UNDERSTANDING AND CONTROLLING BATCH NORMALIZATION STATISTICS FOR SOURCE-FREE CROSS-SUBJECT MOTOR-IMAGERY EEG DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-subject motor-imagery electroencephalography (MI-EEG) decoders often lose accuracy on unseen users because intersubject distribution shifts affect both the input signals and the intermediate feature distributions. Although batch normalization (BN) is widely used in EEG decoders, the role of its running mean and variance in cross-subject generalization remains unclear. We isolate this role by decomposing a representative source-free test-time adaptation (TTA) procedure into target-statistics recalibration and affine-parameter optimization. Four strictly matched conditions and a source-statistics restoration intervention are evaluated on three datasets with three BN-based backbones. Target BN statistic recalibration alone closely reproduces the composite adaptation gains at both the setting and matched-case levels. Disabling target statistic updates removes the mean gain, and restoring source statistics largely eliminates it in the six BCI settings with restoration results. However, unconditional recalibration can still degrade individual target cases. We therefore treat BN recalibration as a selective action and evaluate a label-free risk gate under nested leave-one-subject-out (LOSO) calibration. Using 23 target-time diagnostics, the gate selects either BN recalibration or source-only inference for each offline target case, after observing its unlabelled evaluation inputs. It reduces negative-transfer rates in all nine evaluated settings, although the retained gain varies. These results identify target BN statistics as a source of adaptation gain in the evaluated TTA procedure and show how the associated risk can be reduced within the studied protocols, without establishing prospective or cross-dataset safety.

1 INTRODUCTION

Motor-imagery electroencephalography (MI-EEG) allows a brain–computer interface (BCI) to infer movement intention without overt muscular activity. In deployment, a decoder trained on previously recorded users must generalize to a user who was not observed during training. Differences in neurophysiology, electrode contact, recording conditions, and imagery strategies alter both the input signals and the internal feature distributions of the decoder (Jayaram et al., 2016; Lotte et al., 2018). A model that performs well on the source subjects may therefore lose accuracy on a held-out target subject. Labelled calibration data can reduce this loss, but collecting labels for every new user is time-consuming and increases the burden of deployment.

Cross-subject transfer learning and domain adaptation can reduce the need for calibration, but many approaches require source recordings, target labels, or joint optimization over source and target data (Jayaram et al., 2016; Liang et al., 2020). These requirements are difficult to satisfy when physiological recordings cannot be retained after training and target labels are unavailable at deployment. Source-free test-time adaptation (TTA) instead updates a pretrained source model using only unlabelled target inputs (Li et al., 2024; Sun et al., 2020). Because normalization layers are common in EEG decoders and often change during test-time updates, understanding the adaptation mechanism requires separating the effect of BN from the effects of other updated components.

A BN layer standardizes intermediate activations with a running mean and variance before applying learned affine scale and shift parameters (Ioffe & Szegedy, 2015). Most cross-subject EEG studies

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.

nevertheless treat BN as an implementation detail and report only the final accuracy of a complete adaptation pipeline. They rarely distinguish the replacement of source running statistics with target statistics from the optimization of trainable affine parameters. As a result, the operation responsible for an adaptation gain remains uncertain, as does the component that causes harm when adaptation fails.

We study this ambiguity with a two-stage mechanism-to-control framework. First, a representative test-time update is decomposed into recalibration of the BN running mean and variance and optimization of the BN affine scale and shift. Strictly matched comparisons, a no-statistics negative control, and a source-statistics restoration intervention isolate the component associated with the observed adaptation gain. Second, that component becomes an optional action for an unseen target case. A nested subject-level LOSO gate uses 23 diagnostics computed from unlabelled prediction distributions and adaptation records to choose between BN statistics recalibration and source-only inference.

Experiments on three datasets and three BN-based backbones show that BN statistics-only recalibration closely reproduces the gain of the composite adaptation procedure in all nine dataset–backbone settings. The no-statistics control does not preserve this improvement. Despite its positive mean effect, unconditional recalibration still degrades individual target cases. The proposed gate reduces exposure to such degradation in every setting, with a setting-dependent cost in retained gain. Our contributions are:

- **Operation-level mechanism analysis.** We separate target-statistics recalibration from affine-parameter optimization through strictly matched conditions, a negative control, and source-statistics restoration.
- **Case-level risk characterization.** Matched-case gains and the negative-transfer rate expose a risk of harm to individual target cases that pooled accuracy does not reveal.
- **Selective label-free control.** We evaluate a nested-LOSO routing rule based on 23 target-time diagnostics under an offline, transductive protocol; the contribution is an empirical audit of this decision, not a new optimization algorithm.

2 Related Work

Cross-subject EEG transfer. EEG signals and their discriminative patterns vary across users, sessions, acquisition systems, and experimental conditions (Jayaram et al., 2016; Lotte et al., 2018). Compact convolutional and attention-based decoders improve the learned representation (Lawhern et al., 2018; Schirrmeister et al., 2017; Song et al., 2023), while Euclidean alignment and its source-free variants reduce covariance mismatch across subjects (He & Wu, 2020; Wang et al., 2025). These methods improve cross-subject decoding, but a method that improves mean accuracy may still harm an individual target, a failure known as negative transfer (Wang et al., 2019).

Source-free and test-time adaptation. Source-free domain adaptation updates a pretrained model without retaining source recordings (Li et al., 2024; Liang et al., 2020). TTA methods instead differ in the signal and parameters used at inference: test-time training optimizes a self-supervised objective (Sun et al., 2020), TENT updates normalization-related parameters by minimizing prediction entropy (Wang et al., 2021), and T3A adjusts class prototypes without backpropagation (Iwasawa & Matsuo, 2021). Reliability-aware approaches filter uncertain samples or stabilize updates (Niu et al., 2022; 2023), while evaluation audits show that model selection and test protocols can change conclusions (Zhao et al., 2023). Online EEG studies have compared alignment, adaptive BN, and entropy minimization for calibration-free MI decoding (Wimpff et al., 2024). Most studies still evaluate complete pipelines, obscuring whether gains arise from statistics replacement, parameter optimization, or their interaction.

BN mechanisms and selective control. Adaptive BN replaces source statistics with target-domain statistics (Li et al., 2016), and prediction-time covariate-shift adaptation shows that recalibration can improve robustness (Schneider et al., 2020). TTN further demonstrates that direct target-batch normalization becomes unreliable when batches are small or nonindependent (Lim et al., 2023). Selective prediction permits a model to abstain when estimated risk is high (Geifman & El-Yaniv, 2017; 2019). These lines of work motivate selective application of an adaptation operation, but they

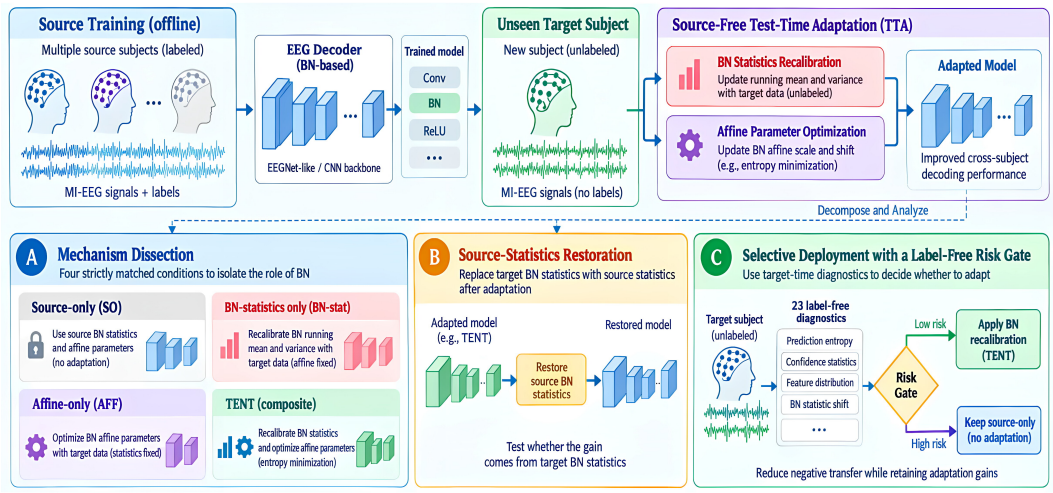


Figure 1: Source-free LOSO framework for mechanism analysis and case-level risk control. Labelled source subjects train a shared decoder; each unseen target subject provides disjoint unlabelled adaptation and held-out evaluation subsets. The evaluation inputs, but not their locked labels, are available for gate diagnostics.

do not determine whether BN-statistics recalibration should be applied to an unseen EEG subject or measure the resulting negative transfer against a paired source-only model. We address these questions for the evaluated TENT procedure by controlling target-statistics recalibration and affine optimization separately, restoring source statistics as an intervention, and learning a label-free binary decision from historical LOSO cases. The present comparisons isolate the BN-statistics pathway rather than benchmark the broader set of source-free adaptation methods.

3 Methodology

3.1 Overview and source-free protocol

Figure 1 summarizes the two-stage methodology. Labelled source subjects train a shared decoder, while the trials of each unseen target subject are divided before cropping into an unlabelled adaptation subset and a locked evaluation subset. The lower-left panel isolates BN statistics through four matched conditions and a restoration intervention. The lower-right panel treats the identified operation as a case-level decision and selects BN-only recalibration or source-only inference.

Let $\mathcal{S}$ denote the source subjects and $t \notin \mathcal{S}$ the target subject held out in an outer LOSO fold. The source predictor is trained only on labelled source trials:

$$\mathcal{D}_{\mathcal{S}} = \bigcup_{r \in \mathcal{S}} \mathcal{D}_r, \tag{1}$$

$$\theta_s = \arg\min_{\theta} \frac{1}{|\mathcal{D}_{\mathcal{S}}|} \sum_{(x,y) \in \mathcal{D}_{\mathcal{S}}} \mathcal{L}_{\text{sup}}(f_\theta(x), y). \tag{2}$$

Neither source recordings nor target labels are accessible when the target-time decision is made. The target trials form disjoint sets $\mathcal{D}_t^u = \{x_i^t\}_{i=1}^{n_t^u}$ and $\mathcal{D}_t^e = \{(x_j^t, y_j^t)\}_{j=1}^{n_t^e}$. Only $\mathcal{D}_t^u$ updates BN statistics; evaluation labels affect neither adaptation nor risk estimation. Unlabelled evaluation inputs are used to compute prediction-marginal diagnostics for both candidate predictors before the case-level decision. This offline, transductive use of the evaluation inputs does not imply a pre-inference decision for the first target trial.

3.2 BN STATISTICS AND CONTROLLED ISOLATION

For BN layer ℓ, running mean μ_ℓ, variance σ_ℓ^2, affine scale γ_ℓ, and shift β_ℓ, the transformation is

$$\widehat{h}_{\ell,j} = \frac{h_{\ell,j} - \mu_\ell}{\sqrt{\sigma_\ell^2 + \epsilon}}, \tag{3}$$

$$\mathrm{BN}_\ell(h_{\ell,j}) = \gamma_\ell \widehat{h}_{\ell,j} + \beta_\ell. \tag{4}$$

BN-only target recalibration freezes every trainable parameter and updates the running statistics over target batches. With momentum ρ, batch mean $\mu_{\ell,k}^{\mathcal{B}}$, and variance $v_{\ell,k}^{\mathcal{B}}$,

$$\mu_\ell^{t,(k)} = (1-\rho)\mu_\ell^{t,(k-1)} + \rho\mu_{\ell,k}^{\mathcal{B}}, \tag{5}$$

$$(\sigma_\ell^{t,(k)})^2 = (1-\rho)(\sigma_\ell^{t,(k-1)})^2 + \rho v_{\ell,k}^{\mathcal{B}}. \tag{6}$$

The source statistics initialize the recursion; after K batches, the predictor uses target statistics while retaining all learned parameters.

We use a TENT-style entropy-minimization condition (Wang et al., 2021) as a controlled probe, optimizing BN affine parameters $\phi = \{\gamma_\ell, \beta_\ell\}_{\ell \in \mathcal{L}_{\mathrm{BN}}}$ while updating target BN running statistics. Unlike the original TENT implementation, which uses current test-batch statistics without stored running statistics, this condition retains the EMA updates above so that affine optimization can be compared with BN-only recalibration under the same statistic-update rule:

$$\min_\phi \; -\frac{1}{|\mathcal{B}_k|} \sum_{x \in \mathcal{B}_k} \sum_{c=1}^{C} p_\phi(c \mid x) \log p_\phi(c \mid x). \tag{7}$$

Let Φ^s and S^s denote source affine parameters and running statistics, and let Φ^t and S^t denote their target-time counterparts. Every matched case uses the same source checkpoint, adaptation subset, evaluation subset, and random seed under four conditions:

$$f_{\mathrm{Source}}(x) = f(x, \theta_s, \Phi^s, S^s), \tag{8}$$

$$f_{\mathrm{BN}}(x) = f(x, \theta_s, \Phi^s, S^t), \tag{9}$$

$$f_{\mathrm{TENT}}(x) = f(x, \theta_s, \Phi^t_{\mathrm{TENT}}, S^t), \tag{10}$$

$$f_{\text{no-stat}}(x) = f(x, \theta_s, \Phi^t_{\text{no-stat}}, S^s). \tag{11}$$

BN-only tests whether target statistics are sufficient; no-stat tests whether affine entropy minimization remains effective when source statistics are fixed. For the six BCI settings, we additionally restore the source statistics after full TENT while retaining its affine parameters:

$$f_{\mathrm{restore}}(x) = f(x, \theta_s, \Phi^t_{\mathrm{TENT}}, S^s). \tag{12}$$

3.3 NEGATIVE TRANSFER AND LABEL-FREE RISK CONTROL

Balanced accuracy (BAcc) is the primary metric. For historical matched case i, the retrospective BN gain is $g_i = \mathrm{BAcc}_i^{\mathrm{BN}} - \mathrm{BAcc}_i^{\mathrm{source}}$. A case is harmful when $g_i < 0$, helpful when $g_i > 0.002$, and ambiguous otherwise; ambiguous cases are excluded from binary-gate fitting. The 0.002 boundary removes near-zero positive gains from the training labels; it is not a clinically or statistically established benefit threshold. Reported negative-transfer rates still use the zero-gain boundary, and sensitivity to this training-label choice has not been established.

For each setting, a 23-dimensional diagnostic vector contains three BN-process diagnostics, six summaries each of the source and BN prediction marginals, five measures of their change, and three fields from the archived fixed-gate protocol. All quantities are computed without target labels. Within each outer target-subject fold, feature standardization, logistic regression, and threshold selection use only historical subjects. The main analysis fits a separate gate for each dataset–backbone setting. Appendix C also reports separately fitted gates that pool backbones within a dataset, pool all three datasets, or transfer across datasets; those results are not averages of the setting-specific gates. Given standardized diagnostics $\widetilde{z}_i$, the estimated probability that recalibration is helpful is

$$q_i = P(r_i = 1 \mid \widetilde{z}_i) = \frac{1}{1 + \exp[-(w^\top \widetilde{z}_i + b)]}. \tag{13}$$

The fold-specific threshold τ selects $f_{\mathrm{BN},t}$ when $q_t \geq \tau$ and otherwise falls back to f_{source}. We measure safety by the negative-transfer rate and utility by retained BN gain:

$$\mathrm{NTR} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}[\mathrm{BAcc}_i^{\mathrm{adapt}} < \mathrm{BAcc}_i^{\mathrm{source}}], \tag{14}$$

$$R_{\mathrm{gain}} = \frac{\sum_i g_i^{\mathrm{gate}}}{\sum_i g_i^{\mathrm{BN}}}. \tag{15}$$

Appendix A gives the complete diagnostic definitions.

4 Experiments

4.1 Experimental conditions

We evaluate BCI Competition IV 2a (BCI2a), BCI Competition IV 2b (BCI2b) (Tangermann et al., 2012), and the OpenBMI Motor Imagery Dataset (Lee et al., 2019). BCI2a contains nine subjects and four classes, BCI2b contains nine subjects and two classes, and OpenBMI provides two-class recordings from 54 subjects. EEG signals are resampled to 100 Hz and bandpass filtered between 0.5 and 30 Hz. Each BCI epoch spans -0.2 to 3.2 s relative to the task cue, with the precue interval used for baseline correction; OpenBMI spans 0 to 4 s. Standardization parameters are estimated only from source training data. Half of each target subject's trials form the unlabelled adaptation subset and half are reserved for evaluation.

The backbones are EEGNet (Lawhern et al., 2018), ShallowConvNet with BN (Schirrmeister et al., 2017), and ATCNet with BN (Altaheri et al., 2023). Five source-model seeds and five target-split seeds yield 225 matched cases per backbone for BCI2a and BCI2b; three of each yield 486 cases per backbone for OpenBMI, for 2,808 cases across nine settings. Every condition shares the source checkpoint, preprocessing, adaptation and evaluation trials, and random seeds. Target-time experiments use batch size 32, ten adaptation steps, BN momentum 0.1, and learning rate 10^{-4} for affine optimization. We hold these choices and the 50/50 target split fixed to study the BN mechanism rather than tune the adaptation protocol. Momentum 0.1 is the PyTorch BN default; the other settings are fixed study choices, not universal defaults. Prior BN adaptation studies motivate target-statistic estimation but also show its dependence on target batch size and composition (Lim et al., 2023; Schneider et al., 2020). TENT optimizes 80 target-time parameters for EEGNet and ShallowConvNet-BN and 192 for ATCNet-BN. Three nonstandard ATCNet smoke-test rows found during the audit are excluded.

At the setting level, paired equivalence testing uses two one-sided tests with a prespecified ± 0.005 BAcc margin. The reported intervals describe differences over matched subject–seed–split cases. Because repeated cases from one subject are correlated, these intervals and tests are descriptive at the case level and do not establish subject-level equivalence. Risk control uses a logistic gate and a gradient-boosted decision-tree (GBDT) reference under nested subject-level LOSO; feature processing, fitting, and threshold selection use only historical subjects.

4.2 Mechanism evidence

Figure 2 and Table 1 report the main results. BN-only recalibration improves mean BAcc over source-only inference in all nine settings, with gains from 0.0070 to 0.0523. TENT produces nearly identical gains of 0.0068 to 0.0525. In contrast, TENT without target BN statistic updates has a negative mean gain in every setting between -0.0003 and -0.0038. BN-only updates no trainable parameters, while the additional affine optimization of TENT does not consistently improve it.

Table 1: Mean BAcc (%) over strictly matched cases.

Dataset	Backbone	Source	BN	TENT	No stat.
BCI2a	EEGNet	43.30	47.46	47.54	43.00
BCI2a	Shallow-BN	46.44	51.67	51.69	46.30
BCI2a	ATCNet-BN	51.54	54.30	54.28	51.16
BCI2b	EEGNet	59.97	61.18	61.20	59.77
BCI2b	Shallow-BN	60.87	62.65	62.67	60.84
BCI2b	ATCNet-BN	66.93	67.63	67.61	66.79
OpenBMI	EEGNet	76.73	78.69	78.66	76.43
OpenBMI	Shallow-BN	72.36	73.57	73.55	72.25
OpenBMI	ATCNet-BN	82.64	83.83	83.82	82.49

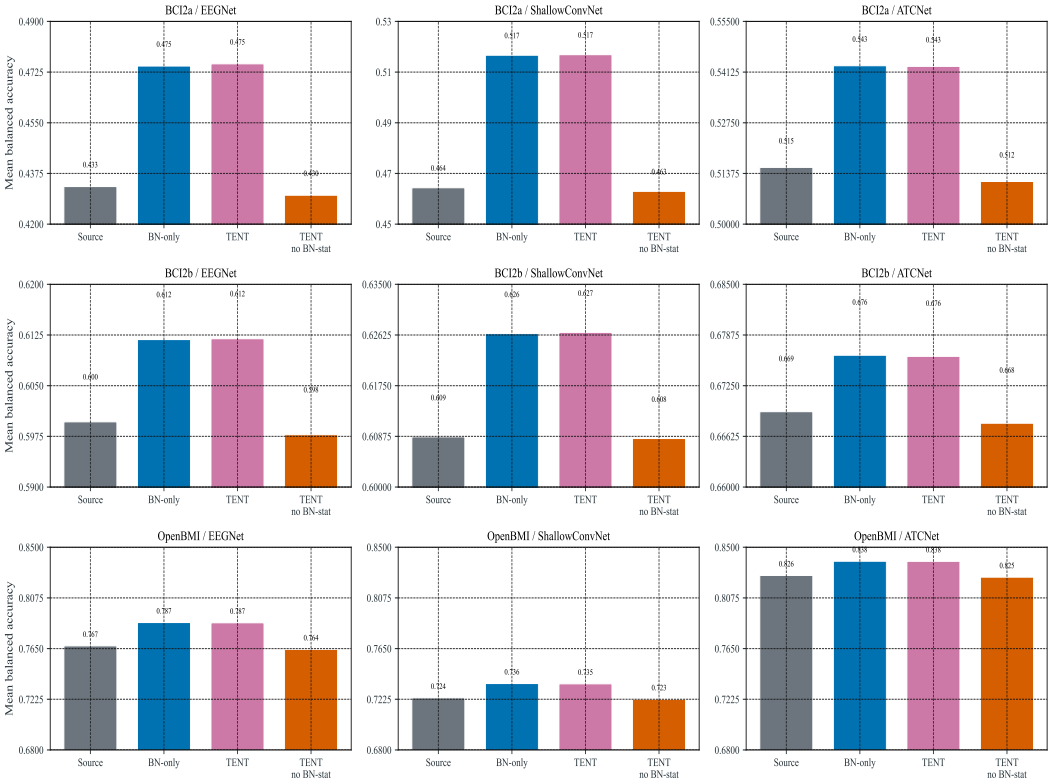


Figure 2: Four-condition mechanism scorecard across nine dataset–backbone settings. BN-only recalibration closely matches TENT, whereas affine entropy optimization without target BN statistic updates does not preserve the gain.

Setting-level averages do not reveal whether two procedures benefit the same targets. Across matched subject–seed–split cases, the gains of TENT and BN-only recalibration have Spearman correlations from 0.84 to 0.98, and their mean differences lie between -0.0003 and 0.0008 BAcc (Figure 3). In all nine settings, the reported case-level confidence interval of the paired TENT-minus-BN mean lies within the prespecified equivalence margin; this does not establish equivalence across independent target subjects. The no-stat negative control yields an average gain of approximately -0.002 BAcc. Restoring source BN statistics after TENT reduces the gain to approximately zero in the six BCI settings for which restoration is available (Figure 4). These comparisons support practical similarity over the matched cases. The negative control shows that affine optimization alone is insufficient in this procedure, and restoration shows that its adapted predictor remains dependent on target statistics after affine optimization.

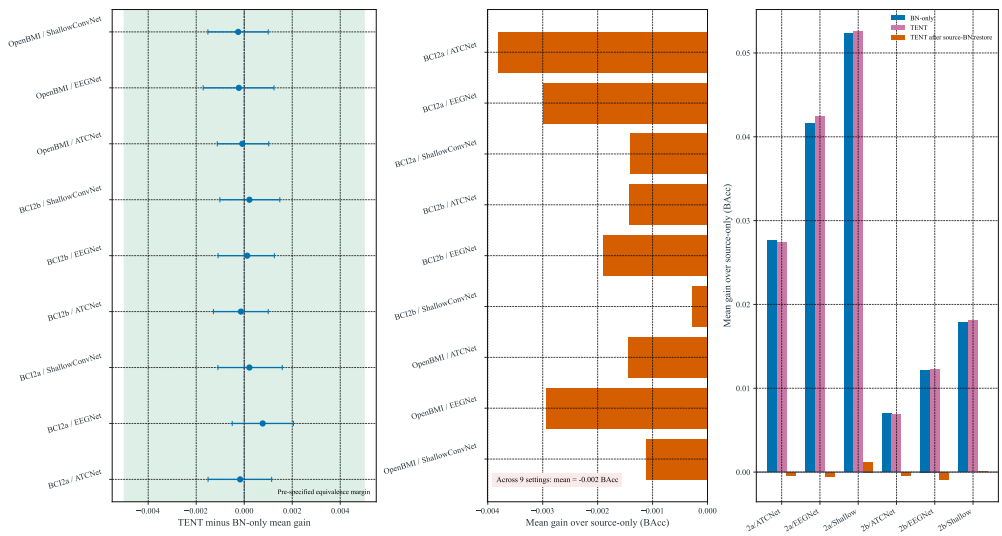


Figure 4: Intervention evidence for target BN statistics. (A) Matched-case TENT-minus-BN differences relative to the ± 0.005 BAcc margin in all nine settings; intervals use repeated subject–seed–split cases. (B) Negative control without target-statistic updates. (C) Source-statistics restoration after TENT in six BCI settings.

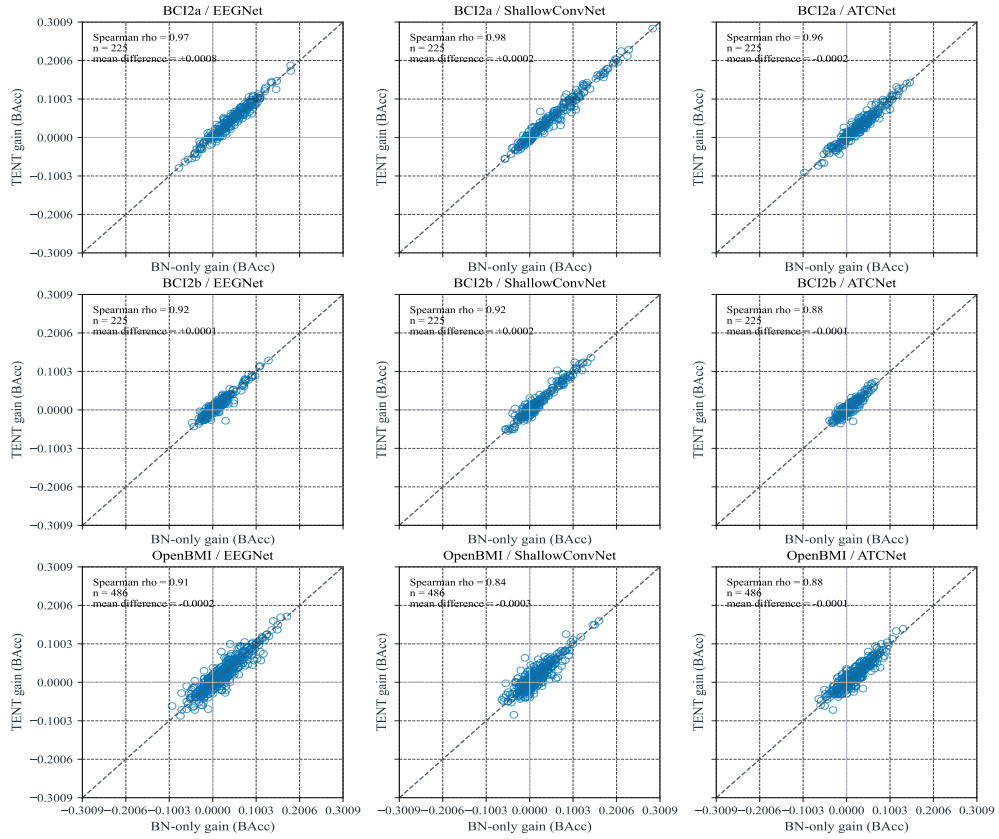


Figure 3: Matched-case alignment between TENT and BN-only gains. Each point is one shared target subject–source seed–split seed case; the dashed diagonal denotes exact agreement.

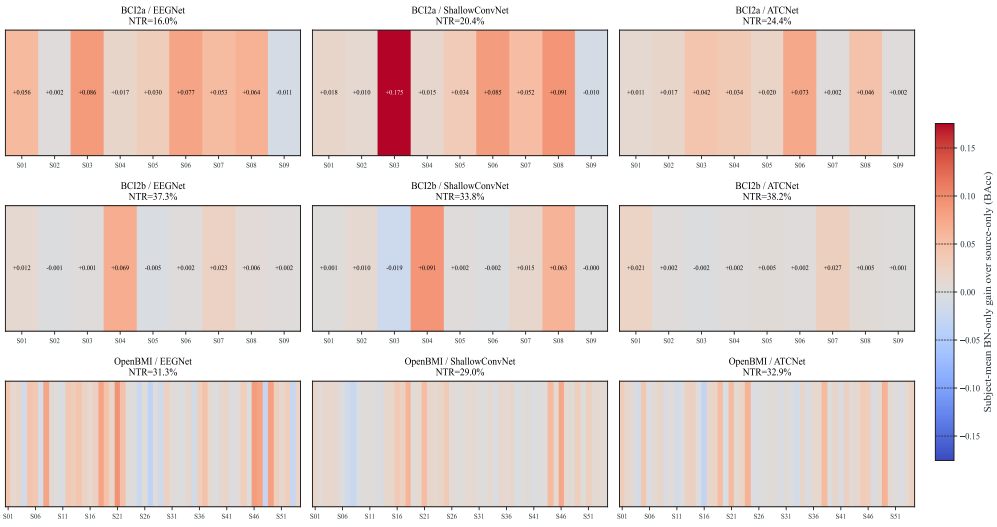


Figure 5: Subject-mean BN gain and matched-case NTR across nine settings. Each heat-map cell averages all source and target-split seeds for one held-out subject.

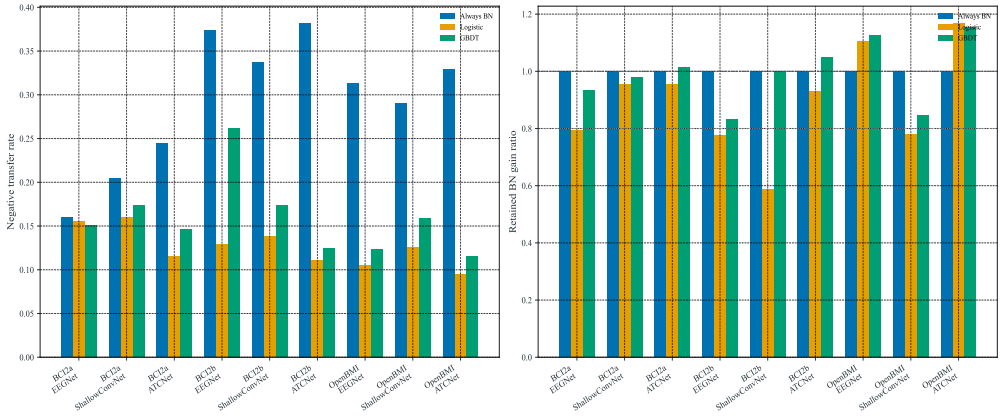


Figure 6: Setting-specific nested-LOSO risk-control audit using 23 saved label-free diagnostics. (A) NTR under unconditional BN recalibration and logistic/GBDT gates. (B) Retained BN gain; ratios above one can occur when harmful cases are rejected.

4.3 Negative transfer and risk control

The same useful BN operation is not safe for every target. Matched-case NTR ranges from 16.0% for BCI2a with EEGNet to 38.2% for BCI2b with ATCNet-BN; across OpenBMI it lies between 29.0% and 32.9%. Positive mean gains therefore coexist with substantial degradation risk for a particular source-seed and target-split realization. Subject-level aggregation shows that some subjects remain beneficial or harmful across seeds and splits, while cases near zero change outcome across experimental realizations (Figure 5).

Figure 6 and Table 2 evaluate the setting-specific gate. The logistic gate reduces NTR in all nine settings. The largest reductions are from 38.2% to 11.1% for BCI2b with ATCNet-BN and from 32.9% to 9.5% for OpenBMI with ATCNet-BN. The smallest is from 16.0% to 15.6% for BCI2a with EEGNet, where retained gain is 79.6%. Retained gain ranges from 58.8% to 116.9%; ratios above one occur when rejecting harmful cases increases average gain. The GBDT generally retains more utility but leaves a higher NTR in several settings. Neither gate is uniformly superior; the two gates define different operating points in the safety–utility trade-off.

When a single gate is refitted across all three backbones within each dataset, NTR falls from 20.3% to 12.9% on BCI2a, from 36.4% to 10.1% on BCI2b, and from 31.1% to 10.6% on OpenBMI; these are not averages of Table 2. Cross-dataset transfer retains only 0.06 of the BN gain on BCI2a and 0.30 on OpenBMI. The controller is therefore protocol-specific rather than a universal safety guarantee. Appendix C reports the representation and transfer analyses.

Table 2: Setting-specific logistic-gate results. NTR values are percentages; retained gain is relative to unconditional BN recalibration.

Dataset	Backbone	BN NTR	Gate NTR	Retained
BCI2a	EEGNet	16.0	15.6	0.796
BCI2a	Shallow-BN	20.4	16.0	0.955
BCI2a	ATCNet-BN	24.4	11.6	0.956
BCI2b	EEGNet	37.3	12.9	0.777
BCI2b	Shallow-BN	33.8	13.8	0.588
BCI2b	ATCNet-BN	38.2	11.1	0.930
OpenBMI	EEGNet	31.3	10.5	1.104
OpenBMI	Shallow-BN	29.0	12.6	0.780
OpenBMI	ATCNet-BN	32.9	9.5	1.169

5 Conclusion and Limitations

This work examines batch normalization as an explicit mechanism for source-free cross-subject MI-EEG generalization. Across three datasets and three BN-based backbones, target BN-statistics recalibration closely reproduces the gain of the evaluated TENT-style condition at setting and matched-case levels. Disabling target statistic updates removes the mean gain, and restoring source statistics largely eliminates it in the six BCI settings with restoration results. These comparisons identify the BN-statistics pathway under the fixed protocol, not the effect of an individual BN layer.

Unconditional recalibration yields NTR values from 16.0% to 38.2% despite a positive mean gain. A nested-LOSO gate lowers case-level NTR in every setting, but retained gain varies and cross-dataset transfer is unreliable. The gate is an offline routing rule calibrated from historical subjects: it uses the unlabelled evaluation inputs after both candidate predictors have been run, though never their labels. It does not establish a prospective first-trial safety guarantee.

These results do not cover other normalization schemes, and repeated cases within subjects do not support independent-user inference. Individual diagnostic contributions and sensitivity to the 0.002 label threshold and fixed BN protocol remain untested.

AI use statement

AI tools were used only to check word choice and grammar and to polish the writing. They were not used to develop the hypotheses, experimental design, data, results, or scientific claims. The authors take full responsibility for the scientific content and final manuscript.

Ethics statement

This study uses previously published, publicly available MI-EEG benchmark datasets and involves no new data collection or participant interaction. The evaluated offline protocol is source-free: source recordings and target labels are unavailable at target time. The experiments are not clinical evaluations. Any use beyond benchmark research would require separate validation and safeguards for physiological data.

Reproducibility statement

Section 3 defines the matched conditions, restoration intervention, risk metrics, and gate-calibration scopes. Section 4 specifies the datasets, preprocessing, model families, target split, seeds, adaptation hyperparameters, statistical tests, and case counts. Appendix A lists the complete diagnostic set. The remaining appendices report the negative-transfer, representation, and cross-dataset analyses.

REFERENCES

Hamdi Altaheri, Ghulam Muhammad, and Mansour Alsulaiman. Physics-informed attention temporal convolutional network for EEG-based motor imagery classification. *IEEE Transactions on Industrial Informatics*, 19(2):2249–2258, 2023.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 2151–2159, 2019.

He He and Dongrui Wu. Transfer learning for brain–computer interfaces: A euclidean space data alignment approach. *IEEE Transactions on Biomedical Engineering*, 67(2):399–410, 2020.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 448–456, 2015.

Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Vinay Jayaram, Morteza Alamgir, Yasemin Altun, Bernhard Schölkopf, and Moritz Grosse-Wentrup. Transfer learning in brain–computer interfaces. *IEEE Computational Intelligence Magazine*, 11(1): 20–31, 2016.

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.

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

Jingjing Li, Zhiqi Yu, Zhekai Du, Lei Zhu, and Heng Tao Shen. A comprehensive survey on source-free domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8): 5743–5762, 2024.

Yanghao Li, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu. Revisiting batch normalization for practical domain adaptation. *arXiv preprint arXiv:1603.04779*, 2016.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 6028–6039, 2020.

Hyesu Lim, Byeonggeun Kim, Jaegul Choo, and Sunghoon Choi. TTN: A domain-shift aware batch normalization in test-time adaptation. In *International Conference on Learning Representations*, 2023.

Fabien Lotte, Laurent Bougrain, Andrzej Cichocki, Maureen Clerc, Marco Congedo, Alain Rakotomamonjy, and Florian Yger. A review of classification algorithms for EEG-based brain–computer interfaces: A 10 year update. *Journal of Neural Engineering*, 15(3):031005, 2018.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 16888–16905, 2022.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *International Conference on Learning Representations*, 2023.

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

Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. EEG Conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2023.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 9229–9248, 2020.

Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J. Miller, Gernot R. Müller-Putz, Guido Nolte, Gert Pfurtscheller, Hubert Preissl, Gerwin Schalk, Alois Schlögl, Carmen Vidaurre, Stephan Waldert, and Benjamin Blankertz. Review of the BCI competition IV. *Frontiers in Neuroscience*, 6:55, 2012.

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.

Han Wang, Hao Han, John Q. Gan, and Hongtao Wang. Lightweight source-free domain adaptation based on adaptive euclidean alignment for brain–computer interfaces. *IEEE Journal of Biomedical and Health Informatics*, 29(2):909–922, 2025.

Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11293–11302, 2019.

Maximilian Wimpff, Mario Döbler, and Bin Yang. Calibration-free online test-time adaptation for electroencephalography motor imagery decoding. In *Proceedings of the 12th International Winter Conference on Brain–Computer Interface*, pp. 1–6, 2024.

Hao Zhao, Yuejiang Liu, Alexandre Alahi, and Tao Lin. On pitfalls of test-time adaptation. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 42058–42080, 2023.

A Complete Gate Diagnostics

For case i, let $a \in \{s, b\}$ index the source and BN-only predictors. From the same n_i^e locked evaluation inputs, the label-free class-prediction marginal is

$$m_i^a(c) = \frac{1}{n_i^e} \sum_{j=1}^{n_i^e} p_i^a(c \mid x_{i,j}^e). \tag{16}$$

Let $m_{i,(1)}^a$ and $m_{i,(2)}^a$ be its two largest components. Six summaries for each candidate are

$$H_i^a = -\sum_c m_i^a(c) \log m_i^a(c), \tag{17}$$

$$\overline{H}_i^a = H_i^a / \log C, \tag{18}$$

$$M_i^a = m_{i,(1)}^a, \tag{19}$$

$$S_i^a = \sqrt{C^{-1} \sum_c (m_i^a(c) - C^{-1})^2}, \tag{20}$$

$$G_i^a = m_{i,(1)}^a - m_{i,(2)}^a, \tag{21}$$

$$E_i^a = \exp(H_i^a). \tag{22}$$

Five additional diagnostics quantify source-to-BN marginal change:

$$D_{i,1} = \sum_c |m_i^s(c) - m_i^b(c)|, \tag{23}$$

$$D_{i,s\to b} = \sum_c \widetilde{m}_i^s(c) \log \frac{\widetilde{m}_i^s(c)}{\widetilde{m}_i^b(c)}, \tag{24}$$

$$D_{i,b\to s} = \sum_c \widetilde{m}_i^b(c) \log \frac{\widetilde{m}_i^b(c)}{\widetilde{m}_i^s(c)}, \tag{25}$$

$$\Delta H_i = H_i^b - H_i^s, \tag{26}$$

$$\Delta M_i = M_i^b - M_i^s, \tag{27}$$

where $\widetilde{m}_i^a(c) = \max\{m_i^a(c), \epsilon_m\}$. The remaining six fields are the mean and final BN adaptation objective, BN runtime, configured threshold, maximum marginal risk value, and entropy risk value retained by the archived fixed-gate protocol. Runtime depends on the platform, and the final three fields are legacy protocol variables. Their separate contributions have not been tested here. The decision is transductive at the matched-case level rather than a per-trial pre-inference gate: both candidate predictors are run on the unlabelled evaluation inputs before a choice is made for that case.

B Negative-Transfer Diagnostics

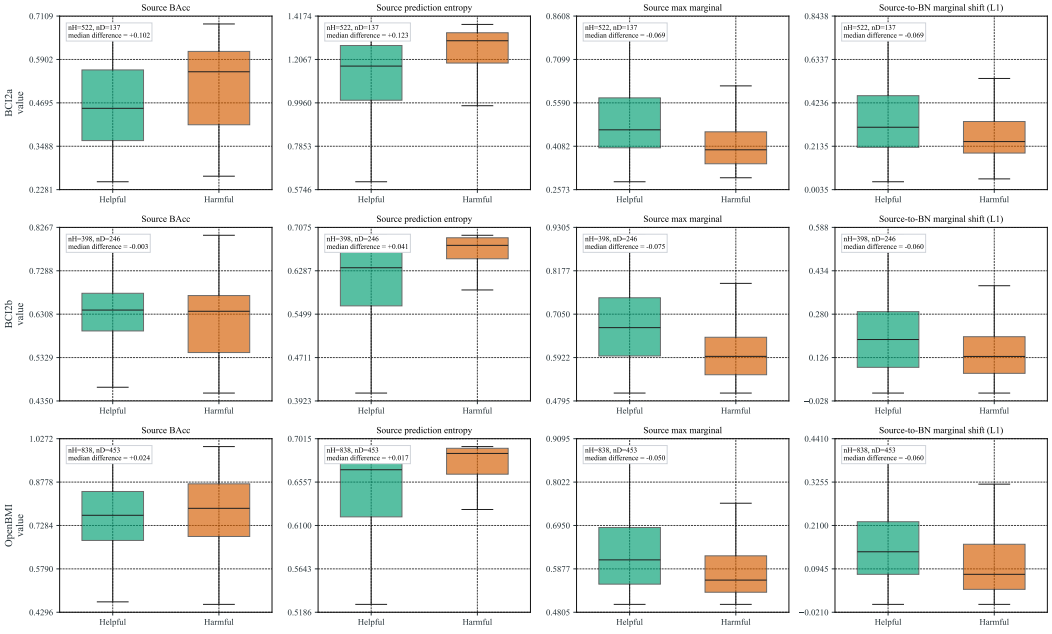


Figure 7: Retrospective comparison of helpful and harmful BN cases. Each dataset row pools the three backbones, yielding 675 cases for each BCI dataset and 1,458 cases for OpenBMI. The four columns show source BAcc, source-prediction entropy, maximum predicted marginal, and source-to-BN marginal change. The latter three are label-free target-time diagnostics; source BAcc is post hoc context and is excluded from the gate.

Figure 7 compares helpful and harmful cases after pooling the three backbones within each dataset. The source BAcc column provides retrospective context. On BCI2a, harmful cases have a higher median source BAcc than helpful cases, with a harmful-minus-helpful difference of $+0.102$. The

corresponding difference is -0.003 on BCI2b and $+0.024$ on OpenBMI. Source BAcc therefore does not separate the two outcome groups in a consistent direction across datasets. It is unavailable when the decision for a new target is made and is excluded from the gate.

The remaining columns use unlabelled predictions. Harmful cases have higher source-prediction entropy, a lower maximum source-prediction marginal, and a smaller source-to-BN change in the predicted class marginal across all three datasets. On BCI2a, the median harmful-minus-helpful differences are $+0.123$ for source-prediction entropy, -0.069 for the maximum marginal, and -0.069 for the marginal L_1 change. The same directions occur for BCI2b and OpenBMI, although the magnitudes differ. In particular, a smaller observed change after recalibration does not by itself imply that the operation was safe; the outcome must still be evaluated against the paired source-only predictor.

These quantities are operational signals, not universal causal explanations. Helpful and harmful labels are assigned retrospectively from evaluated historical gains and are used to fit and audit historical gate folds. The offline case-level decision for a new target uses neither BAcc nor its evaluation labels, but does use its unlabelled evaluation inputs. The overlap between the box plots also explains why the gate reduces risk without eliminating negative transfer.

C Representation Analysis and Protocol Boundary

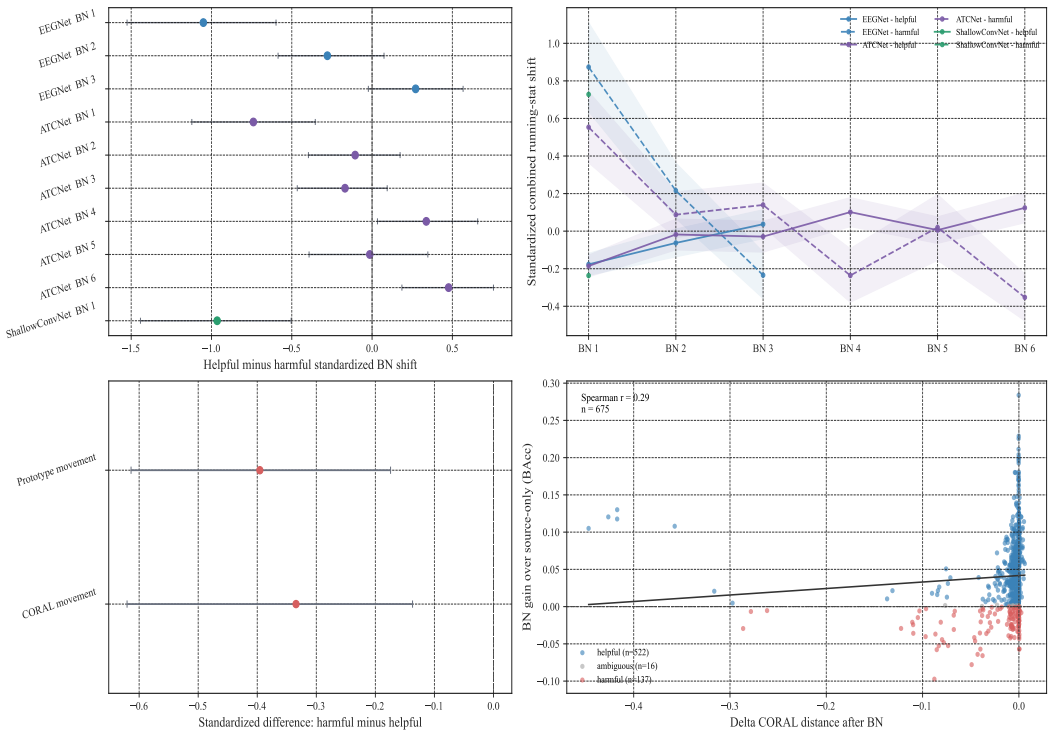


Figure 8: BCI2a representation fingerprints for helpful and harmful BN recalibration across three backbones. Panel A compares standardized BN-shift differences by layer and backbone; Panel B shows their variation across BN layers. Panel C compares prototype and CORAL movements. Panel D relates the change in CORAL distance to BN gain. Outcome groups are defined retrospectively and are not used during unsupervised adaptation.

Label-free prediction diagnostics distinguish harmful cases operationally but do not describe the underlying representation changes. All 675 matched BCI2a cases from the three backbones are analysed in Figure 8. Of these, 522 are helpful, 16 are ambiguous, and 137 are harmful under the thresholds used for the gate audit.

Panel A compares standardized helpful-minus-harmful BN-shift differences by layer and backbone. Their directions vary, so no single layer separates the outcome groups consistently across EEGNet, ShallowConvNet-BN, and ATCNet-BN. Panel B follows the differences across layers; the profiles are neither monotonic nor shared by the backbones. These layerwise measurements are not among the 23 diagnostics used by the current gate.

Panel C compares prototype and CORAL movement between the outcome groups. Harmful cases exhibit greater movement than helpful cases do. Under the sign convention in the figure, the standardized harmful-minus-helpful effects are approximately -0.40 and -0.33, respectively. Harmful recalibration is therefore associated with a distinct and often larger shift in the learned representation, rather than merely an absence of adaptation. This comparison is retrospective: the helpful and harmful groups are defined by the subsequent BAcc change.

Panel D relates the change in CORAL distance to the paired BN gain. Across the 675 BCI2a cases plotted here, the Spearman correlation is 0.29. Separate geometry-replication records yield approximately 0.14 on BCI2b ($n = 675$) and 0.01 on OpenBMI ($n = 4{,}050$); the latter uses more cases than the $1{,}458$ in the main risk-control audit. These two replication results are not plotted in Panel D. The BCI2a result is therefore an explanatory fingerprint rather than a universal risk relation, and the current gate does not use these layerwise or geometric measurements.

Figure 9 compares gate-calibration scopes. Panels A and B refit one gate across the three backbones within each dataset, excluding the held-out subject. In Panel A, logistic NTR falls from 20.3% to 12.9% on BCI2a, 36.4% to 10.1% on BCI2b, and 31.1% to 10.6% on OpenBMI. Averaging the separately fitted Gate NTR entries in Table 2 instead gives 14.4%, 12.6%, and 10.9%. The BN baselines share the same cases; the gate values differ because the fitting scope changes.

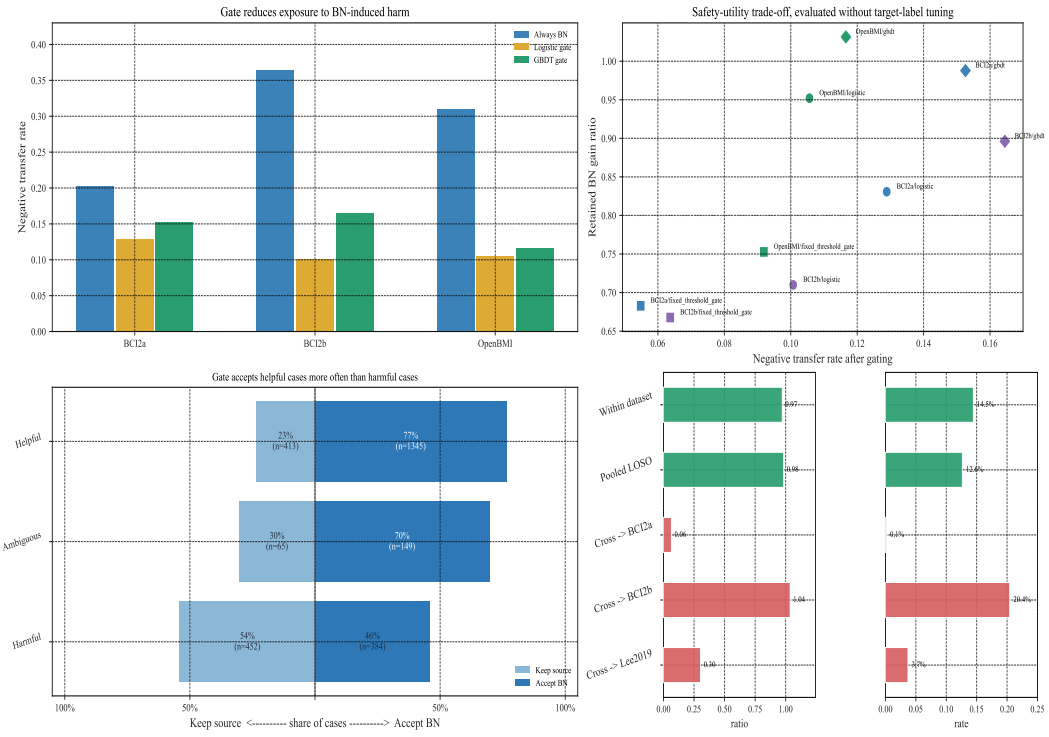


Figure 9: Aggregate risk-control behavior and protocol boundaries. Panels A and B use dataset-level gates fitted across the three backbones within each dataset, unlike the setting-specific gates in Table 2. Panel C audits acceptance decisions for historical cases. Panel D contrasts within-dataset and all-dataset pooled LOSO calibration with cross-dataset stress tests. "Lee2019" denotes OpenBMI.

Panel B places gated NTR beside retained BN gain. Neither logistic nor GBDT is uniformly superior on this trade-off. The archived fixed-threshold gate is a reference, not the controller fitted in nested LOSO.

Panel C audits decisions against post hoc outcomes. The GBDT accepts recalibration for 77% of helpful and 46% of harmful historical cases; ambiguous cases lie between. Thus the gate reduces but does not remove risk. These outcome labels do not route the held-out target.

Panel D compares within-dataset backbone-pooled calibration, calibration across all datasets, and cross-dataset transfer. The first two separately fitted protocols retain 0.97 and 0.98 of BN gain, with NTR values of 14.5% and 12.6%; neither is an average of Table 2. Training on other datasets retains only 0.06 of the gain on BCI2a, 1.04 on BCI2b with 20.4% NTR, and 0.30 on OpenBMI. The controller is therefore protocol-specific, not a universal safety guarantee.