ± 0.105	$+0.198$
	-10	0.509 ± 0.016	0.753 ± 0.101	$+0.244$
	-20	0.500 ± 0.000	0.648 ± 0.115	$+0.148$
F2 modulated	10	0.695 ± 0.130	0.760 ± 0.084	$+0.065$
	0	0.559 ± 0.087	0.710 ± 0.099	$+0.151$
	-10	0.503 ± 0.006	0.666 ± 0.089	$+0.163$
	-20	0.500 ± 0.000	0.597 ± 0.076	$+0.097$
F3 pulsed	10	0.734 ± 0.115	0.735 ± 0.108	$+0.001$
	0	0.621 ± 0.104	0.682 ± 0.095	$+0.061$
	-10	0.528 ± 0.041	0.604 ± 0.067	$+0.077$
	-20	0.503 ± 0.006	0.557 ± 0.051	$+0.053$
F4 broadband	10	0.749 ± 0.135	0.748 ± 0.131	-0.001
	0	0.573 ± 0.095	0.570 ± 0.090	-0.003
	-10	0.500 ± 0.000	0.500 ± 0.000	$+0.000$
	-20	0.500 ± 0.000	0.500 ± 0.000	$+0.000$

Table 3: BCIC-LR full-family stress test, nine LOSO folds on left/right-hand imagery. Accuracy is mean ± fold standard deviation for a frozen ERM decoder.

family	SIR	raw	sensing-gated	Δ
F1 narrowband	10	0.647 ± 0.117	0.543 ± 0.059	-0.103
	0	0.593 ± 0.078	0.554 ± 0.054	-0.039
	-10	0.552 ± 0.067	0.546 ± 0.063	-0.006
	-20	0.508 ± 0.050	0.539 ± 0.055	$+0.030$
F2 modulated	10	0.639 ± 0.131	0.582 ± 0.070	-0.057
	0	0.620 ± 0.094	0.555 ± 0.065	-0.065
	-10	0.546 ± 0.068	0.535 ± 0.047	-0.011
	-20	0.512 ± 0.023	0.502 ± 0.033	-0.011
F3 pulsed	10	0.648 ± 0.115	0.566 ± 0.058	-0.082
	0	0.622 ± 0.100	0.542 ± 0.072	-0.080
	-10	0.539 ± 0.051	0.546 ± 0.086	$+0.006$
	-20	0.497 ± 0.024	0.540 ± 0.037	$+0.043$
F4 broadband	10	0.651 ± 0.114	0.568 ± 0.085	-0.083
	0	0.606 ± 0.100	0.600 ± 0.078	-0.007
	-10	0.549 ± 0.056	0.566 ± 0.056	$+0.018$
	-20	0.522 ± 0.027	0.508 ± 0.029	-0.014

C BASELINES AND ABLATIONS

Table 4: Additional STEW baselines, ten-fold grouped LOSO. TENT adapts on and is evaluated on the same unlabeled target batch; all entries are accuracy means over folds.

method	clean	F1 0 dB	F1 −10 dB	F4 −10 dB
ERM	0.839	0.573	0.508	0.500
Band-stop augmentation	0.828	0.603	0.505	0.500
TENT	**0.855**	**0.660**	**0.544**	**0.501**
SimpleCNN	0.787	0.537	0.506	0.500

D SENSING DIAGNOSTICS

Table 5: STEW sensing summary across the four SIR levels. Detection is reported as the observed range across SIR; SIR MAE is averaged over non-clean rows.

family	detection (2 s)	detection (4 s)	SIR MAE (dB)
F1 narrowband	100%	100%	16.9 / 18.1
F2 modulated	86–100%	98–100%	7.8 / 8.0
F3 pulsed	100%	100%	8.2 / 12.2
F4 broadband	0–4%	0–20%	44.9 / 43.4

The two SIR-MAE entries in each row correspond to 2 s and 4 s windows. The large F4 error reflects the low spectral contrast of broadband interference after whitening.

D.1 SENSOR AND MASK SENSITIVITY

We sweep the two sensor thresholds on the same fixed sample of 50 STEW windows used in Figure 2, with a separate fixed interference stream at -10 dB. F1 detection stays at 100% throughout the grid, whereas clean false positives and F4 detections change substantially (Table 6). The default $(\alpha_{\mathrm{dB}}, q_{\mathrm{floor}}) = (5, 0.5)$ was fixed before this evaluation and uses no test labels.

Table 6: Threshold sensitivity on 50 real STEW windows. Entries are detection rates; clean detection is a false positive. Bold marks the default setting.

α_{dB}	q_{floor}	clean	F1	F4
4	0.4	46%	100%	88%
4	0.5	26%	100%	30%
4	0.6	18%	100%	0%
5	0.4	22%	100%	24%
5	**0.5**	**10%**	**100%**	**4%**
5	0.6	8%	100%	0%
6	0.4	10%	100%	0%
6	0.5	4%	100%	0%
6	0.6	2%	100%	0%

Mask dilation provides the expected suppression–distortion trade-off (Table 7). For each contaminated window, the detected F1 mask is held fixed and applied to both the clean and contaminated signal. Suppression is the reduction in residual interference power after gating; distortion is $\|G_m(x) - x\|_2 / \|x\|_2$. The default two-bin margin gains 2.8 dB over no dilation while increasing average relative clean-signal distortion from 0.080 to 0.114.

Table 7: F1 mask-dilation sensitivity on 50 real STEW windows at -10 dB. Bold marks the default two-bin margin.

dilation (bins)	removed bins	suppression (dB)	clean distortion
0	11.4%	18.45	0.080
1	16.3%	20.09	0.099
2	**21.2%**	**21.25**	**0.114**
3	26.2%	22.32	0.131
4	31.1%	23.29	0.149

E SOURCE-SUBJECT CALIBRATION CHECK

The main coverage analysis uses a target-subject split. Here we examine calibration using source subjects only. For each held-out subject, calibration subjects stay outside model training and no target window contributes to the threshold. STEW uses five calibration subjects and BCIC-LR uses two. Evaluation uses 96 evenly spaced target windows on STEW and every other trial on BCIC-LR (72 trials). The table reports mean $\pm$ between-subject standard deviation.

Table 8: Source-subject conformal calibration check. The tuple is the number of training subjects, calibration subjects, and evaluated target windows per fold.

Dataset	train/cal/test	clean cov.	F1 ungated	F1 gated	gated $\lvert S\rvert$
STEW	$42/5/96$	0.902 ± 0.085	0.475 ± 0.055	0.878 ± 0.063	1.424 ± 0.154
BCIC-LR	$6/2/72$	0.872 ± 0.082	0.864 ± 0.032	0.877 ± 0.046	1.688 ± 0.088

F FULL 4-CLASS BCIC IV 2A BENCHMARK AND NEGATIVE RESULTS

The complete four-class task uses all 9 subjects and four labels (left hand, right hand, feet, tongue). We use the artifact-free set of 2,328 trials for the primary analysis and report all 2,592 trials as a sensitivity check. Each model is trained on eight source subjects and evaluated on the held-out subject; all seeds are fixed before model construction. Chance accuracy is 0.25.

Table 9 gives the complete nine-fold raw/gated values for seed 0 at F1 and F4, -10 dB. The three-seed summaries in the text average the same fold-level quantities over seeds 0, 1, and 2.

Table 9: Four-class BCIC accuracy by held-out subject (seed 0, artifact-free, SIR -10 dB). Raw and gate are evaluated on the same subject windows.

subject	F1 raw	F1 gate	SGC raw	SGC gate	F4 raw	F4 gate
1	.322	.359	.286	.330	.293	.286
2	.222	.319	.237	.267	.248	.248
3	.381	.367	.344	.296	.315	.311
4	.324	.302	.252	.263	.271	.271
5	.271	.248	.267	.302	.225	.229
6	.311	.279	.301	.297	.324	.324
7	.277	.317	.328	.317	.269	.266
8	.314	.383	.314	.356	.292	.292
9	.321	.354	.266	.333	.350	.350
mean	.305	.325	.288	.307	.288	.286

Across seeds 0–2, the ERM F1 gate-minus-raw difference is $+0.018 \pm 0.044$ (95% descriptive t interval $[+0.001, +0.036]$), and the SGC difference is $+0.035 \pm 0.043$ (interval $[+0.017, +0.052]$). The corresponding F4 differences are below 0.001 in magnitude. Increasing EEGNet width/depth, training for 100 epochs, changing dropout, or increasing augmentation ratios did not improve the

F1–F4 trade-off in the 3-subject, two-seed pilot (Table 10). Multi-family F1/F4 augmentation raised F4 to only 0.315 gated in the full nine-fold run, with no raw-to-gated gain. These negative results indicate that the observed limits arise from spectral observability and task information, rather than a simple capacity shortfall.

Table 10: Four-class model and augmentation screening (pilot means over three held-out subjects and two seeds).

configuration	clean	F1 raw	F1 gate	F4 raw	F4 gate
base EEGNet, 60 ep	.472	.303	.335	.291	.291
base EEGNet, 100 ep	.471	.319	.335	.284	.284
wide EEGNet	.451	.315	.312	.283	.282
deep EEGNet	.441	.296	.317	.284	.282
multi-family augmentation	.454	.309	.325	.302	.301

G Data and evaluation details

BCIC-LR comprises 1,296 left- and right-hand imagery trials, including 133 trials marked as containing artifacts in the original release. The clean condition denotes recordings without added EMI and retains these existing artifacts. The analysis windows start 0.504 seconds after the cue and span 3.5 seconds; the 4 ms difference from the nominal start is one original sampling interval. The full four-class training-session data contain 2,592 trials, of which 264 carry artifact flags.

Raw and gated evaluations use the same held-out subjects with separate interference draws. Paired comparisons therefore pair subject-level accuracies across conditions. STEW clean-gating results use a separate training run, as described in Appendix A. The original runs fixed the data-order seed but did not separately fix model initialization; the reported values come from the recorded runs.