EPITAR-BENCH: BENCHMARKING TRANSFERABILITY, ADAPTATION, AND RELIABILITY OF EPILEPSY MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Automated seizure analysis increasingly relies on deep learning and EEG foundation models, yet their transferability, adaptation, and reliability under heterogeneous recording conditions remain poorly characterized. In this paper, we introduce EpiTAR-Bench, a unified benchmark for the **T**ransferability, **A**daptation, and **R**eliability of epilepsy models across seizure detection and prediction. EpiTAR-Bench spans five datasets, seven transfer directions covering cross-dataset, cross-modality, and cross-time shifts, sixteen models from four families, two window lengths, and five levels of target supervision. We further present RIVER, a transfer-specific model designed for partially observed and heterogeneous electrode configurations. Across zero-shot cross-dataset transfer, EEG foundation models form the strongest baseline family, while longer window lengths provide no consistent advantage. Adaptation gains are concentrated at lower target supervision levels for both cross-dataset tasks and cross-modality detection ($p \leq 0.0004$ after Holm correction), whereas cross-modality prediction shows a positive but nonsignificant contrast ($p = 0.1605$) and cross-time transfer shows no comparable pattern. For seizure detection, scalp EEG transfer followed by target iEEG supervision narrows the gap to direct iEEG training, while prediction retains a larger residual gap. Finally, discrimination, calibration, and seizure onset zone (SOZ) evidence are not consistently aligned. These results establish EpiTAR-Bench as a systematic framework for evaluating how epilepsy models transfer, adapt, and remain reliable across heterogeneous EEG settings.

1 INTRODUCTION

Epilepsy affects around 50 million people worldwide (World Health Organization, 2024), making reliable seizure detection and prediction an important clinical objective. Electroencephalography (EEG) is central to diagnosis and monitoring and is therefore a key signal source for automated epilepsy models (Ramgopal et al., 2014). For instance, EEGNet (Lawhern et al., 2018) is a compact decoder designed for specific EEG tasks. Large-scale pre-trained models, including BENDR, BIOT, LaBraM, EEGPT, and CBraMod, aim to learn reusable EEG representations (Kostas et al., 2021; Yang et al., 2023; Jiang et al., 2024; Wang et al., 2024; 2025a). However, a model may encounter substantial shifts across cohorts, sensing modalities, or recording periods. Intracranial EEG (iEEG) poses an additional constraint because labeled recordings are sparse and harder to acquire. This raises a practical question of whether epilepsy models can transfer across heterogeneous EEG settings and adapt effectively when target supervision is limited, particularly for iEEG.

Current approaches use representation learning, pre-training at scale, and specialized transfer mechanisms. Yet two assumptions often remain implicit: a representation that performs well in its original evaluation setting will remain useful under distributional change, and additional target supervision fine-tuning will progressively recover transfer performance. Prior studies challenge both assumptions. Performance can degrade when the test distribution changes, and further adaptation does not always eliminate the gap (Koh et al., 2021; Taori et al., 2020; Kumar et al., 2022). Broad EEG comparisons likewise show that large-scale pre-trained models do not consistently dominate compact neural or classical decoders across tasks and evaluation protocols (Yang et al., 2026). Cross-modality epilepsy transfer requires explicit treatment of heterogeneous sensing domains (Wang et al., 2025b),

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.

while probability calibration may deteriorate under distributional change even when discrimination remains competitive (Guo et al., 2017; Ovadia et al., 2019; Minderer et al., 2021). These questions leave no clear consensus on when epilepsy models transfer, how much target supervision is needed to recover performance, and whether discrimination achieved through direct transfer or recovered through target adaptation is accompanied by reliable confidence and clinically grounded evidence.

Our work makes three main contributions: We first establish EpiTAR-Bench, a unified benchmark of transferability, adaptation, and reliability spanning cross-dataset EEG transfer, scalp EEG to iEEG transfer, and cross-time EEG transfer, covering seizure detection and prediction across five datasets, seven directions, 12s and 60s window lengths, and five target supervision levels from 0% to 100%. Based on this benchmark, we then provide a systematic evaluation of 16 models from four families, together with statistical testing, calibration analysis, and SOZ evidence analysis at the patient level. Finally, we present RIVER as a concrete methodological instantiation of the benchmark's heterogeneous transfer challenge: learning robust representations under partially observed electrode configurations. These contributions lead to the following main findings:

1. **EEG foundation models show the strongest cross-dataset transferability among baseline groups.** Under zero-shot cross-dataset transfer, EEG foundation models achieve the strongest aggregate baseline performance, while RIVER obtains the highest mean AUROC and best average rank. Yet this advantage remains transfer dependent rather than universal across all settings.
2. **Scalp EEG can support adaptation when iEEG labels are scarce.** For seizure detection in the cross-modality setting, evaluated models can approach the direct iEEG reference as target iEEG supervision fine-tuning is introduced, whereas seizure prediction retains a larger gap across supervision levels under heterogeneous scalp EEG to iEEG transfer.
3. **Lower target supervision levels often provide the largest performance recovery, while higher levels do not necessarily yield further gains.** Both cross-dataset tasks and cross-modality detection benefit disproportionately at the initial target supervision level, whereas subsequent levels provide much smaller or inconsistent improvements. This pattern is weaker for cross-modality prediction and is not clearly observed under cross-time transfer.
4. **Longer window lengths show no consistent effects.** Increasing window length does not consistently improve average performance and can even degrade detection under certain transfers. Model rankings can also change substantially despite little change in the overall metric means, showing that longer window lengths do not uniformly translate into more robust transfer.
5. **Reliability cannot be characterized solely by discrimination.** The relationship between discrimination and calibration varies across experimental settings, and the cross-modality detection SOZ evidence analysis further shows that classification rankings do not necessarily align with SOZ evidence across patients.

In general, our findings show that transferability, adaptation, and reliability in epilepsy models cannot be inferred from model family alone. Transfer direction, task, target supervision level, and window length can substantially alter performance, while discrimination, calibration, and spatial evidence may follow different trends. We hope EpiTAR-Bench can encourage broader and deeper evaluation of epilepsy models across heterogeneous datasets, modalities, and recording periods.

2 Related Work

2.1 EEG decoding for specific tasks and pre-trained representations

Epilepsy models traditionally include decoders designed for specific EEG tasks. EEGNet (Lawhern et al., 2018) provides a compact convolutional architecture, whereas ScatterFormer (Zheng et al., 2023), BF-EML (Wen et al., 2026), and EvoBrain (Kotoge et al., 2025) introduce more specialized inductive biases for EEG and seizure analysis. More recent work seeks reusable EEG representations through large-scale pre-training. BENDR (Kostas et al., 2021), BIOT (Yang et al., 2023), LaBraM (Jiang et al., 2024), EEGPT (Wang et al., 2024), CBraMod (Wang et al., 2025a), and ST-EEGFormer (Yang et al., 2026) span contrastive, masked reconstruction, and transformer architectures. Nevertheless, recent comparative evidence shows that larger-scale pre-trained models do not uniformly outperform compact neural or classical decoders (Yang et al., 2026). Thus, pre-training at scale alone does not guarantee robust transferability across heterogeneous EEG settings.

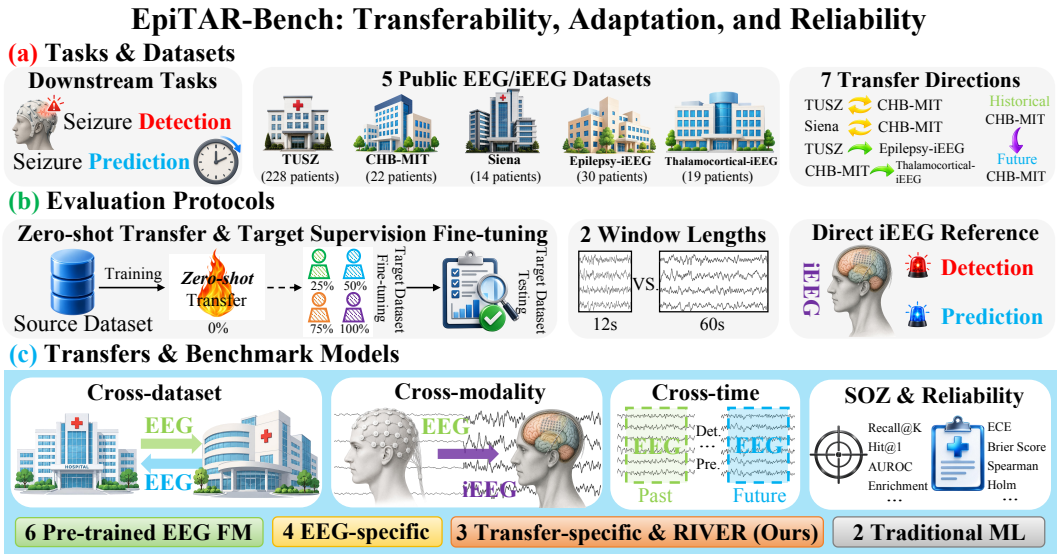

Figure 1: EpiTAR-Bench overview. (a) 2 tasks across 5 public EEG/iEEG datasets and 7 transfer directions. (b) Zero-shot and 25%–100% target supervision levels, 12s/60s window lengths, and direct iEEG references. (c) Cross-dataset, cross-modality, and cross-time transfer evaluation across 16 models: 6 pre-trained EEG foundation models, 4 EEG-specific models, 3 transfer-specific comparators, RIVER, and 2 traditional ML models, together with SOZ and reliability analyses.

2.2 Transfer across heterogeneous EEG settings

Most EEG transfer studies examine variation across subjects or datasets while preserving the sensing modality. Broader transfer settings include shifts in cohort, acquisition context, longitudinal recordings, and sensing modality. Transfer-specific methods such as CMMN (Gnassounou et al., 2023), TSMNet (Kobler et al., 2022), and CST (Wang et al., 2025b) address distribution mismatch and heterogeneous electrophysiological alignment, while recent work has also evaluated EEG foundation models under cross-dataset and transfer-learning settings, including ST-EEGFormer (Yang et al., 2026) and a broader FM generalization study (Kommineni et al., 2026). These challenges are particularly relevant for iEEG (Qamar et al., 2025): a key clinical question is whether knowledge learned from more accessible scalp EEG can support adaptation when labeled iEEG data are limited.

2.3 Adaptation under limited target supervision

After transfer, additional target supervision can support adaptation, especially when target data are limited. Prior work has explored transfer learning for seizure prediction with restricted target data (Lopes et al., 2024) and has shown more broadly that further fine-tuning does not always lead to continued improvement under distributional change (Kumar et al., 2022). Existing studies typically focus on a specific task, transfer, or adaptation setting, so the effect of increasing target supervision remains less systematically characterized. Our benchmark provides a comprehensive framework for characterizing when epilepsy models transfer, how effectively limited target supervision recovers performance, and whether such recovery remains robust and reliable across transfer settings.

3 Benchmark Methods

3.1 Evaluation protocols

Transfer Settings. Figure 1 provides an overview of our benchmark. We consider three transfer tracks: cross-dataset, cross-modality, and cross-time transfer, covering seven ordered directions. Cross-dataset transfer includes reciprocal TUSZ↔CHB-MIT detection and CHB-MIT↔Siena prediction. The cross-modality track evaluates heterogeneous scalp EEG to iEEG transfer using TUSZ→Epilepsy-iEEG detection and CHB-MIT→Thalamocortical-iEEG prediction, while cross-

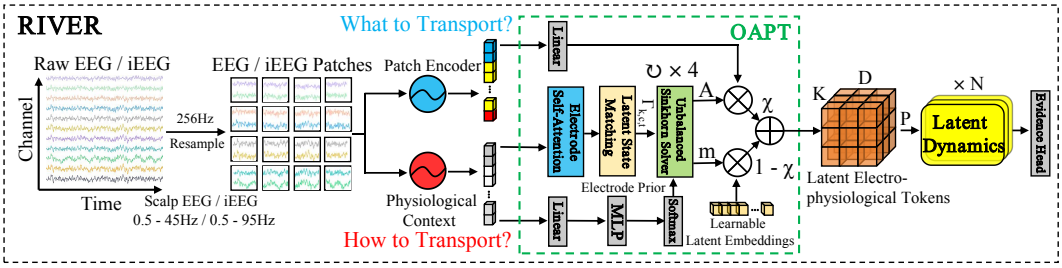

Figure 2: Overview of **R**obust **I**nference across **V**ariable **E**lectrophysiological **R**ecordings (RIVER). Observability-Aware Partial Transport (OAPT) maps variable native electrodes to shared latent states and retains transported mass as an observability signal. Electrode-wise content defines what is transported, while physiological state features guide adaptive electrode–latent pairing and partial transport before downstream latent dynamics.

time transfer evaluates historical→future CHB-MIT recordings within the same patients for both tasks. Each setting is evaluated under 12s and 60s window lengths and five target supervision levels.

Target Supervision Fine-tuning. We evaluate five target supervision levels: 0% (zero-shot), 25%, 50%, 75%, and 100%. For cross-dataset and cross-modality transfer, positive levels use nested target patient subsets, whereas cross-time transfer follows the within-patient chronological adaptation blocks described in Appendix A. Zero-shot assesses direct source-to-target transfer without target training data, with model selection and early stopping based on source validation data and target labels used only for final evaluation. At positive target supervision levels, the target validation labels used for checkpoint and threshold selection are counted as target supervision, while the target test set remains held out for final evaluation. Further details on data construction and transfer protocols are provided in Appendix A. Training and adaptation details are provided in Appendix B.

Metrics and Statistics. For all datasets and runs, we choose AUROC as our primary discrimination metric, complemented by F1 and balanced accuracy. Complementary adaptation analyses use AUPRC, expected calibration error (ECE), and Brier score. We also assess spatial evidence related to the SOZ in zero-shot cross-modality detection. P-values use paired sign-flip tests across 16 models for marginal supervision and window length effects, paired two-sided Wilcoxon tests for budget versus zero and model comparisons, and one-sided Wilcoxon tests for SOZ enrichment, with Holm correction within each comparison family. These tests quantify consistency within the evaluated model set, whereas analyses at the patient level use individual patients as the unit of inference. Further details on statistical testing and evaluation are provided in Appendix E and Appendix H.

3.2 Downstream Tasks and Datasets

We study two epilepsy tasks across five public EEG and iEEG datasets. **Seizure detection** distinguishes ictal from non-ictal activity and is evaluated on TUSZ (Shah et al., 2018), CHB-MIT (Guttag, 2010; Shoeb, 2009), and Epilepsy-iEEG (Li et al., 2026). **Seizure prediction** distinguishes preictal from interictal activity and is evaluated on CHB-MIT, Siena (Detti, 2020; Detti et al., 2020), and Thalamocortical-iEEG (Panchavati et al., 2026). After preprocessing, our benchmark retains 228 TUSZ patients, 22 CHB-MIT patients, 14 Siena patients, 30 Epilepsy-iEEG patients, and 19 Thalamocortical-iEEG patients, while 16 CHB-MIT patients satisfy the cross-time protocol.

Preprocessing. All datasets follow a standardized preprocessing pipeline. Signals are resampled to 256Hz, with a 0.5–45Hz passband for cross-dataset and cross-time experiments and a 0.5–95Hz passband for cross-modality experiments. Scalp EEG is represented with a common bipolar channel configuration, whereas iEEG retains its native electrode set to preserve heterogeneous spatial coverage. Signals are normalized independently for each channel and window length. To accommodate differences among benchmark models, we use input adapters tailored to each model that map the standardized signals to the input format required by each architecture while keeping the underlying benchmark samples unchanged. Further details on preprocessing and dataset construction are provided in Appendix A. Input adaptation for each model is documented in Appendix B.

3.3 BENCHMARK MODELS

3.3.1 EEG FOUNDATION MODELS

The large-scale pre-trained EEG foundation model group contains BENDR (Kostas et al., 2021), BIOT (Yang et al., 2023), LaBraM (Jiang et al., 2024), EEGPT (Wang et al., 2024), CBraMod (Wang et al., 2025a), and ST-EEGFormer (Yang et al., 2026). We use the officially released pre-trained weights and evaluate them under the common benchmark protocol. Further details on released weights and optimization settings are listed in Appendix B.

3.3.2 EEG-SPECIFIC MODELS

We include EEGNet (Lawhern et al., 2018), EvoBrain (Kotoge et al., 2025), ScatterFormer (Zheng et al., 2023), and BF-EML (Wen et al., 2026) in our EEG-specific model family. These models encompass compact convolutional architectures, dynamic graph models, scattering transformers, and debiasing methodologies designed for seizure analysis. Further details are in Appendix B.

3.3.3 TRANSFER-SPECIFIC MODELS

The transfer-specific group includes CST (Wang et al., 2025b), CMMN (Gnassounou et al., 2023), TSMNet (Kobler et al., 2022), and RIVER. CST targets heterogeneous electrophysiological alignment, CMMN addresses spectral mismatch, and TSMNet learns transfer-oriented geometric representations. Building on optimal transport for domain alignment (Courty et al., 2017), RIVER introduces Observability-Aware Partial Transport (OAPT), which maps variable native electrodes to a shared latent electrophysiological space and uses transported mass to represent latent-state observability without fixed EEG/iEEG electrode correspondence. Figure 2 illustrates the mechanism of RIVER. Further details are provided in Appendix C, Appendix D, and Appendix B.

3.3.4 TRADITIONAL MACHINE LEARNING

To examine whether classical approaches remain effective under transfer scenarios, we additionally include two traditional machine learning baselines: a support vector machine (SVM) (Cortes & Vapnik, 1995) and Random Forest (Breiman, 2001). More details are listed in Appendix B.

4 RESULTS

Complete transfer results for all directions and both window lengths are provided in Appendix F.

4.1 DO EPILEPSY MODELS GENERALIZE CONSISTENTLY ACROSS DATASETS?

Figure 3 summarizes zero-shot cross-dataset transfer by model family. Foundation models form the strongest baseline family (mean AUROC 0.649), while EEG-specific and traditional ML models show lower aggregate performance and transfer-specific methods exhibit greater variability, with lower AUROC generally corresponding to worse ranks. Across families, 12s window length generally achieves higher balanced accuracy than 60s, indicating no consistent benefit from longer window length in cross-dataset transfer. RIVER achieves the highest mean AUROC (0.690) and best average rank (4.50), with no significant differences from foundation or transfer-specific comparators after Holm correction. Additional cross-dataset spatial evidence appears in Appendix G.

4.2 DO EPILEPSY MODELS REMAIN STABLE UNDER CROSS-TIME TRANSFER?

Figure 4 shows that the relationship between discrimination and decision metrics varies across tasks and window lengths under cross-time transfer. In 12s detection, AUROC and balanced accuracy are most closely aligned, with several foundation and transfer-specific models concentrated in the upper-right quadrant. At 60s, models become more dispersed, including within the same family, and high AUROC does not always coincide with high balanced accuracy. Prediction shows weaker AUROC–F1 alignment across all model families. RIVER follows a distinct pattern, lying toward the upper-left in detection but moving into the upper-right at 60s prediction. Overall, cross-time robustness depends on task, window length, and evaluation metric rather than model family alone.

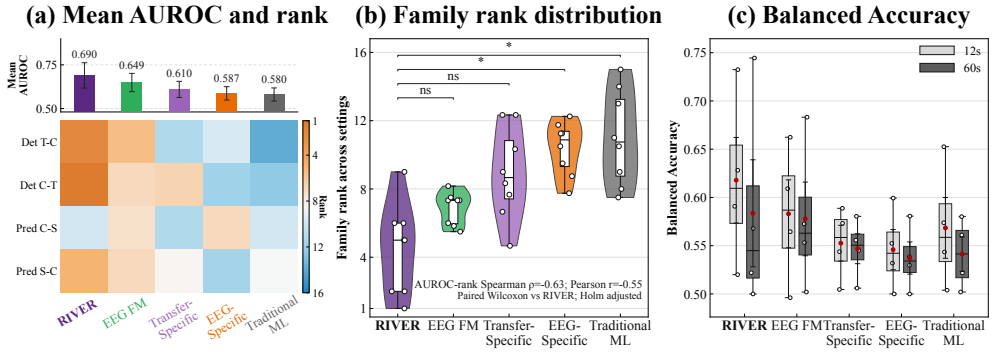

Figure 3: Zero-shot cross-dataset transfer by model group. (a) Mean AUROC and rank across four directions (T = TUSZ, C = CHB-MIT, S = Siena), aggregated over 12s and 60s. (b) Rank distributions across directions and window lengths; white points denote individual settings and the inset reports AUROC–rank correlations. RIVER–family Wilcoxon tests use eight seed-averaged window length×task×direction settings ($n = 8$), with Holm correction across four family comparisons. (c) Balanced accuracy at 12s/60s; red points and error bars show mean and SEM. Significance: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, and ns.

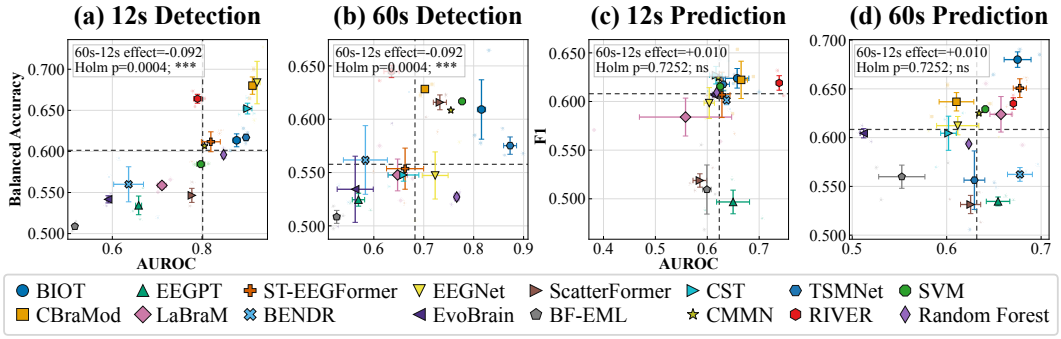

Figure 4: Zero-shot cross-time transfer. (a,b) Detection: AUROC versus balanced accuracy at 12s and 60s. (c,d) Prediction: AUROC versus F1 at 12s and 60s. Vertical and horizontal dashed lines mark the median model AUROC and task metric, respectively, within each panel, partitioning models into relative performance quadrants. Window length effects use AUROC for detection and F1 for prediction. Significance: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, and ns.

4.3 Can scalp EEG support adaptation when iEEG labels are scarce?

Figure 5 shows that gains are concentrated at lower target supervision levels for cross-modality detection, whereas prediction exhibits a smaller, nonsignificant contrast. RIVER achieves the strongest zero-shot detection performance, while the other model groups begin lower and improve partially as target iEEG supervision increases. The first-versus-later contrast is $+0.036$ for detection ($p = 0.0004$ after Holm correction) and $+0.018$ for prediction ($p = 0.1605$), with statistical evidence only for detection. This indicates that the initial target supervision stage provides most of the measurable recovery for detection, while prediction benefits more modestly and less consistently across supervision levels. In the evaluated cross-modality settings, target iEEG supervision narrows the gap to direct iEEG training for detection, while prediction retains a larger residual gap. Further direct iEEG results are reported in Appendix F.4.

Spatial evidence for the SOZ. Figure 6 summarizes native contact SOZ evidence, SOZ enrichment at the patient level, and aggregated rankings across eight SOZ metrics. SOZ-based rankings differ from the model ordering by zero-shot cross-modality detection AUROC. RIVER achieves the best aggregated SOZ rank at both 12s and 60s, with average ranks of 1.25 and 1.00, respectively. Foundation models rank second at 12s (2.25), while the transfer-specific comparator group ranks second at 60s (2.00). Further SOZ metrics are provided in Appendix H.

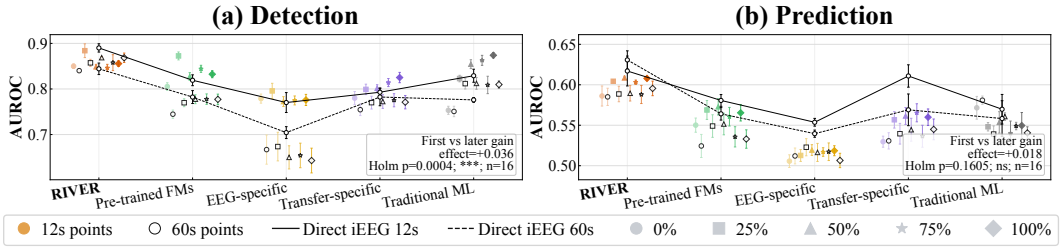

Figure 5: Cross-modality AUROC across target supervision levels for 12s and 60s: (a) detection and (b) prediction. "First" is the 0–25% gain; "later" is the mean of the 25–50%, 50–75%, and 75–100% gains. Effect is First minus later. Holm-corrected p-values use $n = 16$ model blocks. Significance: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, and ns.

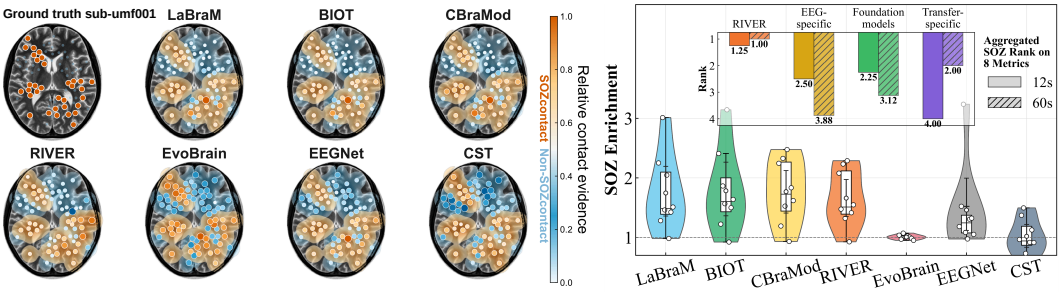

Figure 6: SOZ evidence for 12s zero-shot cross-modality detection. Left: clinically annotated contacts and native contact evidence maps for patient sub-umf001. Right: SOZ enrichment across $n = 8$ patients with bootstrap 95% confidence intervals; the dashed line at one denotes chance enrichment. Inset: aggregated rank over eight SOZ evidence metrics (SOZ enrichment, SOZ AUROC, SOZ average precision, Recall@K, and Hit@1, Hit@3, Hit@5, and Hit@10) under 12s and 60s window lengths. Detailed SOZ metrics are reported in Tables H.2 and H.3.

4.4 How do target supervision levels affect adaptation and calibration after transfer?

Figure 7 summarizes cross-dataset performance and calibration across target supervision levels. The 0–25% interval yields the dominant gain for both tasks (+0.052 AUROC for detection and +0.134 F1 for prediction), while later increments are near zero or negative; the matched first versus later effects are +0.051 and +0.136, respectively ($p = 0.0002$ for both after Holm correction). ECE likewise drops sharply by 25% and then largely plateaus, indicating that early target supervision improves both discrimination and calibration.

Cross-time transfer shows no comparable performance recovery: detection changes only marginally (+0.005, approximately zero, −0.001, and 0.000), while prediction alternates between small losses and gains (−0.005, −0.003, −0.007, and +0.005), with first versus later effects of +0.005 and −0.003 ($p = 1.0$ for both). However, calibration differs: detection ECE remains essentially stable, whereas prediction ECE drops sharply by 25% and then plateaus, showing that calibration can improve without corresponding F1 recovery.

4.5 How do window length and calibration affect model evaluation?

Figure 8 summarizes the relationship between discrimination and calibration across transfer settings and how calibration evolves across increasing target supervision levels. The Spearman association between AUROC and ECE changes both in strength and sign, from −0.65 for cross-modality detection to +0.19 for cross-modality prediction, while ECE follows distinct trajectories across supervision levels. Improvements in discrimination therefore cannot be assumed to imply better calibration, and the effect of adaptation on calibration depends on the task.

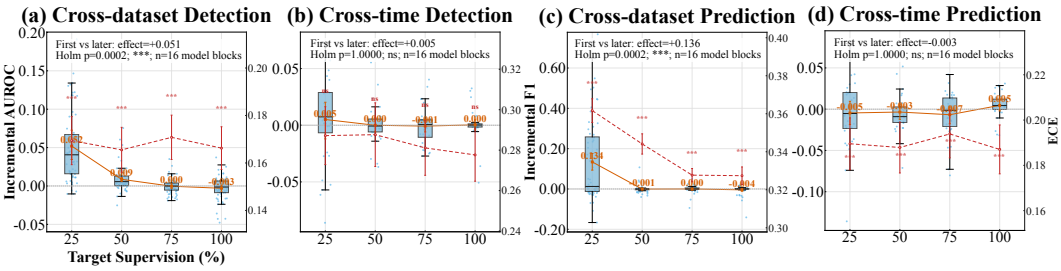

Figure 7: Marginal performance gains and calibration across target supervision levels for cross-dataset and cross-time transfer. Light-blue boxplots/points show incremental AUROC for detection (a,b) and incremental F1 for prediction (c,d); orange markers/lines show aggregate gain estimates. Red dashed curves show ECE aggregated over 12s and 60s on the right axis (lower is better; error bars show SEM), with 0% as the unplotted baseline. "First" is the 0–25% gain and "later" is the mean of the 25–50%, 50–75%, and 75–100% gains. ECE significance uses paired two-sided Wilcoxon tests versus 0%, matched by window length, direction, model, and seed, with Holm correction over four budgets. Significance: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, and ns.

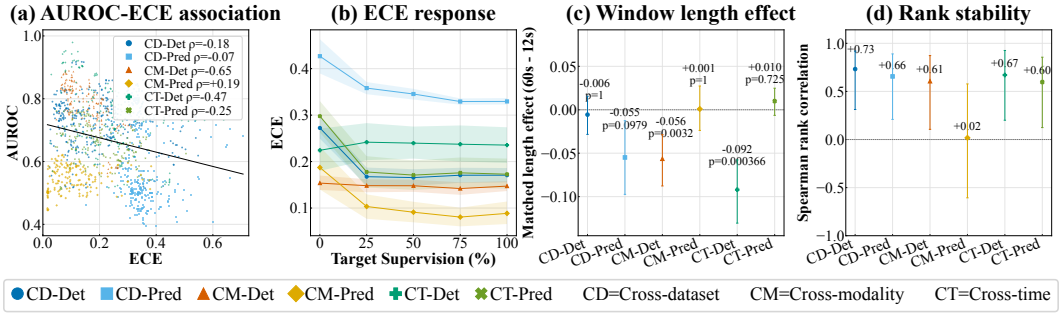

Figure 8: Calibration and window length sensitivity. (a) AUROC–ECE association. (b) ECE across target supervision levels. (c) Matched 60s minus 12s window length effect with bootstrap 95% confidence intervals in the zero-shot setting. (d) Spearman rank correlation between the two window lengths. Window length effects use AUROC for detection and F1 for prediction.

Increasing window length from 12s to 60s produces little mean change for cross-dataset detection (-0.006), cross-modality prediction ($+0.001$), and cross-time prediction ($+0.010$), while the effect is negative for cross-dataset prediction (-0.055), cross-modality detection (-0.056), and cross-time detection (-0.092). A small mean effect can still coincide with unstable model rankings: cross-modality prediction changes by only $+0.001$ on average and has a near-zero Spearman rank correlation across window lengths ($\rho = +0.02$), with the largest declines occurring in cross-modality and cross-time detection. Longer window lengths are therefore not uniformly beneficial.

5 Discussion

Across our benchmark, transferability depends on transfer setting rather than model family alone. Under zero-shot cross-dataset transfer, RIVER achieves the best aggregate AUROC and rank, while foundation models form the strongest baseline, although differences from foundation and transfer-specific comparators are not statistically significant after Holm correction. This extends evidence on distribution shift (Koh et al., 2021; Taori et al., 2020) and EEG comparisons showing no uniform foundation model dominance (Yang et al., 2026) to dataset, modality, and longitudinal shifts.

This dependence is particularly evident under cross-modality transfer. In the evaluated settings, detection shows stronger recovery with target iEEG supervision and can approach the direct iEEG reference, whereas prediction retains a larger gap. Nevertheless, the pattern is compatible with pre-ictal discrimination being more sensitive to spatial coverage and patient context, while electrode and feature-space mismatch are established barriers in heterogeneous electrophysiology (Wang et al.,

Table 1: Complementary recovery from the 0% to 25% target supervision levels. Each entry reports cross-dataset / cross-modality / cross-time recovery, averaged over the 12s and 60s window lengths and the available directions within each transfer and task setting. Positive values denote improvement: AUPRC uses $\mathrm{AUPRC}_{25} - \mathrm{AUPRC}_0$, whereas ECE and Brier use $\mathrm{Metric}_0 - \mathrm{Metric}_{25}$.

Detection			
Model group	ΔAUPRC ↑	ECE rec. ↑	Brier rec. ↑
Foundation models	+0.060 / +0.013 / +0.021	+0.103 / +0.027 / -0.025	+0.081 / +0.012 / -0.010
EEG-specific	+0.028 / 0.000 / +0.004	+0.067 / +0.009 / 0.000	+0.051 / -0.001 / +0.008
Transfer-specific	+0.034 / +0.010 / -0.004	+0.172 / -0.004 / -0.025	+0.109 / 0.000 / -0.016
Traditional ML	+0.100 / +0.052 / -0.001	+0.115 / -0.077 / -0.023	+0.077 / +0.013 / -0.015
RIVER	+0.054 / -0.005 / +0.005	+0.046 / +0.086 / -0.011	+0.042 / +0.041 / -0.001
Prediction			
Model group	ΔAUPRC ↑	ECE rec. ↑	Brier rec. ↑
Foundation models	+0.007 / +0.013 / +0.069	+0.131 / +0.107 / +0.125	+0.133 / +0.035 / +0.074
EEG-specific	+0.010 / +0.007 / +0.023	+0.027 / +0.039 / +0.133	+0.038 / +0.012 / +0.076
Transfer-specific	+0.026 / -0.001 / +0.075	+0.079 / +0.056 / +0.109	+0.087 / +0.030 / +0.075
Traditional ML	+0.043 / -0.047 / +0.108	+0.012 / +0.062 / +0.107	+0.015 / +0.016 / +0.062
RIVER	+0.015 / +0.017 / +0.027	-0.056 / +0.148 / +0.093	-0.030 / +0.045 / +0.053

2025b; Kuhlmann et al., 2018; Li et al., 2021). For detection, the RIVER ablation study and electrode scarcity analysis further suggest that modeling incomplete electrode observability contributes to robustness under heterogeneous sensing. Detailed analyses are provided in Appendix D.

Variation across target supervision levels further separates transferability from recoverability. Both cross-dataset tasks and cross-modality detection concentrate performance gains in the 0–25% interval, whereas cross-modality prediction shows a weaker contrast and cross-time performance shows no comparable pattern. Calibration follows a distinct trajectory: ECE improves early for cross-dataset tasks and cross-time prediction but remains unchanged for cross-time detection. Thus, limited target exposure can affect discrimination and probability quality differently under shift, consistent with the sensitivity of fine-tuning (Kumar et al., 2022). Table 1 likewise shows metric-dependent recovery, supporting evaluation across multiple dimensions of model performance and reliability.

Window length, calibration, and SOZ evidence further show why discrimination alone is insufficient. Longer window lengths can reorder models without improving average performance across transfer settings, consistent with prior evidence that EEG window length can materially alter seizure-detection performance (Reisinger et al., 2025; Jeong et al., 2026). Discrimination and calibration are likewise not consistently aligned under transfer (Guo et al., 2017; Ovadia et al., 2019; Minderer et al., 2021), and SOZ evidence does not mirror classification rankings. These results therefore characterize model behavior within the evaluated transfer settings rather than universal properties of epilepsy models. From a practical perspective, epilepsy models should be evaluated as a multidimensional profile spanning transfer setting, target supervision level, window length, discrimination, calibration, and clinically relevant evidence rather than as a single aggregate score.

6 Conclusion

We present EpiTAR-Bench, a unified benchmark for transferability, adaptation, and reliability of epilepsy models under cross-dataset, cross-modality, and cross-time transfer, together with RIVER. Across the 16 evaluated models, transferability and adaptation depend on transfer direction, task, window length, and available target supervision level rather than on model family alone. The initial 0–25% target supervision interval captures much of the recoverable gain under cross-dataset transfer and cross-modality detection; this pattern is weaker for cross-modality prediction, and cross-time adaptation is less consistent. Longer window lengths, however, do not guarantee better performance. Scalp EEG transfer provides useful initialization for iEEG detection, whereas prediction retains a larger gap in the evaluated cross-modality setting. Discrimination, calibration, and SOZ evidence also need not align. Overall, EpiTAR-Bench provides a unified view of how epilepsy models transfer, adapt, and remain reliable across heterogeneous datasets, modalities, and recording periods.

AI USE STATEMENT

In this work, we used generative AI tools to polish the writing and improve the clarity and fluency of the manuscript. We also used generative AI tools for limited auxiliary coding tasks, such as generating simple scripts and instructions to help identify and debug implementation issues. All AI-assisted code was manually reviewed and verified by the authors before use. We did not use generative AI tools for research design, experimental design, data analysis, scientific interpretation, or scientific conclusions, and generative AI tools were not used to autonomously determine, modify, or interpret experimental results. We reviewed all AI-assisted work to ensure that it preserved the intended scientific meaning and did not introduce unsupported claims or alter the reported results. We take full responsibility for the final content of this work.

ETHICS STATEMENT

This study uses previously released and de-identified EEG and iEEG datasets and does not involve the collection of new human participant data. We follow the access conditions, data use requirements, and ethics and consent information provided with the original dataset releases. All analyses are conducted retrospectively for research purposes. The authors take responsibility for the responsible use, analysis, and reporting of the datasets included in this work.

REPRODUCIBILITY STATEMENT

We have taken several steps to support reproducibility. The benchmark uses fixed preprocessing procedures, explicit source and target isolation, deterministic construction of target supervision fine-tuning subsets, and fixed target test cohorts. Dataset construction and preprocessing are documented in Appendix A. Model initialization, optimization, and input adaptation are described in Appendix B. The formulation and implementation details of RIVER are provided in Appendix C. Ablation and mechanism analyses are provided in Appendix D. Statistical procedures are specified in Appendix E. Complete transfer summaries are provided in Appendix F. Additional analyses of window length sensitivity, calibration, and SOZ evidence are included in Appendix H. Code is provided anonymously within the supplementary material.

REFERENCES

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Sy21R9JAW.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

Yoon Gi Chung, Yonghoon Jeon, Sun Ah Choi, Anna Cho, Hunmin Kim, Hee Hwang, and Ki Joong Kim. Deep convolutional neural network based interictal-preictal electroencephalography prediction: Application to focal cortical dysplasia type-II. *Frontiers in Neurology*, 11:594679, 2020. doi: 10.3389/fneur.2020.594679.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995. doi: 10.1007/BF00994018. URL https://doi.org/10.1007/BF00994018.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2017. doi: 10.1109/TPAMI.2016.2615921.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, pp. 2292–2300, 2013. URL https://proceedings.neurips.cc/paper/2013/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html.

Paolo Detti. Siena scalp EEG database. *PhysioNet*, 2020. doi: 10.13026/5d4a-j060. URL https://doi.org/10.13026/5d4a-j060. Version 1.0.0.

Paolo Detti, Giampaolo Vatti, and Garazi Zabalo Manrique de Lara. EEG synchronization analysis for seizure prediction: A study on data of noninvasive recordings. *Processes*, 8(7):846, 2020. doi: 10.3390/pr8070846. URL https://doi.org/10.3390/pr8070846.

Nadine El-Dajani, Tim Friedrich Lutz Wilhelm, Jan Baumann, Rainer Surges, and Bernd T. Meyer. Patient-independent epileptic seizure detection with reduced EEG channels and deep recurrent neural networks. *Information*, 16(1):20, 2025. doi: 10.3390/info16010020.

Kilian Fatras, Thibault Séjourné, Rémi Flamary, and Nicolas Courty. Unbalanced minibatch optimal transport; applications to domain adaptation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3186–3197. PMLR, 2021. URL https://proceedings.mlr.press/v139/fatras21a.html.

Théo Gnassounou, Rémi Flamary, and Alexandre Gramfort. Convolution Monge mapping normalization for learning on sleep data. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0459. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/21718991f6acf19a42376b5c7a8668c5-Abstract-Conference.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

John Guttag. CHB-MIT scalp EEG database. *PhysioNet*, 2010. doi: 10.13026/C2K01R. URL https://doi.org/10.13026/C2K01R. Version 1.0.0.

Robert Hogan, Sean R. Mathieson, Aurel Luca, Soraia Ventura, Sean Griffin, Geraldine B. Boylan, and John M. O'Toole. Scaling convolutional neural networks achieves expert level seizure detection in neonatal EEG. *npj Digital Medicine*, 8(1):17, 2025. doi: 10.1038/s41746-024-01416-x.

Hyewon Jeong, Kwanhyung Lee, Seyun Kim, Hoon-Chul Kang, and Donghwa Yang. Deep learning-based real-time seizure detection and multi-seizure classification on pediatric EEG. *Frontiers in Neurology*, 17:1726258, 2026. doi: 10.3389/fneur.2026.1726258.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=QzTpTRVtrP.

Ximiao Jiang, Xiaotong Liu, Youjun Liu, Qingyun Wang, Bao Li, and Liyuan Zhang. Epileptic seizures detection and the analysis of optimal seizure prediction horizon based on frequency and phase analysis. *Frontiers in Neuroscience*, 17:1191683, 2023. doi: 10.3389/fnins.2023.1191683.

Reinmar J. Kobler, Jun-ichiro Hirayama, Qibin Zhao, and Motoaki Kawanabe. SPD domain-specific batch normalization to crack interpretable unsupervised domain adaptation in EEG. In *Advances in Neural Information Processing Systems*, volume 35, pp. 6219–6235, 2022. doi: 10.52202/068431-0450. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/28ef7ee7cd3e03093acc39e1272411b7-Abstract-Conference.html.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664. PMLR, 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Aditya Kommineni, Emily Zhou, Kleanthis Avramidis, Tiantian Feng, and Shrikanth Narayanan. A multi-dimensional framework for evaluating generalization in EEG foundation models. In Elizabeth Healey, Jason Fries, Tom Pollard, Shengpu Tang, Anna Zink, Tom Hartvigsen, Monica Agrawal, Sam Finlayson, Benjamin Glicksberg, Brett Beaulieu-Jones, Kai Wang, Daseyra

Fontalvo, Tasmie Sarker, Irene Chen, and Emily Alsentzer (eds.), *Proceedings of the 7th Conference on Health, Inference, and Learning*, volume 333 of *Proceedings of Machine Learning Research*, pp. 221–244. PMLR, 29–30 Jun 2026. URL https://proceedings.mlr.press/v333/kommineni26a.html.

Demetres Kostas, Stéphane Aroca-Ouellette, and Frank Rudzicz. BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data. *Frontiers in Human Neuroscience*, 15:653659, 2021. doi: 10.3389/fnhum.2021.653659. URL https://doi.org/10.3389/fnhum.2021.653659.

Rikuto Kotoge, Zheng Chen, Tasuku Kimura, Yasuko Matsubara, Takufumi Yanagisawa, Haruhiko Kishima, and Yasushi Sakurai. EvoBrain: Dynamic multi-channel EEG graph modeling for time-evolving brain networks. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4837. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/d51ce6040fe3dafd260411593f05a1fa-Abstract-Conference.html.

Levin Kuhlmann, Klaus Lehnertz, Mark P. Richardson, Björn Schelter, and Hitten P. Zaveri. Seizure prediction—ready for a new era. *Nature Reviews Neurology*, 14(10):618–630, 2018. doi: 10.1038/s41582-018-0055-2.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=UYneFzXSJWh.

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. URL https://doi.org/10.1088/1741-2552/aace8c.

Adam Li, Sara Inati, Kareem Zaghloul, Nathan Crone, William Anderson, Emily Johnson, Iahn Cajigas, Damian Brusko, Jonathan Jagid, Angel Claudio, Andres Kanner, Jennifer Hopp, Stephanie Chen, Jennifer Haagensen, and Sridevi Sarma. Epilepsy-iEEG-multicenter-dataset. NEMAR, 2026. URL https://doi.org/10.82901/nemar.on003029. Version v1.0.0; derived from OpenNeuro ds003029 v1.0.5.

Zan Li, Madeline Fields, Fedor Panov, Saadi Ghatan, Bülent Yener, and Lara Marcuse. Deep learning of simultaneous intracranial and scalp EEG for prediction, detection, and lateralization of mesial temporal lobe seizures. *Frontiers in Neurology*, 12:705119, 2021. doi: 10.3389/fneur.2021.705119.

Fábio Lopes, Mauro F. Pinto, António Dourado, Andreas Schulze-Bonhage, Matthias Dümpelmann, and César Teixeira. Addressing data limitations in seizure prediction through transfer learning. *Scientific Reports*, 14:14169, 2024. doi: 10.1038/s41598-024-64802-1. URL https://doi.org/10.1038/s41598-024-64802-1.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 15682–15694, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/8420d359404024567b5aefda1231af24-Abstract.html.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, pp. 13991–14002, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html.

Saarang Panchavati, Atsuro Daida, Sotaro Kanai, Shingo Oana, Hiroya Ono, Masaki Izumi, Kikuko Kaneko, Aria Fallah, Joe X. Qiao, Noriko Salamon, Raman Sankar, Corey Arnold, William Speier, and Hiroki Nariai. Thalamocortical ictal iEEG dataset. OpenNeuro, 2026. URL

https://doi.org/10.18112/openneuro.ds007445.v1.0.2. Dataset ds007445, Version v1.0.2.

Wasi Ur Rehman Qamar, Min-Ho Lee, and Berdakh Abibullaev. Deep learning in intracranial EEG for seizure detection: advances, challenges, and clinical applications. *Frontiers in Neuroscience*, 19:1677898, 2025. doi: 10.3389/fnins.2025.1677898.

Sriram Ramgopal, Sigride Thome-Souza, Michele Jackson, Navah Ester Kadish, Iván Sánchez Fernández, Jacquelyn Klehm, William Bosl, Claus Reinsberger, Steven Schachter, and Tobias Loddenkemper. Seizure detection, seizure prediction, and closed-loop warning systems in epilepsy. *Epilepsy & Behavior*, 37:291–307, 2014. doi: 10.1016/j.yebeh.2014.06.023. URL https://doi.org/10.1016/j.yebeh.2014.06.023.

Patrick Reisinger, Jonathan Larochelle, Ciamak Abkai, Sotirios Kalousios, Nico Zabler, Matthias Dümpelmann, Andreas Schulze-Bonhage, Peter Woias, and Laura Comella. The impact of EEG preprocessing parameters on ultra-low-power seizure detection. *Epilepsia*, 66(10):3895–3906, 2025. doi: 10.1111/epi.18521.

Khaled Saab, Jared Dunnmon, Christopher Ré, Daniel L. Rubin, and Christopher Lee-Messer. Weak supervision as an efficient approach for automated seizure detection in electroencephalography. *npj Digital Medicine*, 3:59, 2020. doi: 10.1038/s41746-020-0264-0.

Vinit Shah, Eva von Weltin, Silvia Lopez, James Riley McHugh, Lillian Veloso, Meysam Golmohammadi, Iyad Obeid, and Joseph Picone. The Temple University Hospital seizure detection corpus. *Frontiers in Neuroinformatics*, 12:83, 2018. doi: 10.3389/fninf.2018.00083. URL https://doi.org/10.3389/fninf.2018.00083.

Ali Hossam Shoeb. *Application of Machine Learning to Epileptic Seizure Onset Detection and Treatment*. PhD thesis, Massachusetts Institute of Technology, September 2009. URL http://hdl.handle.net/1721.1/54669.

Rosa Q. So, Vibhor Krishna, Nicolas Kon Kam King, Huijuan Yang, Zhuo Zhang, Francesco Sammartino, Andres M. Lozano, Richard A. Wennberg, and Cuntai Guan. Prediction and detection of seizures from simultaneous thalamic and scalp electroencephalography recordings. *Journal of Neurosurgery*, 126(6):2036–2044, 2017. doi: 10.3171/2016.7.JNS161282.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d8330f857a17c53d217014ee776bfd50-Abstract.html.

Nhan Duy Truong, Anh Duy Nguyen, Levin Kuhlmann, Mohammad Reza Bonyadi, Jiawei Yang, Samuel Ippolito, and Omid Kavehei. Convolutional neural networks for seizure prediction using intracranial and scalp electroencephalogram. *Neural Networks*, 105:104–111, 2018. doi: 10.1016/j.neunet.2018.04.018.

Prabhat Kumar Upadhyay, Priyaranjan Kumar, Manoj Kumar Panda, and Aswini Kumar Samantaray. Seamless integration for enhanced seizure prediction using HybridConvMobileNet on typhoon HIL. *Scientific Reports*, 15(1):45480, 2025. doi: 10.1038/s41598-025-29393-5.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. EEGPT: Pretrained transformer for universal and reliable representation of EEG signals. In *Advances in Neural Information Processing Systems*, volume 37, pp. 39249–39280, 2024. doi: 10.52202/079017-1239. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/4540d267eeec4e5dbd9dae9448f0b739-Abstract-Conference.html.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBraMod: A criss-cross brain foundation model for EEG decoding. In *International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=NPNUHgHF2w.

Xiaoshuang Wang, Guanghui Zhang, Ying Wang, Lin Yang, Zhanhua Liang, and Fengyu Cong. One-dimensional convolutional neural networks combined with channel selection strategy for seizure prediction using long-term intracranial EEG. *International Journal of Neural Systems*, 32 (2):2150048, 2022. doi: 10.1142/S0129065721500489.

Ziwei Wang, Siyang Li, and Dongrui Wu. Canine EEG helps human: Cross-species and cross-modality epileptic seizure detection via multi-space alignment. *National Science Review*, 12 (6):nwaf086, 2025b. doi: 10.1093/nsr/nwaf086. URL https://doi.org/10.1093/nsr/nwaf086.

Ziyu Wang, Jie Yang, Hemmings Wu, Junming Zhu, and Mohamad Sawan. Power efficient refined seizure prediction algorithm based on an enhanced benchmarking. *Scientific Reports*, 11(1): 23498, 2021. doi: 10.1038/s41598-021-02798-8.

Ziqi Wen, Cai Xu, Wanqing Zhao, Jie Zhao, and Wei Zhao. Universal EEG epilepsy detection via evidential multi-view de-biasing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 26849–26857, 2026. doi: 10.1609/aaai.v40i32.39896. URL https://ojs.aaai.org/index.php/AAAI/article/view/39896.

World Health Organization. Epilepsy. WHO Fact Sheet, February 2024. URL https://www.who.int/news-room/fact-sheets/detail/epilepsy. 7 February 2024.

Jie Xiang, Yanan Li, Xubin Wu, Yanqing Dong, Xin Wen, and Yan Niu. Synchronization-based graph spatio-temporal attention network for seizure prediction. *Scientific Reports*, 15(1):4080, 2025. doi: 10.1038/s41598-025-88492-5.

Chaoqi Yang, M. Brandon Westover, and Jimeng Sun. BIOT: Biosignal transformer for cross-data learning in the wild. In *Advances in Neural Information Processing Systems*, volume 36, pp. 78240–78260, 2023. doi: 10.52202/075280-3420. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f6b30f3e2dd9cb53bbf2024402d02295-Abstract-Conference.html.

Liuyin Yang, Qiang Sun, Ang Li, and Marc Van Hulle. Are EEG foundation models worth it? Comparative evaluation with traditional decoders in diverse BCI tasks. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/f0f39b7686634fc81ca0112566b2c05f-Abstract-Conference.html.

Ruizhe Zheng, Jun Li, Yi Wang, Tian Luo, and Yuguo Yu. ScatterFormer: Locally-invariant scattering transformer for patient-independent multispectral detection of epileptiform discharges. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 148–158, 2023. doi: 10.1609/aaai.v37i1.25086. URL https://ojs.aaai.org/index.php/AAAI/article/view/25086.

APPENDIX

A BENCHMARK DATA AND TRANSFER PROTOCOLS

This section shows the data contract used throughout our benchmark. CD, CM, and CT denote cross-dataset, cross-modality, and cross-time transfer, respectively. Tables A.1 and A.2 summarize the retained cohorts and the effective samples available after task construction.

Table A.1: Benchmark datasets and retained cohorts.

Dataset release	Modality	Retained cohort	Key recording property
TUSZ v2.0.6 (Shah et al., 2018)	scalp EEG	228 patients	clinical EEG with seizure annotations
CHB-MIT v1.0.0 (Guttag, 2010; Shoeb, 2009)	scalp EEG	22 patients	prolonged pediatric scalp EEG
Siena v1.0.0 (Detti, 2020; Detti et al., 2020)	scalp EEG	14 patients	47 seizures; 512Hz; 10–20 system
Epilepsy-iEEG ds003029 v1.0.5 (Li et al., 2026)	iEEG	30 patients	multicenter ECoG/SEEG
Thalamocortical-iEEG ds007445 v1.0.2 (Panchavati et al., 2026)	iEEG	19 patients	66 recordings with thalamic sampling

Table A.2: Retained benchmark cohorts for 12s and 60s window lengths. Clips are total (positive/negative), and channels are minimum/median/maximum after preprocessing.

Track	Task	Dataset	Patients	12s clips (+/-)	60s clips (+/-)	Channels	Band (Hz)
CD	Det.	CHB-MIT	22	53,781 (1,074/52,707)	10,747 (351/10,396)	16/16/16	0.5–45
CD	Det.	TUSZ	228	67,061 (15,913/51,148)	13,145 (4,460/8,685)	16/16/16	0.5–45
CD	Pred.	CHB-MIT	22	84,764 (21,087/63,677)	16,924 (4,198/12,726)	16/16/16	0.5–45
CD	Pred.	Siena	14	39,682 (6,415/33,267)	7,902 (1,280/6,622)	16/16/16	0.5–45
CM	Det.	TUSZ	228	67,063 (15,913/51,150)	13,145 (4,460/8,685)	16/16/16	0.5–95
CM	Det.	Epilepsy-iEEG	30	2,177 (1,101/1,076)	397 (290/107)	30/71/136	0.5–95
CM	Pred.	CHB-MIT	22	84,764 (21,087/63,677)	16,924 (4,198/12,726)	16/16/16	0.5–95
CM	Pred.	Thalamocortical-iEEG	19	18,252 (8,971/9,281)	3,598 (1,774/1,824)	41/112/216	0.5–95
CT	Det.	CHB-MIT Hist.	16	36,491 (695/35,796)	7,294 (235/7,059)	16/16/16	0.5–45
CT	Det.	CHB-MIT Fut.	16	29,424 (610/28,814)	5,887 (206/5,681)	16/16/16	0.5–45
CT	Pred.	CHB-MIT Hist.	16	53,760 (14,410/39,350)	10,752 (2,888/7,864)	16/16/16	0.5–45
CT	Pred.	CHB-MIT Fut.	16	60,715 (12,349/48,366)	12,066 (2,397/9,669)	16/16/16	0.5–45

The two TUSZ detection entries differ by two 12s non-ictal clips because CD and CM use separate preprocessing caches under their 0.5–45Hz and 0.5–95Hz signal contracts. The source cohort and positive seizure events are unchanged, and this small difference arises from track-specific cache construction and quality control.

A.1 SIGNAL PREPROCESSING

All tracks share the same preprocessing order: channel quality control, dataset-specific spectral fitting for line-noise removal, bandpass filtering, antialiased resampling, normalization, and benchmark slicing. Filtering is applied to continuous recordings before clip construction. Only the track-specific signal contract changes, as summarized in Table A.3.

Table A.3: Signal processing contract.

Track	Bandpass	Sampling	Normalization	Channel contract
CD	0.5–45Hz	256Hz	per-clip/channel z-score	16 scalp bipolar derivations
CM	0.5–95Hz	256Hz	per-clip/channel z-score	16 scalp source; native iEEG target
CT	0.5–45Hz	256Hz	per-clip/channel z-score	16 scalp bipolar derivations

For better signal quality, we require channel standard deviation of at least 0.1μV, no more than 20% flat channels, and no continuously bad-channel segment longer than 120s.

A.2 TASK LABELS

Detection treats ictal activity as positive and non-ictal activity as negative. Complete seizure events are retained, while non-ictal training clips are dynamically sampled up to a benchmark-specific 1:2

positive/negative ratio. Related undersampling and class-balancing strategies have been used in prior seizure studies (Saab et al., 2020; Hogan et al., 2025; El-Dajani et al., 2025). The 12s and 60s variants share the same patient and event split before clips are formed.

For prediction, preictal activity is treated as positive and interictal activity as negative. Across all prediction datasets, the final 5min before seizure onset is excluded, and the preceding 30min interval (35–5min before onset) is labeled preictal. Interictal segments are constructed according to the recording structure of each dataset: CHB-MIT uses periods from 4h after the previous seizure offset to 1h before the next onset; Siena uses within-session periods from 5min after the previous seizure offset to 35min before the next onset; and Thalamocortical-iEEG uses eligible within-component intervals ending 35min before an onset and beginning 5min after a preceding seizure offset, with component boundaries used at the beginning and end of each recording component. These timing definitions are consistent with prior seizure-prediction studies (Wang et al., 2021; Jiang et al., 2023; Xiang et al., 2025; Upadhyay et al., 2025; So et al., 2017; Truong et al., 2018; Wang et al., 2022).

Prediction uses dynamic 1:1 preictal/interictal sampling during source training (Chung et al., 2020; Xiang et al., 2025). During cross-dataset and cross-modality target adaptation, all preictal clips from selected target patients are retained and interictal clips are sampled at the same ratio with patient balancing, while cross-time adaptation follows the chronological adapt prefixes defined below.

A.3 Transfer directions and split construction

The benchmark contains seven ordered source-to-target directions, listed in Table A.4.

Table A.4: Ordered benchmark directions.

ID	Transfer	Task	Source	Target
D1	CD	Detection	TUSZ	CHB-MIT
D2	CD	Detection	CHB-MIT	TUSZ
D3	CD	Prediction	CHB-MIT	Siena
D4	CD	Prediction	Siena	CHB-MIT
D5	CM	Detection	TUSZ	Epilepsy-iEEG
D6	CM	Prediction	CHB-MIT	Thalamocortical-iEEG
D7	CT	Detection + Prediction	historical CHB-MIT	future CHB-MIT, same patients

TUSZ follows the official corpus partition (Shah et al., 2018), using the development and evaluation portions as validation and test sets. The remaining CD and CM datasets use fixed patient-level train/validation/test partitions summarized in Tables A.6 and A.7. For CT, positive clips are grouped by `patient_id::event_id` and events are ordered by their earliest `timeline_clip_start_seconds`. Patients with fewer than four positive events are excluded, and the latest 30% (`FUTURE_EVENT_FRACTION`=0.30) form the fixed future block, while the earlier events are partitioned into historical, dev (20% of the earlier block), and adapt (30% of the earlier block). The first dev, adapt, and future event starts define the temporal boundaries. Negative clips are assigned by clip start time and any clip crossing these boundaries is excluded. `CHBMIT-Historical` maps historical/dev/future to train/validation/test with adapt unused, whereas `CHBMIT-Future` maps adapt/dev/future to train/validation/test with historical unused. Target supervision budgets use chronological prefixes of positive events in adapt, with dev and future test fixed.

A.4 Target supervision fine-tuning level construction

For cross-dataset and cross-modality transfer, target supervision is defined at the patient level. For $p \in \{25, 50, 75, 100\}$, the selected target training patient count is

$$K_p = \max\Big(1, \Big\lfloor N_{\text{patient}} \frac{p}{100} + 0.5 \Big\rfloor\Big). \tag{1}$$

For zero-shot evaluation, $K_0 = 0$. The patient subsets are nested, so larger fine-tuning levels extend the smaller subset. Selected patients contribute all positive events and clips for detection or prediction, with dynamic 1:2 and 1:1 negative sampling (Hogan et al., 2025; El-Dajani et al., 2025; Chung et al., 2020; Xiang et al., 2025), respectively. For CD/CM, the percentage p is applied separately to

the target training and target validation patient pools: for example, the 25% level uses 25% of target training patients for adaptation and 25% of target validation patients for checkpoint and threshold selection. Target-test labels are excluded. Cross-time transfer instead uses chronological prefixes of positive events in the within-patient adapt block defined above. The resulting cross-modality patient quotas are shown in Table A.5.

Table A.5: Patient-level quotas for cross-modality target adaptation.

Target dataset	Total train patients	25%	50%	75%	100%
Epilepsy-iEEG	16	4	8	12	16
Thalamocortical-iEEG	11	3	6	8	11

A.5 Split statistics

Tables A.6 and A.7 expose the effective train, validation, and test partitions after task construction. The 12s and 60s statistics share patient splits within each row, making changes across window lengths in recordings, events, and clip counts directly visible.

Table A.6: Cross-dataset split statistics.

Task	Dataset	Split	Patients	Recordings (12/60)	Pos. events (12/60)	Clips 12s / 60s
Det.	CHB-MIT	train	14	78/78	114/114	36,713 (690/36,023) / 7,336 (220/7,116)
Det.	CHB-MIT	val	4	26/26	36/36	7,782 (246/7,536) / 1,555 (77/1,478)
Det.	CHB-MIT	test	4	25/25	34/34	9,286 (138/9,148) / 1,856 (54/1,802)
Det.	TUSZ	train	161	682/680	1,721/1,456	38,031 (10,895/27,136) / 7,426 (2,940/4,486)
Det.	TUSZ	val	37	217/214	492/475	18,286 (2,678/15,608) / 3,609 (850/2,759)
Det.	TUSZ	test	30	165/164	360/324	10,744 (2,340/8,404) / 2,110 (670/1,440)
Pred.	CHB-MIT	train	14	185/183	100/98	74,188 (12,446/61,742) / 14,820 (2,478/12,342)
Pred.	CHB-MIT	val	4	41/41	36/36	5,205 (4,267/938) / 1,034 (848/186)
Pred.	CHB-MIT	test	4	36/36	32/32	5,371 (4,374/997) / 1,070 (872/198)
Pred.	Siena	train	8	26/26	30/30	24,916 (4,042/20,874) / 4,961 (807/4,154)
Pred.	Siena	val	3	7/7	9/9	7,435 (1,173/6,262) / 1,481 (233/1,248)
Pred.	Siena	test	3	8/8	8/8	7,331 (1,200/6,131) / 1,460 (240/1,220)

Cross-dataset patient partitions remain fixed across window lengths. The 60s setting nevertheless contains fewer valid clips because the same recordings and event boundaries support fewer non-overlapping windows.

Table A.7: Cross-modality split statistics.

Task	Dataset	Split	Patients	Recordings (12/60)	Pos. events (12/60)	Clips 12s / 60s
Det.	TUSZ	train	161	682/680	1,721/1,456	38,032 (10,895/27,137) / 7,426 (2,940/4,486)
Det.	TUSZ	val	37	217/214	492/475	18,287 (2,678/15,609) / 3,609 (850/2,759)
Det.	TUSZ	test	30	165/164	360/324	10,744 (2,340/8,404) / 2,110 (670/1,440)
Det.	Epilepsy-iEEG	train	16	55/55	55/55	1,285 (698/587) / 234 (182/52)
Det.	Epilepsy-iEEG	val	6	19/19	19/19	428 (193/235) / 77 (54/23)
Det.	Epilepsy-iEEG	test	8	18/18	18/18	464 (210/254) / 86 (54/32)
Pred.	CHB-MIT	train	14	185/183	100/98	74,188 (12,446/61,742) / 14,820 (2,478/12,342)
Pred.	CHB-MIT	val	4	41/41	36/36	5,205 (4,267/938) / 1,034 (848/186)
Pred.	CHB-MIT	test	4	36/36	32/32	5,371 (4,374/997) / 1,070 (872/198)
Pred.	Thalamocortical-iEEG	train	11	38/38	38/38	10,329 (5,075/5,254) / 2,036 (1,005/1,031)
Pred.	Thalamocortical-iEEG	val	4	15/15	15/15	4,113 (2,055/2,058) / 808 (404/404)
Pred.	Thalamocortical-iEEG	test	4	13/13	13/13	3,810 (1,841/1,969) / 754 (365/389)

Cross-modality target cohorts are fixed across all fine-tuning levels: Epilepsy-iEEG uses 16/6/8 train/validation/test patients for detection, and Thalamocortical-iEEG uses 11/4/4 for prediction. At each positive fine-tuning level, the same percentage is applied separately within the target training and target validation cohorts.

B Benchmark Model Implementation Details

This section specifies the implementation choices needed to place heterogeneous models under one benchmark protocol. Shared schedules are stated once. Differences among models are limited to initialization, optimization, input interfaces, and fixed feature baselines.

B.1 WEIGHT SOURCES AND OPTIMIZATION SETTINGS

Initialization and the learning-rate settings that differ across models are consolidated in Table B.1.

Table B.1: Model initialization and optimization settings. "scratch" denotes benchmark initialization.

Model	Family	Initialization / weight source	Source LR	Target LR	Target backbone LR	Weight decay
BENDR (Kostas et al., 2021)	foundation	pre-trained weights (Kostas et al., 2021)	1×10^{-5}	1×10^{-5}	1×10^{-6}	0.01
BIOT (Yang et al., 2023)	foundation	pre-trained weights (Yang et al., 2023)	1×10^{-3}	1×10^{-4}	1×10^{-5}	1×10^{-5}
CBraMod (Wang et al., 2025a)	foundation	pre-trained weights (Wang et al., 2025a)	1×10^{-4}	1×10^{-5}	1×10^{-6}	0.05
EEGPT (Wang et al., 2024)	foundation	pre-trained weights (Wang et al., 2024)	5×10^{-4}	1×10^{-4}	1×10^{-5}	0.05
LaBraM (Jiang et al., 2024)	foundation	pre-trained weights (Jiang et al., 2024)	5×10^{-4}	1×10^{-4}	1×10^{-5}	0.05
ST-EEGFormer (Yang et al., 2026)	foundation	pre-trained weights (Yang et al., 2026)	5×10^{-4}	1×10^{-4}	1×10^{-5}	0.05
EEGNet (Lawhern et al., 2018)	EEG-specific	scratch	1×10^{-3}	1×10^{-4}	1×10^{-5}	–
EvoBrain (Kotoge et al., 2025)	EEG-specific	scratch	3×10^{-4}	1×10^{-4}	1×10^{-5}	5×10^{-4}
ScatterFormer (Zheng et al., 2023)	EEG-specific	scratch	5×10^{-4}	1×10^{-4}	1×10^{-5}	0.05
BF-EML (Wen et al., 2026)	EEG-specific	scratch	5×10^{-4}	1×10^{-4}	1×10^{-5}	0.05
CST (Wang et al., 2025b)	transfer-specific	scratch	1×10^{-3}	1×10^{-4}	1×10^{-5}	0
CMMN (Gnassounou et al., 2023)	transfer-specific	scratch	–	–	–	–
TSMNet (Kobler et al., 2022)	transfer-specific	scratch	5×10^{-4}	1×10^{-4}	1×10^{-5}	0.05
SVM (Cortes & Vapnik, 1995)	traditional ML	scratch	–	–	–	–
Random Forest (Breiman, 2001)	traditional ML	scratch	–	–	–	–
RIVER	transfer-specific	scratch	3×10^{-4}	1×10^{-4}	1×10^{-5}	0.05

Foundation models use the pre-trained checkpoints associated with the cited public implementations, consistent with recent EEG foundation-model evaluations under downstream transfer and adaptation settings (Yang et al., 2026; Kommineni et al., 2026). Here, "zero-shot" denotes evaluation without target-data training or adaptation. Gradient clipping uses a maximum norm of 1.0 for BENDR, BF-EML, EEGPT, LaBraM, RIVER, ScatterFormer, and TSMNet, and 5.0 for EvoBrain.

B.2 TRAINING AND ADAPTATION PROTOCOL

Source training uses 100 epochs, and source-validation AUROC selects the checkpoint used for zero-shot evaluation. The 0% condition evaluates this checkpoint directly on the target test set without target training labels, and the same checkpoint initializes all positive fine-tuning levels.

At positive levels, neural adaptation uses 5 epochs of linear probing followed by 50 epochs of full fine-tuning. For CD/CM, target validation AUROC selects the adapted checkpoint using the same percentage of the target validation patient pool as the target training percentage. CT retains the fixed dev block defined in Appendix A. Batch size is 128 for 12s inputs and 64 for 60s inputs. Adaptation retains 25% of source data for rehearsal, using a 1:1 class ratio for prediction and 1:2 for detection (Hogan et al., 2025; El-Dajani et al., 2025; Chung et al., 2020; Xiang et al., 2025). SVM, CMMN, and Random Forest are refit on the same target patient fraction (or chronological positive-event adapt prefix for CT) with source rehearsal. Table B.2 collects the shared neural schedule.

Table B.2: Unified neural training protocol.

Window length (s)	Source epochs	Patience	Linear probing	Fine-tuning	Batch size
12	100	20	5	50	128
60	100	20	5	50	64

B.3 COMPUTE ENVIRONMENT

Table B.3 records the software and hardware environment used throughout the benchmark. The NVIDIA driver supports CUDA 12.8, while the PyTorch build itself uses CUDA 12.1.

B.4 INPUT ADAPTATION ACROSS TRANSFER TRACKS

All models operate on the same benchmark samples after preprocessing, and only the interface required by each architecture differs. Table B.4 summarizes this adaptation contract.

The scalp contract uses 16 longitudinal bipolar derivations: `FP1-F7`, `F7-T7`, `T7-P7`, `P7-O1`, `FP2-F8`, `F8-T8`, `T8-P8`, `P8-O2`, `FP1-F3`, `F3-C3`, `C3-P3`, `P3-O1`,

Table B.3: Software and compute environment.

Item	Version / configuration
Operating system	Ubuntu 22.04.5 LTS
CPU	4 × Intel Xeon Gold 5220 @ 2.20GHz
GPU	4 × NVIDIA A40 (46,068MiB each) & 4 × NVIDIA A100 (40GB HBM2 each)
NVIDIA driver	570.172.08
Python	3.11.15
PyTorch	2.5.1+cu121 (CUDA 12.1)
scikit-learn	1.5.2

Table B.4: Input and adaptation contract by transfer track.

Track	Source input	Target input	Adapter policy	Normalization
Cross-dataset	16 standard scalp bipolar derivations	same 16 derivations	common scalp EEG contract	z-score
Cross-modality	16 standard scalp bipolar derivations	native iEEG contacts	audited channel bridges specific to each model, see Table B.5	z-score
Cross-time	historical CHB-MIT, 16 bipolar derivations	future CHB-MIT, same derivations	common scalp EEG contract	z-score
direct iEEG	native iEEG contacts	native iEEG contacts	same native contact contract as target modality	z-score

`FP2-F4, F4-C4, C4-P4, P4-O2`. In cross-modality, scalp recordings use this bridge where the required derivations are available, whereas target iEEG retains its native electrode axis before adaptation to each model. The interfaces fall into a small set of patterns: native-set attention, model-specific slot or vocabulary bridges, union-channel projection or convolution, structured resize/covariance/image bridges, and permutation-invariant statistical aggregation. RIVER instead passes the native electrode axis directly to OAPT. Table B.5 gives the model-level contract.

B.5 Statistical and traditional ML comparators

The statistical comparators use fixed, permutation-invariant summaries where appropriate, so their cross-modality interface does not depend on native contact order.

The SVM representation combines activity, Hjorth statistics, mean absolute first difference, RMS, skewness, kurtosis, spectral entropy, band log power, and relative band power over delta (0.5–4Hz), theta (4–8Hz), alpha (8–13Hz), beta (13–30Hz), low gamma (30–45Hz), mid gamma (45–55Hz), and high gamma (65–95Hz). Its spectral template is fixed across tracks. For CD and CT, the two bands above 45Hz are computed after 45Hz low-pass filtering and therefore contain only residual post-filter energy. For CM they are computed within the 0.5–95Hz signal contract. Native iEEG contacts are aggregated permutation-invariantly, and CMMN estimates its normalization from training-domain features before classifier fitting.

LinearSVC outputs are converted to the shared positive class score by $s = \sigma(f(x)) = [1 + \exp(-f(x))]^{-1}$. We do not fit Platt scaling, isotonic regression, or `CalibratedClassifierCV`. SVM ECE and Brier scores therefore assess the sigmoid-transformed margin rather than a separately calibrated probability.

C RIVER

This section organizes the architectural overview of RIVER. We present RIVER around native patch encoding, state-conditioned transport, mass-aware latent construction, shared latent dynamics, and temporal evidence aggregation.

C.1 Problem formulation

We view each EEG or iEEG recording as a partial observation of a shared epileptic process through a recording-dependent sensing system (Courty et al., 2017). RIVER therefore does not assume channel identity across recordings. Instead, it maps the available native electrodes into a fixed set of latent electrophysiological slots and carries transported observation mass forward as an explicit measure of support.

Table B.5: Cross-modality interface audit and statistical comparator summary. The bridge names the input adapter used by each model. The interface summary states the main masking or projection rule, the backbone entry point, and whether native contact identity is retained beyond the adapter.

Model	Group	Bridge	Interface summary
BENDR	EEG FM	union projection	masked zero fill, 1×1 conv to the 20-channel backbone input, with order dependence after projection
BIOT	EEG FM	native set attention	explicit mask and no truncation, descriptor attention to 18 latent channels with mean token initialization, with identity retained only at adapter input
CBraMod	EEG FM	native set attention	explicit mask and no truncation, descriptor attention to 18 latent channels before the patch backbone, with identity retained only at adapter input
EEGPT	EEG FM	union convolution	no fixed native padding, union channel convolution with target weights mean initialized from source channels, with identity following union indices
LaBraM	EEG FM	position bridge	valid-channel mask, union channels mapped to LaBraM slots; unknown target slots use mean positional initialization, with slots ordered after bridging
ST-EEGFormer	EEG FM	slot bridge	valid-channel mask, target-only slots use mean initialization from source or pre-trained channel embeddings, with slots ordered after bridging
EEGNet	EEG-spec.	native set attention	padded to native capacity, validity inferred from nonzero signal; descriptor attention to 18 latent channels before EEGNet; identity retained only at adapter input
EvoBrain	EEG-spec.	set attention + graph	batch padding with native mask from raw adjacency, descriptor attention to 18 latent nodes; graph rebuilt from latent signals; identity retained only at adapter input
ScatterFormer	EEG-spec.	image bridge	masked input, union channel–time array mapped to normalized, difference, and magnitude images and resized to 64×64; physical identity not preserved
BF-EML	EEG-spec.	23-channel projection	masked input, 1×1 conv from union channels to 23 channels with identity initialization; physical identity not preserved
CST	Transfer-spec.	ResizeNet bridge	mask before padding, native capacity resized to 16 channels; auxiliary branch uses the first 16 padded slots; identity retained only at ResizeNet input
TSMNet	Transfer-spec.	covariance bridge	masked input, union channels processed by temporal Conv1d and covariance features; order-dependent union interface
CMMN	Transfer-spec.	statistical features	native statistics and validity features with permutation-invariant aggregation, training-domain normalization, and logistic regression, contact identity discarded
SVM	Trad. ML	statistical features	handcrafted time and frequency features plus channel availability, permutation-invariant aggregation, StandardScaler, and LinearSVC; contact identity discarded
Random Forest	Trad. ML	statistical features	same handcrafted feature family as SVM, permutation-invariant aggregation and RandomForestClassifier; contact identity discarded
RIVER	Transfer-spec.	native transport	explicit channel mask, no fixed scalp padding; native contacts enter OAPT using temporal unbalanced optimal transport (UOT) and map to 12 latent slots; identity retained through transport

C.2 NOTATION

Table C.1 lists the main symbols used in the RIVER formulation. Here, $\sigma(\cdot)$ denotes the logistic sigmoid, $\odot$ denotes elementwise multiplication, LN denotes layer normalization, and FFN denotes the feed-forward sublayer.

C.3 NATIVE ENCODING AND TEMPORAL MATCHING

Each native electrode is divided into 1s patches and encoded by a shared patch encoder to produce $\mathbf{H}_{c,t} \in \mathbb{R}^D$. In parallel, the state branch computes a 16-dimensional descriptor vector from mean absolute first difference, zero-crossing fraction, Hjorth mobility and complexity, skewness, excess kurtosis, lag-one autocorrelation, spectral entropy, relative band power in seven frequency bands, and correlation with the leave-one-out valid-channel reference. The descriptor template is fixed across tracks. In cross-dataset and cross-time experiments, the 45–65Hz and 65–95Hz entries are computed after low-pass filtering to 45Hz and therefore contain only residual energy above the passband; in cross-modality experiments, the same descriptors are computed within the 0.5–95Hz signal contract.

The descriptor vector is concatenated with its 4s and 16s temporal averages and encoded by an MLP. A second MLP encodes the content token, and a sigmoid fusion gate combines the physiological and

Table C.1: Core RIVER notation.

Symbol	Meaning
C	number of native electrodes or contacts
P	number of 1s temporal patches
D	latent embedding dimension
K	number of shared latent slots
$\mathbf{H}_{c,t}$	native content token for electrode c at patch t
$\widetilde{\mathbf{s}}_{c,t}$	contextualized electrode state used for transport
M_c	validity mask for native electrode c
$\Gamma_{k,c,t}$	transport cost between slot k and electrode c
$\Pi^{\mathrm{raw}}_{k,c,t}$	raw unbalanced transport plan
$m_{k,t}$	transported mass received by slot k
$\mathbf{z}_{k,t}$	latent token after transport and support interpolation
$\mathbf{U}^{\star}_{k,t}$	final shared latent field after the dynamics stack
$\mathbf{Y}_{k,t}$	task readout field after the adaptation residual
o_{clip}	final clip logit

content branches before masked attention across valid electrodes produces $\widetilde{\mathbf{s}}_{c,t}$. Learned normalized projections give a slot query $\overline{\mathbf{q}}_k$ and electrode key $\overline{\mathbf{r}}_{c,t}$, with instantaneous matching cost $\gamma^{\mathrm{inst}}_{k,c,t} = 1 - \langle \overline{\mathbf{q}}_k, \overline{\mathbf{r}}_{c,t} \rangle$. Let $\overline{\gamma}^{(3)}_{k,c,t}$ be its three-patch moving average. We measure state change by $\delta^{\mathrm{chg}}_t = \mathrm{RMS}_{c,h}(\widetilde{s}_{c,t,h} - \widetilde{s}_{c,t-1,h})$ for $t > 1$ and set it to zero at the first patch. A learned continuity gate $\kappa_t = \sigma(g_{\mathrm{cont}}(\delta^{\mathrm{chg}}_t))$ then mixes instantaneous and smoothed matching costs:

$$\Gamma_{k,c,t} = \kappa_t \gamma^{\mathrm{inst}}_{k,c,t} + (1 - \kappa_t)\overline{\gamma}^{(3)}_{k,c,t}. \tag{2}$$

The resulting match therefore depends on the sample, time point, native electrode, and latent slot.

C.4 Observability-Aware Partial Transport

OAPT uses unbalanced transport so that weakly observed latent slots are not forced to receive full mass (Fatras et al., 2021). A learned observability map produces $\rho_{c,t} = g_{\mathrm{obs}}(\widetilde{\mathbf{s}}_{c,t})$, and the masked electrode prior is $b_{c,t} = M_c \exp(\rho_{c,t}/\theta_{\mathrm{mass}}) / \sum_{c'} M_{c'} \exp(\rho_{c',t}/\theta_{\mathrm{mass}})$. The latent reference mass is uniform, $a_k = 1/K$, and the log transport kernel is $\Lambda_{k,c,t} = -\Gamma_{k,c,t}/\varepsilon$. With LSE denoting log-sum-exp over the indexed variable, one Sinkhorn iteration (Cuturi, 2013) is

$$\log u_{k,t} \leftarrow \tau[\log a_k - \mathrm{LSE}_c(\Lambda_{k,c,t} + \log v_{c,t})]\,. \tag{3}$$

$$\log v_{c,t} \leftarrow \tau[\log b_{c,t} - \mathrm{LSE}_k(\Lambda_{k,c,t} + \log u_{k,t})]\,. \tag{4}$$

After the configured iterations, the masked raw transport plan is

$$\Pi^{\mathrm{raw}}_{k,c,t} = M_c \exp(\log u_{k,t} + \Lambda_{k,c,t} + \log v_{c,t})\,. \tag{5}$$

The same plan determines assignment, support, and transported content. Specifically, $A_{k,c,t} = \Pi^{\mathrm{raw}}_{k,c,t} / \sum_{c'} \Pi^{\mathrm{raw}}_{k,c',t}$, $m_{k,t} = \sum_c \Pi^{\mathrm{raw}}_{k,c,t}$, and $\mathbf{z}_{\mathrm{obs},k,t} = \sum_c A_{k,c,t} \mathcal{P}_{\mathrm{val}}(\mathbf{H}_{c,t})$. The support confidence is $\chi_{k,t} = \sigma((\log m_{k,t} - \log(1/K))/\theta_{\mathrm{mass}})$, which interpolates between transported evidence and the learned slot prior:

$$\mathbf{z}_{k,t} = \chi_{k,t}\mathbf{z}_{\mathrm{obs},k,t} + (1 - \chi_{k,t})\mathbf{p}^{\mathrm{slot}}_k. \tag{6}$$

If a row carries negligible mass, the implementation substitutes a uniform assignment over valid electrodes for numerical stability. Here, $\mathcal{P}_{\mathrm{val}} : \mathbb{R}^D \to \mathbb{R}^D$ is the learned value projection and $\mathbf{p}^{\mathrm{slot}}_k$ is the learned prior for slot k. The shared latent dynamics start from $\mathbf{U}^{(0)}_{k,t} = \mathbf{z}_{k,t} + \boldsymbol{\phi}_t$, where $\boldsymbol{\phi}_t$ is a sinusoidal temporal encoding (Vaswani et al., 2017).

C.5 Latent dynamics and evidence aggregation

RIVER uses $N = 6$ latent dynamics blocks. For block j, let $\mathbf{F}^{(j)} = \mathrm{LN}(\mathbf{U}^{(j-1)})$. The local operator $\mathcal{L}$ is a depthwise temporal convolution followed by a pointwise convolution, while the global operator $\mathcal{G}$ is a low-rank Fourier operator along the patch axis. The gate $\boldsymbol{\lambda}^{(j)} = \sigma(g_{\mathrm{scale}}(\mathbf{F}^{(j)}))$

forms $\mathbf{T}^{(j)} = \boldsymbol{\lambda}^{(j)} \odot \mathcal{L}(\mathbf{F}^{(j)}) + (1 - \boldsymbol{\lambda}^{(j)}) \odot \mathcal{G}(\mathbf{F}^{(j)})$. Multihead self-attention over the K latent slots gives a cross-slot update $\mathbf{B}^{(j)}$. Writing $\widehat{\mathbf{U}}^{(j)} = \mathbf{U}^{(j-1)} + (\mathbf{T}^{(j)} + \mathbf{B}^{(j)})/\sqrt{2}$, each block applies

$$\mathbf{U}^{(j)} = \widehat{\mathbf{U}}^{(j)} + \mathrm{FFN}\Big(\mathrm{LN}(\widehat{\mathbf{U}}^{(j)})\Big). \tag{7}$$

The final shared field is $\mathbf{U}^{\star} = \mathrm{LN}(\mathbf{U}^{(N)})$.

The evidence head contains a bottleneck residual branch whose output projection is initialized to zero. Its correction is $\mathbf{c}^{\mathrm{res}}_{k,t} = W_{\mathrm{up}}\,\mathrm{GELU}(W_{\mathrm{down}}\,\mathrm{LN}(\mathbf{U}^{\star}_{k,t}))$, and the global residual gate is $\beta_{\mathrm{res}} = \sigma(h_{\mathrm{res}}((KP)^{-1}\sum_{k,t}\mathbf{U}^{\star}_{k,t}))$. The task readout is

$$\mathbf{Y}_{k,t} = \mathbf{U}^{\star}_{k,t} + I_{\mathrm{adapt}}\beta_{\mathrm{res}}\mathbf{c}^{\mathrm{res}}_{k,t}, \tag{8}$$

where $I_{\mathrm{adapt}} = 0$ during source training and $I_{\mathrm{adapt}} = 1$ during target adaptation.

The learned evidence map gives slot evidence $\ell^{\mathrm{slot}}_{k,t} = h_{\mathrm{ev}}(\mathbf{Y}_{k,t})$ and patch evidence $e_t = \log\sum_k \exp(\ell^{\mathrm{slot}}_{k,t}) - \log K$. The duration weight is $\omega_t = \mathrm{MA}_3[\sigma(h_{\mathrm{dur}}(K^{-1}\sum_k \mathbf{Y}_{k,t}))]$, while $\mu_{\mathrm{mix}} = \sigma(h_{\mathrm{mix}}((KP)^{-1}\sum_{k,t}\mathbf{Y}_{k,t}))$ controls the mixture of peak and persistent evidence. We use $E_{\mathrm{peak}} = \theta_{\mathrm{pool}}[\log\sum_t \exp(e_t/\theta_{\mathrm{pool}}) - \log P]$ and $E_{\mathrm{persistent}} = (\sum_t \omega_t e_t)/(\sum_t \omega_t)$, giving the final clip logit

$$o_{\mathrm{clip}} = \mu_{\mathrm{mix}}E_{\mathrm{peak}} + (1 - \mu_{\mathrm{mix}})E_{\mathrm{persistent}}. \tag{9}$$

The positive class score is $\sigma(o_{\mathrm{clip}})$. This readout separates brief high-evidence activity from evidence that persists over time.

C.6 Architecture configuration

The fixed architecture used in all benchmark experiments is summarized in Table C.2.

Table C.2: RIVER architecture parameters.

Component	Parameter	Value
Native encoder	raw input	$[B, 1, C, P, 256]$, with $P = 12$ or 60
Native encoder	patch length / content dimension	256 samples (1s) / 128
State branch	descriptor / hidden dimension	16 / 64
State branch	descriptor scales	1s / 4s / 16s
OAPT	latent slots K / embedding D	12 / 128
OAPT	transport mode / Sinkhorn iterations	temporal UOT / 4
OAPT	ε / marginal relaxation τ	0.20 / 0.80
OAPT	shared mass temperature θ_{mass}	0.50
Latent dynamics	depth / heads / feedforward dimension	6 / 8 / 384
Latent dynamics	Fourier modes / rank	32 / 8
Latent dynamics	local kernel / mixer	5 / local spectral
Evidence head	pooling temperature θ_{pool} / smoothing kernel	0.50 / 3
Evidence head	residual bottleneck	32
Evidence head	residual stage switch	source off / target adaptation on

In addition to the clip logit, the evidence head exposes patch evidence, peak and persistent summaries, duration weights, and gating values used in the mechanism analyses below.

C.7 Time and space complexity

Let I_{SK} denote the number of Sinkhorn iterations and D_{ff} the feed-forward width. With patch length and encoder widths fixed, the dominant forward cost is

$$\begin{aligned} T_{\mathrm{forward}} = \mathcal{O}\Big(&P(CD^2 + C^2D) + PKCD + I_{\mathrm{SK}}PKC \\ &+ N\big[KDP\log P + P(KD^2 + K^2D) + PKDD_{\mathrm{ff}}\big]\Big). \end{aligned} \tag{10}$$

Working memory, excluding parameters and optimizer state, scales as

$$M_{\mathrm{work}} = \mathcal{O}\big(PCD + PKC + PKD + PC^2 + PK^2\big). \tag{11}$$

Training additionally stores activations for the N latent blocks. In our configuration ($K = 12$, $N = 6$, $I_{\mathrm{SK}} = 4$, $D = 128$), OAPT is linear in the number of native electrodes for fixed K; the quadratic dependence on C comes from relational attention over native electrodes.

D RIVER Ablations and Mechanism Analysis

We conduct ablations to address three distinct questions: whether variants are parameter matched, whether each component contributes consistently across transfer settings, and whether those contributions persist as electrode observations become sparse.

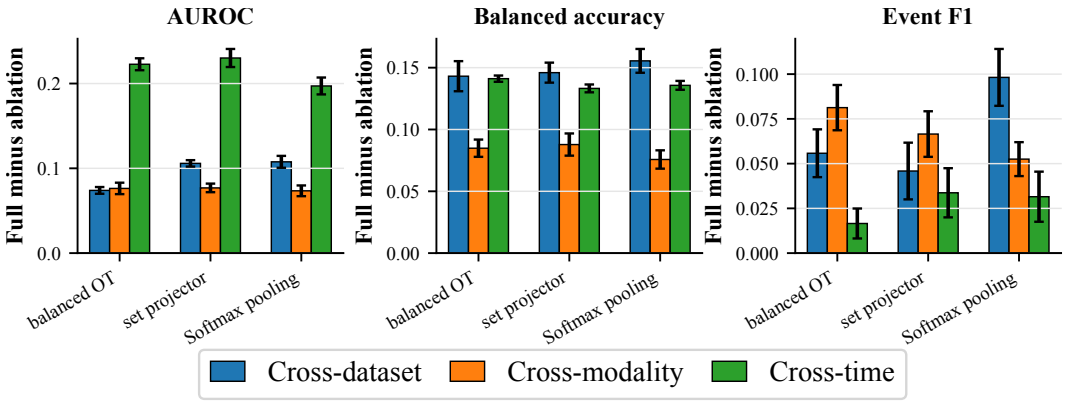

Figure D.1: RIVER component contribution across transfer settings. Bars show paired Full RIVER minus ablation differences aggregated over the five target supervision levels and three seeds within each transfer track; error bars denote SEM, and positive values favor Full RIVER.

D.1 Parameter matching

Parameter matching is assessed with both total trainable parameters and *active* trainable parameters, i.e., parameters that participate in the forward and backward passes. Softmax pooling replaces unbalanced transport with prior-weighted softmax pooling, balanced OT sets the electrode mass relaxation from 0.80 to 1.00, and the set projector variant replaces the transport-based OAPT mapping with a parameter-matched set projector. Table D.1 confirms that all variants remain within the stated 0.5% tolerance.

Table D.1: Parameter matching for the RIVER ablations. Differences in active parameters are measured relative to Full RIVER; all variants fall within the 0.5% matching threshold.

Variant	Total trainable	Active trainable	Active Δ vs. Full	Active Δ (%)
Full RIVER	2,194,314	2,185,577	0	0.000
Softmax pooling	2,194,314	2,185,577	0	0.000
balanced OT	2,194,314	2,185,577	0	0.000
set projector	2,194,668	2,185,931	+354	+0.016

D.2 Component contribution across transfer settings

Figure D.1 shows that each ablation reduces 12s detection performance in the matched comparisons, although the size of the effect depends on transfer track. Table D.2 gives the underlying Full RIVER minus ablation differences after aggregating the five fine-tuning levels and three seeds. Event F1 is computed at the seizure-event level.

The gains are consistently positive for AUROC and balanced accuracy across CD, CM, and CT. Event-level F1 follows the same direction but with smaller and more variable effects.

Table D.2: RIVER component contribution across target supervision levels. Entries are paired Full RIVER minus ablation differences for 12s detection, aggregated over five target supervision levels and three seeds within each transfer track. Values are mean ± SEM. Positive values favor the full model.

Track	Ablation	Δ AUROC	Δ Balanced Accuracy	Δ Event F1
Cross-dataset	balanced OT	+0.074 ± 0.004	+0.143 ± 0.012	+0.056 ± 0.013
	set projector	+0.106 ± 0.004	+0.146 ± 0.008	+0.046 ± 0.016
	Softmax pooling	+0.108 ± 0.007	+0.156 ± 0.010	+0.098 ± 0.016
Cross-modality	balanced OT	+0.076 ± 0.007	+0.085 ± 0.007	+0.081 ± 0.013
	set projector	+0.077 ± 0.005	+0.088 ± 0.009	+0.067 ± 0.013
	Softmax pooling	+0.073 ± 0.006	+0.076 ± 0.007	+0.053 ± 0.009
Cross-time	balanced OT	+0.223 ± 0.007	+0.141 ± 0.003	+0.017 ± 0.008
	set projector	+0.230 ± 0.011	+0.133 ± 0.003	+0.034 ± 0.014
	Softmax pooling	+0.197 ± 0.010	+0.136 ± 0.004	+0.032 ± 0.014

D.3 Cross-modality ablation and electrode scarcity

Figure D.2 separates standard cross-modality ablation performance from a distinct electrode scarcity stress test. The middle panel reports the standard 12s cross-modality detection ablations across target supervision levels, whereas the left and right panels examine robustness to progressively reduced electrode availability. To make the scarcity trajectories directly comparable, all variants are normalized to the same Full RIVER reference with all electrodes. For variant v and retained-electrode ratio r,

$$\mathrm{Retention}(v, r) = \frac{\mathrm{AUROC}(v, r)}{\mathrm{AUROC}(\text{Full RIVER}, 1.00)}. \tag{12}$$

By construction, Full RIVER equals 1.000 at Keep 1.00. An ablated variant can therefore begin below 1.000 even before electrode removal, and its subsequent trajectory reflects both its initial performance deficit relative to Full RIVER and its sensitivity to reduced electrode availability. The degradation slope is fitted to this same common-reference retention curve. Values near zero indicate little additional change as electrodes are removed, whereas positive values indicate a stronger decline. Negative slopes can occur when a weaker variant starts substantially below Full RIVER and rebounds slightly for particular electrode subsets; they should therefore not be interpreted as superior robustness in absolute performance.

The standard ablation and scarcity views answer complementary questions under a shared visualization: Table D.3 reports zero-shot cross-modality performance under the standard ablation setting, while Table D.4 reports the common-reference retention values used by the scarcity panels.

Table D.3: Cross-modality RIVER ablation in the zero-shot setting for 12s detection. AUROC, balanced accuracy, F1, and AUPRC are reported as mean ± SEM over three seeds from the standard ablation runs. Best values are bold and second best values are underlined.

Variant	AUROC ↑	Balanced Accuracy ↑	F1 ↑	AUPRC ↑
balanced OT	0.776 ± 0.003	0.693 ± 0.004	0.682 ± 0.013	0.777 ± 0.007
set projector	0.772 ± 0.002	<u>0.703 ± 0.005</u>	<u>0.704 ± 0.005</u>	<u>0.804 ± 0.001</u>
Softmax pooling	<u>0.778 ± 0.002</u>	<u>0.703 ± 0.009</u>	0.703 ± 0.009	0.783 ± 0.008
Full RIVER	**0.850 ± 0.005**	**0.735 ± 0.010**	**0.742 ± 0.008**	**0.832 ± 0.009**

The standard ablation results show a consistent advantage for the complete RIVER formulation. Full RIVER reaches an AUROC of 0.850 ± 0.005, compared with 0.776 ± 0.003 for balanced OT, 0.772 ± 0.002 for the set projector, and 0.778 ± 0.002 for Softmax pooling. The same pattern holds for balanced accuracy, F1, and AUPRC, where Full RIVER achieves 0.735 ± 0.010, 0.742 ± 0.008, and 0.832 ± 0.009, respectively. Although the relative ordering of the three alternatives varies slightly across metrics, none matches the complete model. Together with the parameter-matched comparisons above, this indicates that the advantage of RIVER is associated with the transport formulation rather than simply additional model capacity.

Table D.4: Cross-modality electrode scarcity stress test using Full RIVER at Keep 1.00 as the common reference. Retention values are mean ± SEM across electrode subsampling repetitions where available. The degradation slope is computed from the same common reference retention curve; values near zero indicate little additional change as electrode availability decreases.

Variant	Degradation slope	Keep 1.00	Keep 0.75	Keep 0.50	Keep 0.25
balanced OT	-0.0478	0.583 ± 0.000	0.676 ± 0.043	0.621 ± 0.013	0.639 ± 0.031
set projector	+0.0342	0.723 ± 0.000	0.647 ± 0.038	0.622 ± 0.025	0.718 ± 0.007
Softmax pooling	-0.0537	0.584 ± 0.000	0.661 ± 0.033	0.638 ± 0.035	0.635 ± 0.036
Full RIVER	**+0.0052**	**1.000 ± 0.000**	**1.003 ± 0.002**	**1.005 ± 0.000**	**0.995 ± 0.005**

The scarcity analysis further distinguishes baseline performance from robustness to missing electrodes. Full RIVER remains close to its all-electrode reference across the entire sweep, with retention values of 1.000, 1.003, 1.005, and 0.995 at Keep 1.00, 0.75, 0.50, and 0.25, respectively, yielding a near-zero degradation slope of +0.0052. In contrast, replacing OAPT with the set projector produces a larger positive slope (+0.0342), while balanced OT and Softmax pooling begin substantially below the common Full RIVER reference and exhibit non-monotonic trajectories, resulting in negative fitted slopes. These negative slopes reflect partial rebound from a lower starting point rather than stronger absolute performance.

Taken together, the two views suggest that the benefit of OAPT is not limited to higher nominal cross-modality performance. Its partial transport and observability-aware representation are also consistent with stable inference when only a subset of electrodes remains available. This behavior is particularly relevant to heterogeneous iEEG recordings, where electrode coverage can vary substantially across patients. At the same time, because the scarcity curves use a common Full RIVER denominator, the slope should be interpreted jointly with the absolute ablation results rather than as an independent measure of model quality.

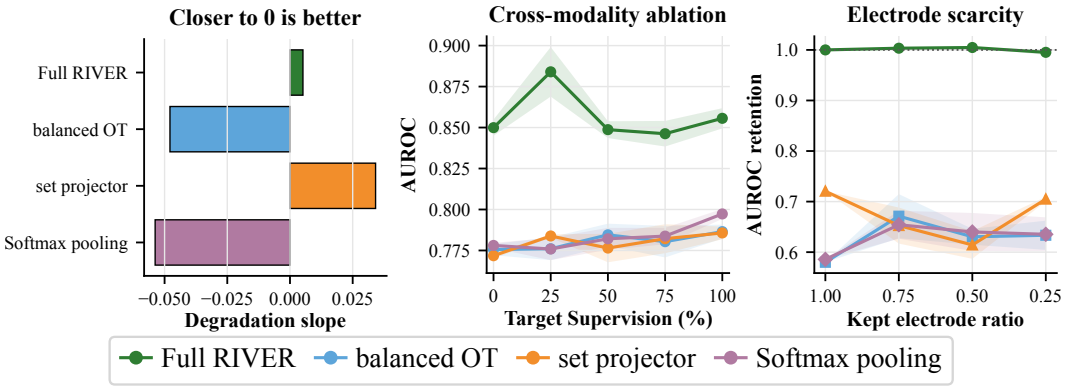

Figure D.2: Cross-modality ablation and electrode scarcity stress test. Left: degradation slope computed from the common reference retention curves. Middle: standard 12s cross-modality detection AUROC across target supervision levels. Right: AUROC relative to the Full RIVER Keep 1.00 reference as the retained electrode ratio decreases. Left and right use the same Full RIVER denominator with all electrodes.

E Statistical Analysis

E.1 Analysis units and estimation

Statistical inference follows the unit represented by each claim. Target supervision level analyses are matched by model, direction, task, window length, and seed. Window length analyses are matched by model, direction, task, and seed in the zero-shot setting. SOZ analyses instead use patients as the inference unit. Bootstrap confidence intervals use 10,000 resamples, with deterministic analysis-specific offsets. We report AUROC, F1, balanced accuracy, AUPRC, ECE, and Brier score

where supported by the corresponding analysis. Thresholded metrics use a clip-level F1-maximizing validation threshold chosen before test evaluation from $\{0, 1\}$ and all observed validation scores, with ties resolved toward the largest threshold; 0% uses source dev (`source_dev_max_f1`), positive budgets use target dev (`target_dev_max_f1` or `fixed_target_dev_max_f1`), and the selected threshold is then fixed for target-test F1, balanced accuracy, precision, and recall. Target test labels never participate in threshold selection.

E.2 Target supervision fine-tuning level marginal gain analysis

For a matched setting, let $\mathcal{J}_b^{(m,d,t,w,s)}$ denote the observed evaluation metric at target supervision level b. The gain between adjacent budgets is

$$\Delta_b^{(m,d,t,w,s)} = \mathcal{J}_b^{(m,d,t,w,s)} - \mathcal{J}_{b^-}^{(m,d,t,w,s)}, \tag{13}$$

where m, d, t, w, and s denote model, direction, task, window length, and seed, and b^- denotes the immediately preceding target supervision level. The primary contrast is

$$C_{\text{budget}} = \Delta_{0\to25} - \frac{\Delta_{25\to50} + \Delta_{50\to75} + \Delta_{75\to100}}{3}. \tag{14}$$

Adjacent gains are formed within matched model, direction, task, window length, and seed trajectories. Seeds and then directions/window lengths are averaged to one contrast per model. At $b = 0$, $\mathcal{J}_0$ is the target test score produced by the source-only zero-shot procedure. For neural models, this score comes from the checkpoint selected using the source validation set. At each positive target supervision level, neural models are initialized from that fixed zero-shot checkpoint and fine-tuned on the corresponding target patient subset for CD/CM or chronological positive-event adapt prefix for CT. For CD/CM, checkpoint selection uses AUROC from the target validation subset at the same percentage p as the target training subset, with the train and validation fractions treated separately rather than added. CT uses its fixed dev block. SVM, CMMN, and Random Forest follow the same budget-specific target subset together with source rehearsal. Consequently, $\Delta_{0\to25}$ measures the first supervised target-adaptation gain relative to zero-shot transfer, while the later increments compare successive positive target-supervision budgets under the same target validation protocol. With $n = 16$, the two-sided sign-flip test enumerates all $2^{16} = 65{,}536$ sign assignments. The six combinations of transfer track and task form one Holm correction family.

E.3 Complementary recovery across model families

The complementary recovery analysis asks whether the first supervised adaptation step changes ranking quality and probability quality in the same direction. In Table 1, AUPRC recovery is $\text{AUPRC}_{25} - \text{AUPRC}_0$, while ECE and Brier use $\text{Metric}_0 - \text{Metric}_{25}$ so that positive values always denote improvement. Gains are averaged over seeds and then over the available directions and window lengths with equal model weight.

E.4 Window length sensitivity and rank stability

Window length sensitivity is evaluated in the zero-shot setting by matching each model, direction, task, and seed across the two window lengths:

$$\Delta_{\text{window}} = \mathcal{J}_{60\text{s}} - \mathcal{J}_{12\text{s}}, \tag{15}$$

using AUROC for detection and F1 for prediction. Contrasts for each seed are averaged within each model and direction and then across directions to one contrast per model. Mean effects use the same exact two-sided test over 16 model blocks, with the six combinations of transfer track and task corrected together using Holm's method. Model ranking stability is measured by Spearman correlation between the 12s/60s model rankings after averaging seeds and directions to model means.

E.5 Calibration and probability quality

Reliability metrics consume a common positive class score $s \in [0, 1]$ and binary label $y \in \{0, 1\}$. Probabilistic models contribute their positive class probability; LinearSVC contributes the sigmoid-transformed margin defined in Appendix B. ECE uses 15 bins of equal width on $[0, 1]$ (Guo et al.,

2017): the first 14 bins are left-closed/right-open, the final bin includes 1, empty bins are skipped, and each nonempty bin contributes its sample fraction times the absolute difference between mean score and empirical positive rate,

$$\mathrm{ECE} = \sum_{b=1}^{15} \frac{n_b}{N} \left| \overline{s}_b - \overline{y}_b \right|. \tag{16}$$

Brier score is

$$\mathrm{Brier} = \frac{1}{N} \sum_{i=1}^{N} (s_i - y_i)^2. \tag{17}$$

The AUROC–ECE association is computed separately for each transfer and task after averaging seeds within each model, direction, window length, and target supervision level. Bootstrap uncertainty uses complete model blocks, preserving all settings belonging to each model.

E.6 SOZ EVIDENCE

We conduct SOZ analysis to ask whether zero-shot TUSZ→Epilepsy-iEEG detection evidence aligns with clinically annotated contacts (Li et al., 2026). Reusing the completed cross-modality detection checkpoints on the fixed target-test cohort, native contact evidence is generated through a common localization interface: neural models use $|x \odot \nabla_x o_+|$ for the positive logit (Ancona et al., 2018), summed over time and non-contact dimensions and normalized within each clip. SVM, Random Forest, and CMMN are omitted from the SOZ ranking because their statistical feature pipelines discard native contact identity before classification and therefore do not yield contact-resolved evidence. SOZ ranking uses positive clips only, averages evidence for each patient–contact pair, and ranks contacts from highest to lowest evidence. RIVER transport diagnostics are analyzed separately and are not substituted for this localization evidence. We use enrichment, SOZ AUROC, average precision, Recall@K, and Hit@1/3/5/10, with K equal to the patient-specific number of annotated SOZ contacts. The main cohort contains $n = 8$ patients. Seeds are averaged within patient–model pairs and patients are bootstrapped for uncertainty. Enrichment is tested against 1 with a one-sided Wilcoxon signed-rank test after seed averaging, with all model-wise tests from the 12s and 60s window lengths forming a single Holm correction family. Paired RIVER–comparator tests use shared patients, two-sided Wilcoxon signed-rank tests, and Holm correction across the SOZ family.

F COMPLETE BENCHMARK RESULTS

This section complements the main results with trajectories for individual models, family calibration distributions, and ranks across all five fine-tuning levels. The marginal target supervision claims use the paired sign-flip analysis defined in Appendix E. The supplementary AUROC and ECE annotations in this section compare each nonzero target supervision level with the 0% level using two-sided paired Wilcoxon signed-rank tests matched by window length, transfer direction, model, and seed. The 12s and 60s window lengths remain separate matched observations. Within each AUROC panel, the four nonzero-budget comparisons form one Holm correction family; within each ECE model family, its four budget comparisons form one Holm correction family.

F.1 CROSS-DATASET TRANSFER

Figure F.1 shows that both detection directions concentrate most recovery at the first positive target supervision level. Averaged over directions, AUROC increases from 0% to 100% by about 0.060 at 12s and 0.055 at 60s, and every nonzero level differs significantly from zero-shot after Holm correction. Consistent with the aggregate result, EEG foundation models are frequently among the stronger baseline ranks, while EEG-specific, transfer-specific, and traditional ML models shift more with transfer direction. RIVER and several foundation or transfer-specific models remain individually competitive, but no family leads both directions. Family ECE distributions do not mirror the AUROC ordering, indicating that the common early recovery trend does not imply a universal family hierarchy in calibration.

Figure F.2 shows stronger directional asymmetry in prediction. Across directions, the aggregate 0%–100% AUROC gain is about 0.045 at 12s and 0.035 at 60s, but family and model rankings

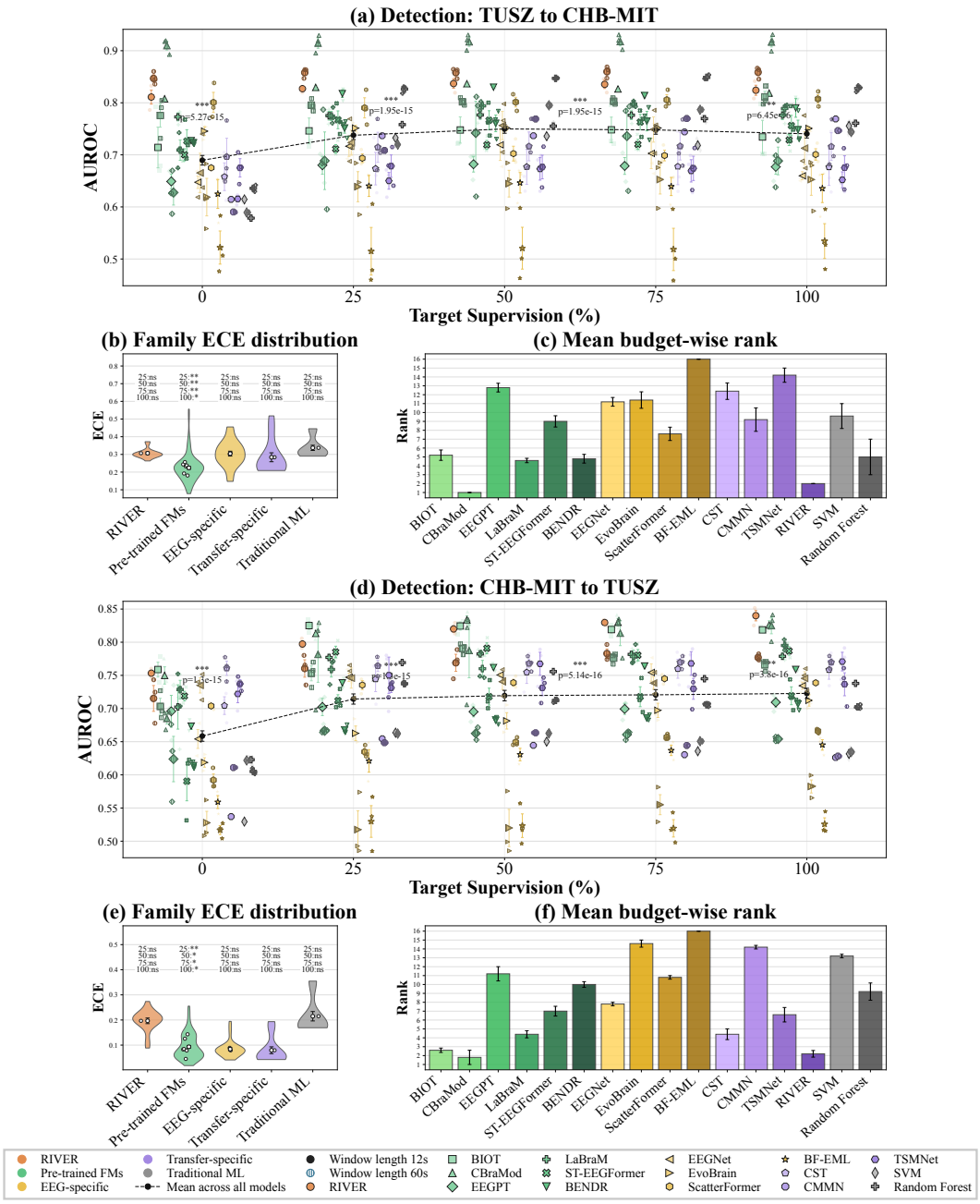

Figure F.1: Cross-dataset detection. Panels (a–c) show D1 TUSZ → CHB-MIT and panels (d–f) show D2 CHB-MIT → TUSZ. For each direction, the left panel shows AUROC across the five target supervision levels with seed-level 12s/60s window lengths, mean ± SEM, and the across-model mean trend. The center panel shows family ECE distributions, and the right panel shows mean model rank across budgets with SEM across budgets. ECE significance uses paired two-sided Wilcoxon tests versus 0%, matched by window length, direction, model, and seed, with Holm correction over the four budgets within each family.

depend strongly on direction. CHB-MIT→Siena changes modestly, with significance at 25%, 50%, and 100%, but not 75%, whereas Siena→CHB-MIT shows an increase with all nonzero levels above zero-shot after correction. Foundation and transfer-specific models remain competitive across parts of both trajectories, but their relative ranks and ECE distributions change between directions.

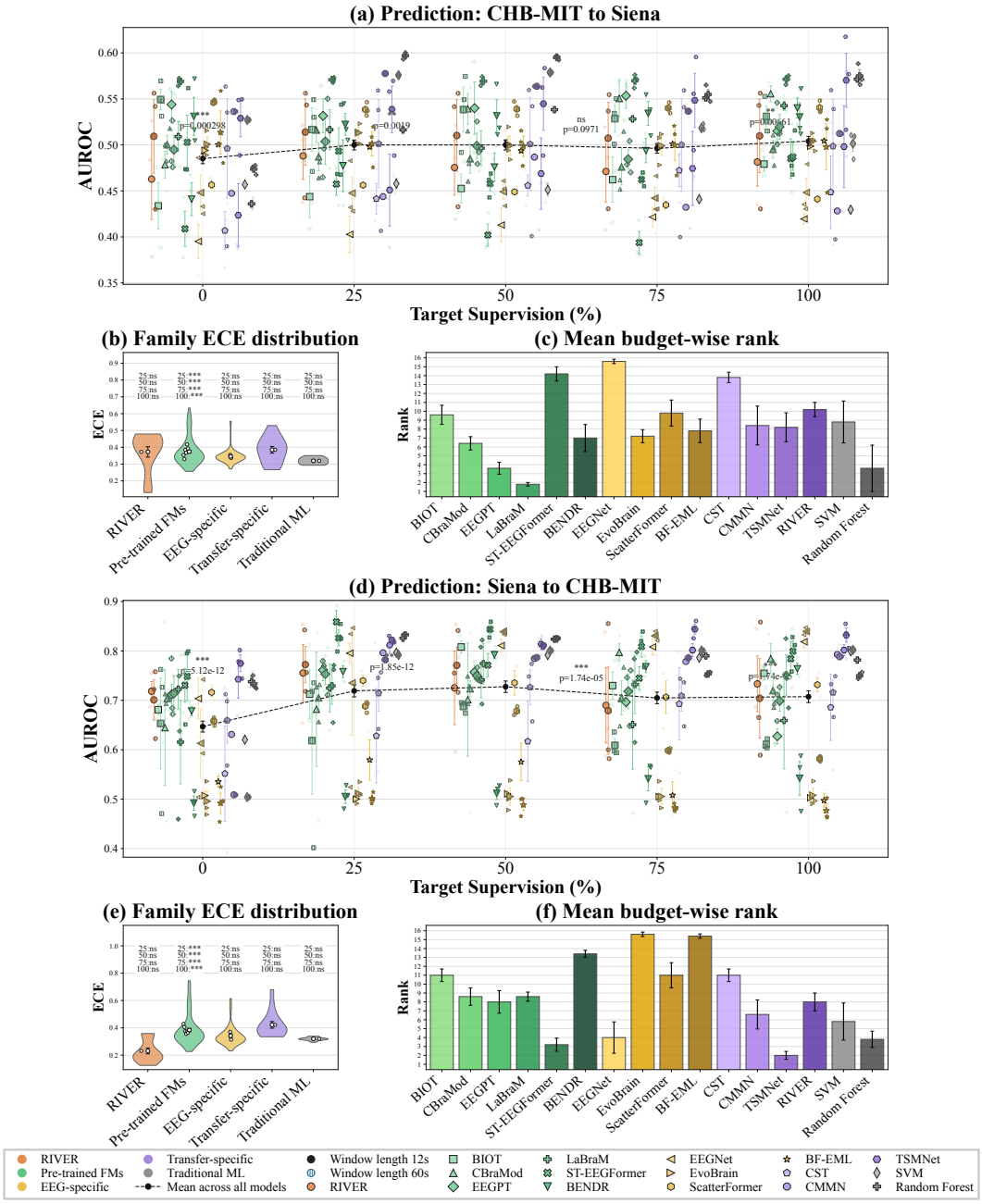

Figure F.2: Cross-dataset prediction. Panels (a–c) show D3 CHB-MIT → Siena and panels (d–f) show D4 Siena → CHB-MIT. AUROC trajectories, family ECE distributions, and mean model ranks across budgets follow the conventions of Figure F.1.

RIVER is competitive for Siena→CHB-MIT without a uniform advantage in the reverse direction, reinforcing that ordered source–target pairing matters more than a fixed family hierarchy.

F.2 CROSS-MODALITY TRANSFER

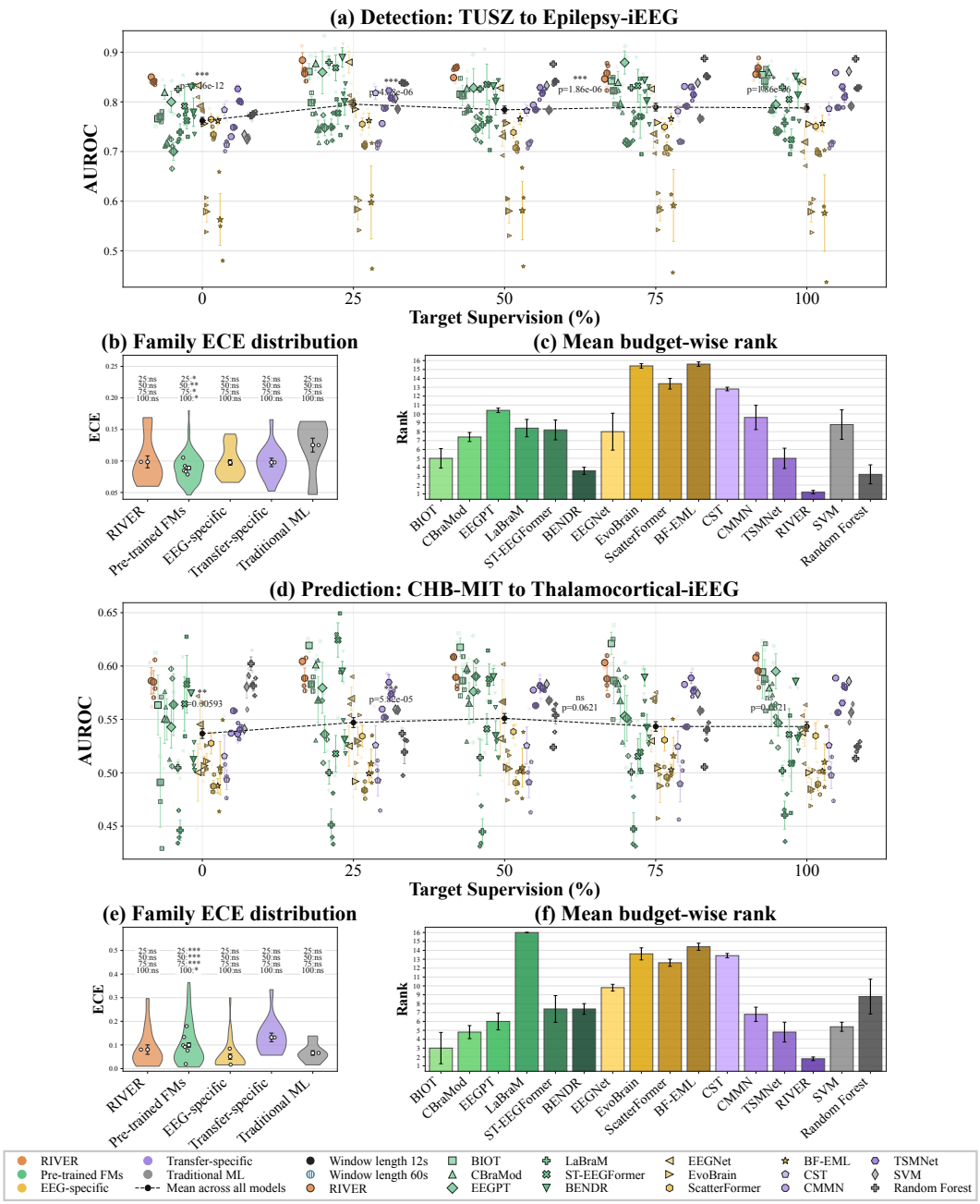

Figure F.3: Cross-modality transfer. Panels (a–c) show D5 TUSZ → Epilepsy-iEEG detection and panels (d–f) show D6 CHB-MIT → Thalamocortical-iEEG prediction. AUROC trajectories, family ECE distributions, and mean model ranks across budgets follow the conventions of Figure F.1.

Figure F.3 separates the family and model trends for cross-modality detection and prediction. Detection shows a clear supervision response, with mean AUROC increasing from 0.790 to 0.824 at 12s and from 0.734 to 0.752 at 60s, and most recovery appearing at the first positive level. RIVER has the strongest mean rank across budgets, while the foundation, EEG-specific, transfer-specific comparator, and traditional ML groups recover from lower starting levels with substantial within-family variation. Prediction is less responsive, increasing from 0.540 to 0.553 at 12s and remaining essentially unchanged at 60s, with significant improvement confined to lower supervision levels.

Family ECE trends differ from discrimination: prediction can improve markedly in calibration despite limited AUROC recovery, and no family shows a uniform calibration advantage.

F.3 CROSS-TIME TRANSFER

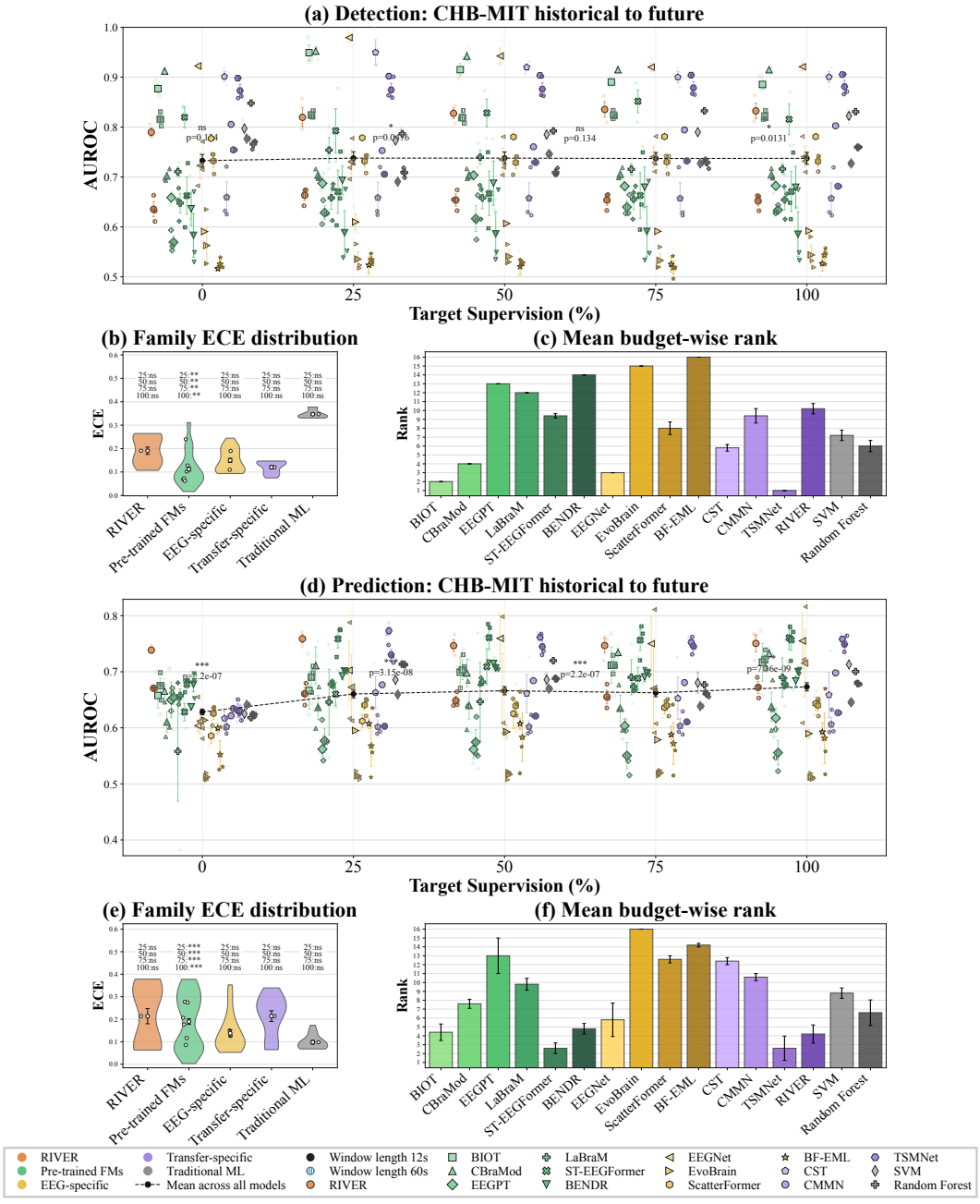

Figure F.4: Cross-time transfer within the same patients. Panels (a–c) show historical → future CHB-MIT detection and panels (d–f) show the corresponding prediction task. AUROC trajectories, family ECE distributions, and mean model ranks across budgets follow the conventions of Figure F.1.

Figure F.4 shows that cross-time transfer does not produce a common family-level adaptation pattern. Detection changes by only about +0.009 from 0% to 100% at 12s and is essentially unchanged at 60s, with significance only at selected budgets. Prediction shows a clearer AUROC increase of

about 0.063 at 12s and 0.027 at 60s, although the F1 recovery emphasized in the main analysis remains limited. Mean ranks are heterogeneous: TSMNet and several foundation models are strong in particular settings, while no family consistently leads both tasks. Calibration also separates the tasks, with detection ECE remaining stable or worsening and prediction ECE improving after future-period supervision.

F.4 Direct iEEG reference

Direct iEEG training establishes the direct reference for the two cross-modality tasks. Each iEEG dataset uses the same fixed patient-level partition as its corresponding benchmark split, and models are trained on the full iEEG training partition with validation-AUROC checkpoint selection. Thresholded metrics use the same validation F1-maximizing threshold rule described in Appendix E, using the direct iEEG validation split. Foundation models retain the initialization in Table B.1, while scratch models, neural schedules, and statistical comparators follow Appendix B.

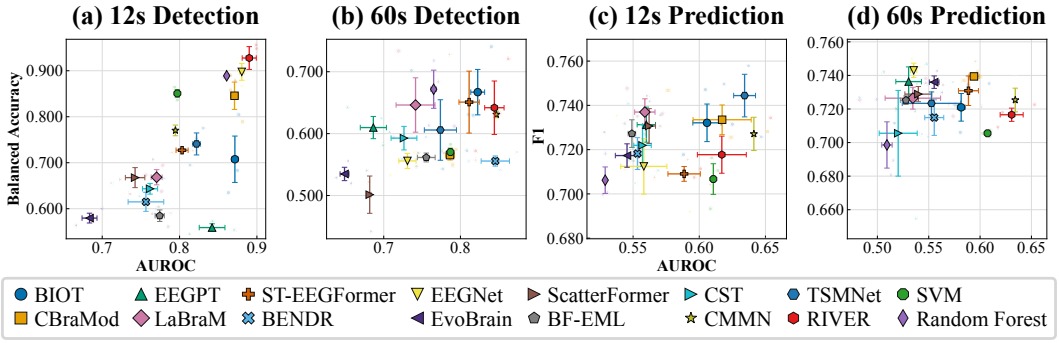

Figure F.5: Direct iEEG reference across tasks and window lengths. The horizontal axis is AUROC. The vertical axis is balanced accuracy for detection and F1 for prediction. Light markers show seed-level results, while larger markers with horizontal and vertical error bars show model means ± SEM.

Figure F.5 shows that direct iEEG training does not produce a single family hierarchy. Foundation models are broadly competitive, EEG-specific EEGNet is strong for detection, and transfer-specific models occupy several leading prediction positions. At the model level, RIVER lies in the upper-right region for 12s detection, while the strongest 60s AUROC and balanced accuracy come from different models. Prediction F1 is more compressed than AUROC, emphasizing that the in-domain reference remains metric dependent.

Table F.1: Direct iEEG reference across detection and prediction. Values are mean ± standard deviation across runs. Within each metric and window length, the best value is bold and the second best is underlined.

	12s				60s			
Model	Det. AUROC	Det. Bal. Acc.	Pred. AUROC	Pred. F1	Det. AUROC	Det. Bal. Acc.	Pred. AUROC	Pred. F1
BENDR	0.757 ± 0.040	0.615 ± 0.036	0.553 ± 0.005	0.718 ± 0.012	0.845 ± 0.032	0.556 ± 0.014	0.555 ± 0.016	0.715 ± 0.018
BIOT	0.872 ± 0.007	0.708 ± 0.087	0.606 ± 0.019	0.732 ± 0.015	0.823 ± 0.016	0.667 ± 0.064	0.582 ± 0.007	0.721 ± 0.014
CBraMod	0.871 ± 0.009	0.846 ± 0.052	0.617 ± 0.038	0.734 ± 0.012	0.786 ± 0.010	0.565 ± 0.014	0.594 ± 0.008	0.739 ± 0.002
EEGPT	0.740 ± 0.036	0.587 ± 0.008	0.560 ± 0.012	0.731 ± 0.014	0.778 ± 0.029	0.558 ± 0.012	0.530 ± 0.022	0.736 ± 0.015
LaBraM	0.729 ± 0.024	0.590 ± 0.030	0.559 ± 0.013	0.737 ± 0.010	0.715 ± 0.013	0.557 ± 0.032	0.535 ± 0.047	0.726 ± 0.011
ST-EEGFormer	0.735 ± 0.010	0.569 ± 0.023	0.589 ± 0.021	0.709 ± 0.006	0.767 ± 0.008	0.545 ± 0.030	0.589 ± 0.017	0.731 ± 0.015
EEGNet	0.883 ± 0.004	0.897 ± 0.041	0.558 ± 0.030	0.712 ± 0.022	0.840 ± 0.011	0.631 ± 0.035	0.535 ± 0.005	**0.743 ± 0.007**
EvoBrain	0.800 ± 0.032	0.641 ± 0.013	0.545 ± 0.015	0.717 ± 0.009	0.697 ± 0.048	0.550 ± 0.026	0.555 ± 0.008	0.736 ± 0.006
ScatterFormer	0.764 ± 0.020	0.663 ± 0.067	0.562 ± 0.008	0.731 ± 0.014	0.809 ± 0.025	0.612 ± 0.024	0.540 ± 0.023	0.729 ± 0.008
BF-EML	0.813 ± 0.002	0.641 ± 0.022	0.549 ± 0.005	0.727 ± 0.011	0.710 ± 0.010	0.586 ± 0.054	0.528 ± 0.010	0.725 ± 0.004
CST	0.777 ± 0.038	0.694 ± 0.018	0.557 ± 0.011	0.722 ± 0.016	0.710 ± 0.072	0.574 ± 0.005	0.520 ± 0.032	0.706 ± 0.044
CMMN	0.857 ± 0.000	0.776 ± 0.020	**0.641 ± 0.000**	0.727 ± 0.013	**0.847 ± 0.000**	0.645 ± 0.019	**0.634 ± 0.000**	0.725 ± 0.012
TSMNet	0.818 ± 0.020	0.741 ± 0.108	0.634 ± 0.014	**0.744 ± 0.017**	0.772 ± 0.056	0.530 ± 0.068	0.553 ± 0.049	0.723 ± 0.012
SVM	0.808 ± 0.002	0.634 ± 0.003	0.611 ± 0.003	0.707 ± 0.012	0.805 ± 0.005	0.564 ± 0.006	0.607 ± 0.004	0.706 ± 0.002
Random Forest	0.715 ± 0.010	0.605 ± 0.011	0.529 ± 0.002	0.706 ± 0.010	0.675 ± 0.020	**0.672 ± 0.011**	0.509 ± 0.009	0.699 ± 0.024
RIVER	**0.890 ± 0.019**	**0.928 ± 0.029**	0.617 ± 0.032	0.718 ± 0.015	0.844 ± 0.018	0.615 ± 0.030	0.631 ± 0.019	0.717 ± 0.007

Table F.1 quantifies the trade-offs across models behind these family trends. RIVER leads 12s detection in AUROC and balanced accuracy. At 60s, transfer-specific CMMN leads AUROC, traditional ML Random Forest leads balanced accuracy, and foundation model BENDR is close to the AUROC maximum. For prediction, CMMN leads AUROC at both window lengths, while TSMNet and EEG-specific EEGNet lead 12s and 60s F1, respectively. RIVER is second in 60s prediction AUROC but does not lead F1, confirming that the direct iEEG reference is metric and window length dependent.

G Cross-Dataset Spatial Evidence

This section compares spatial evidence for the reciprocal 12s zero-shot cross-dataset detection transfers in a single paired view. Figure G.1 places TUSZ→CHB-MIT above CHB-MIT→TUSZ, allowing the direction-dependent channel importance patterns to be compared under the same task and window length. Each block contrasts the ground truth channel importance with the corresponding spatial maps for each model.

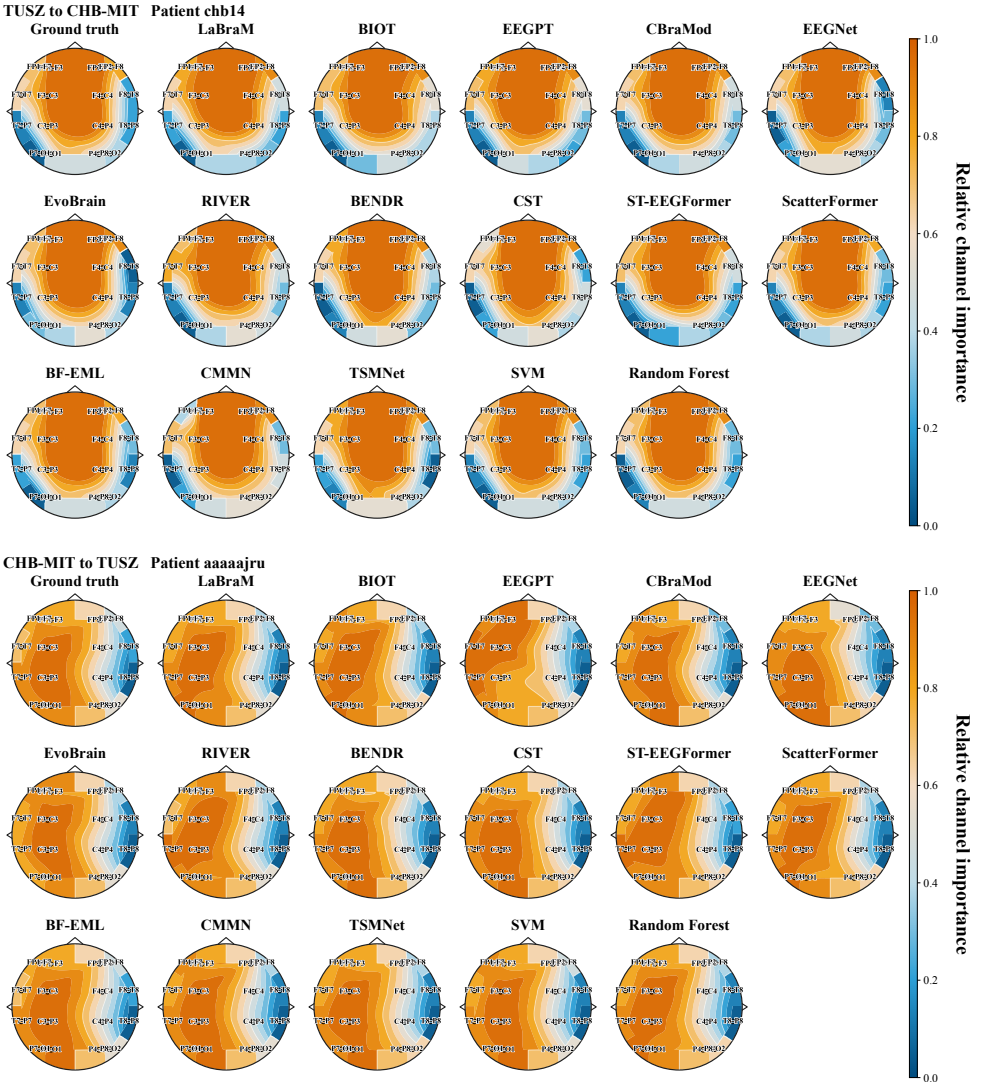

Figure G.1: Bidirectional spatial evidence for 12s zero-shot cross-dataset detection. Upper: D1 TUSZ → CHB-MIT for representative target patient chb14. Lower: D2 CHB-MIT → TUSZ for representative target patient aaaaajru. Each block shows the ground truth channel importance followed by the spatial importance maps of the 16 evaluated models. Colors indicate relative channel importance on a common 0–1 scale within each transfer view.

Figure G.1 shows that transferred spatial evidence preserves coarse organization more consistently than exact channel-level patterns. For TUSZ→CHB-MIT, the reference map is broadly bilateral with elevated importance over central and frontal regions, and most models recover this overall structure despite differences in local weighting. In the reverse CHB-MIT→TUSZ direction, the reference map becomes more left-right asymmetric, a contrast retained by many models. These direction-specific changes suggest that spatial evidence is conditioned by the target cohort and patient, and that similar AUROC does not necessarily imply reliance on the same electrophysiological regions.

H Reliability, Window Length, and Clinical Evidence

H.1 Window length sensitivity and discrimination–calibration association

Table H.1 consolidates the window-length-sensitivity and AUROC–ECE analyses underlying Figure 8. The largest negative window length effects occur for cross-time and cross-modality detection, whereas cross-modality prediction is essentially unchanged despite near-zero rank stability across window lengths. The AUROC–ECE association is strongest and negative for cross-modality detection, while cross-modality prediction is weakly positive.

Table H.1: Window-length-sensitivity, rank stability, and discrimination–calibration association across transfer settings. Window length effects are 60s minus 12s and, together with rank stability, are evaluated under zero-shot transfer, using AUROC for detection and F1 for prediction. Window-length ρ denotes the Spearman rank correlation between the 12s and 60s model rankings. AUROC–ECE ρ denotes the Spearman association between discrimination and calibration across all target supervision levels; 95% CIs use 10,000 bootstrap resamples over complete model blocks. N denotes the number of model-setting observations used for the AUROC–ECE association.

Setting	60s − 12s	Holm p	Window length ρ	AUROC–ECE ρ	95% CI	N
CD Det.	-0.006	1	0.73	-0.18	[-0.358, 0.020]	320
CD Pred.	-0.055	0.0979	0.66	-0.07	[-0.169, 0.036]	320
CM Det.	-0.056	0.0032	0.61	-0.65	[-0.742, -0.539]	160
CM Pred.	+0.001	1	+0.02	+0.19	[-0.252, 0.499]	160
CT Det.	-0.092	0.000366	0.67	-0.47	[-0.726, -0.149]	160
CT Pred.	+0.010	0.725	0.60	-0.25	[-0.622, 0.157]	160

H.2 SOZ evidence

To align the quantitative localization metrics with the window-length-specific spatial maps, we report the 12s and 60s SOZ analyses separately.

Table H.2: SOZ evidence metrics for 12s zero-shot cross-modality detection. Available seeds are first averaged within each patient; results are reported as mean ± standard deviation across $n = 8$ patients. Best values are bold and second best values are underlined.

Model	Enrichment ↑	SOZ AUROC ↑	SOZ AP ↑	Recall@K ↑	Hit@1 ↑	Hit@3 ↑	Hit@5 ↑	Hit@10 ↑
BENDR	0.997 ± 0.029	0.458 ± 0.225	0.305 ± 0.137	0.186 ± 0.210	0.375 ± 0.518	0.625 ± 0.452	0.708 ± 0.452	0.792 ± 0.354
BIOT	**1.784 ± 0.702**	0.722 ± 0.143	0.453 ± 0.221	0.408 ± 0.314	0.625 ± 0.415	0.792 ± 0.354	0.833 ± 0.252	**1.000 ± 0.000**
CBraMod	1.769 ± 0.555	0.737 ± 0.136	0.491 ± 0.223	0.456 ± 0.314	0.625 ± 0.415	0.875 ± 0.248	0.875 ± 0.248	**1.000 ± 0.000**
EEGPT	1.460 ± 0.609	0.720 ± 0.142	**0.547 ± 0.258**	**0.488 ± 0.260**	**0.750 ± 0.463**	0.875 ± 0.354	0.875 ± 0.354	**1.000 ± 0.000**
LaBraM	1.746 ± 0.653	0.732 ± 0.140	0.499 ± 0.180	0.460 ± 0.244	0.625 ± 0.415	0.875 ± 0.248	0.958 ± 0.118	**1.000 ± 0.000**
ST-EEGFormer	1.285 ± 0.335	0.728 ± 0.163	0.495 ± 0.290	0.442 ± 0.326	0.625 ± 0.518	0.750 ± 0.463	0.750 ± 0.463	0.875 ± 0.354
EEGNet	1.460 ± 0.739	0.663 ± 0.168	0.457 ± 0.171	0.377 ± 0.103	**0.750 ± 0.345**	0.875 ± 0.248	**1.000 ± 0.000**	**1.000 ± 0.000**
EvoBrain	1.006 ± 0.039	0.505 ± 0.278	0.385 ± 0.293	0.231 ± 0.271	0.375 ± 0.452	0.500 ± 0.436	0.583 ± 0.427	0.750 ± 0.388
ScatterFormer	1.540 ± 0.749	0.701 ± 0.186	0.537 ± 0.202	0.414 ± 0.253	0.667 ± 0.356	0.875 ± 0.173	0.917 ± 0.154	**1.000 ± 0.000**
BF-EML	1.507 ± 0.641	0.682 ± 0.165	0.507 ± 0.219	0.384 ± 0.266	**0.750 ± 0.295**	0.875 ± 0.248	0.917 ± 0.154	0.958 ± 0.118
CST	1.019 ± 0.297	0.500 ± 0.176	0.281 ± 0.099	0.180 ± 0.170	0.333 ± 0.252	0.458 ± 0.354	0.500 ± 0.356	0.667 ± 0.436
TSMNet	1.434 ± 0.302	**0.751 ± 0.103**	0.525 ± 0.173	0.438 ± 0.279	0.708 ± 0.452	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
RIVER	1.660 ± 0.487	0.748 ± 0.142	0.542 ± 0.231	0.437 ± 0.229	0.583 ± 0.345	0.875 ± 0.354	0.958 ± 0.118	**1.000 ± 0.000**

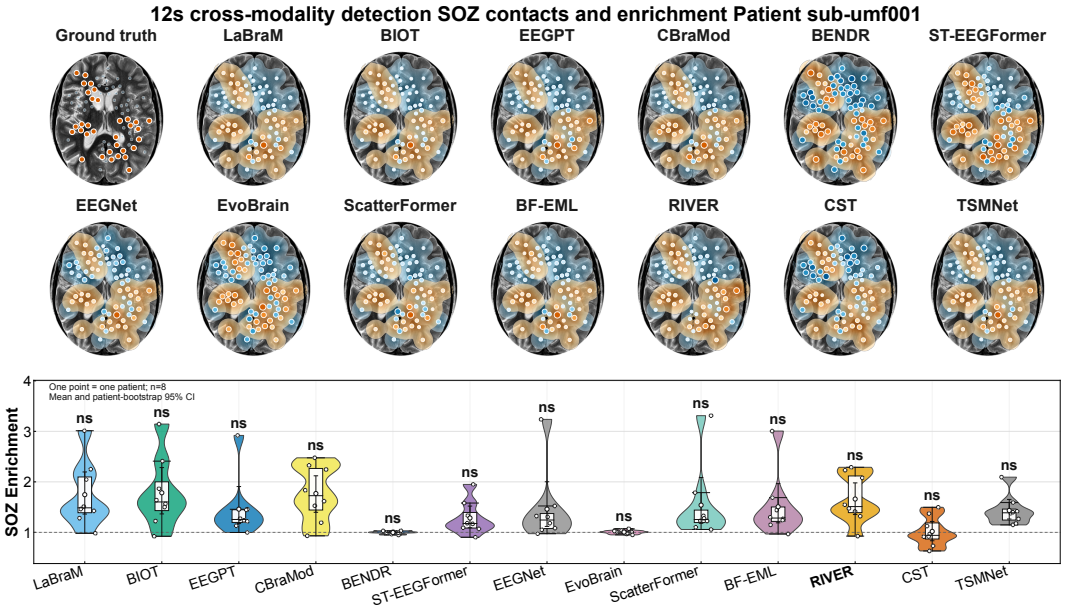

Figure H.1: SOZ evidence for 12s zero-shot cross-modality detection. The contact maps compare clinically annotated SOZ contacts with native contact evidence for patient sub-umf001 across neural models. The violin plots summarize SOZ enrichment across $n = 8$ patients with the mean and patient bootstrap 95% confidence interval; the dashed line at one denotes chance enrichment. Enrichment significance uses one-sided Wilcoxon tests at the patient level after seed averaging with one Holm family across both window lengths. Significance: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, and ns.

Table H.2 and Figure H.1 summarize the 12s localization results. Different models lead different criteria: BIOT has the highest enrichment (1.784), TSMNet the highest SOZ AUROC (0.751), and EEGPT the highest SOZ AP (0.547) and Recall@K (0.488), while RIVER remains near the leading values across several metrics. Although enrichment exceeds one for many models, the patient-level distributions overlap substantially and the enrichment comparisons shown in Figure H.1 do not separate clearly from chance.

Table H.3: SOZ evidence metrics for 60s zero-shot cross-modality detection. Metric definitions and aggregation follow Table H.2. Best values are bold and second best values are underlined.

Model	Enrichment ↑	SOZ AUROC ↑	SOZ AP ↑	Recall@K ↑	Hit@1 ↑	Hit@3 ↑	Hit@5 ↑	Hit@10 ↑
BENDR	0.958 ± 0.065	0.347 ± 0.236	0.231 ± 0.137	0.116 ± 0.135	0.208 ± 0.396	0.333 ± 0.398	0.542 ± 0.502	0.625 ± 0.518
BIOT	1.958 ± 1.245	**0.714 ± 0.207**	**0.515 ± 0.349**	0.337 ± 0.302	0.542 ± 0.502	0.625 ± 0.518	0.750 ± 0.463	0.917 ± 0.236
CBraMod	1.042 ± 0.155	0.581 ± 0.288	0.444 ± 0.292	0.269 ± 0.217	0.625 ± 0.452	0.708 ± 0.452	0.708 ± 0.452	0.750 ± 0.463
EEGPT	0.998 ± 0.046	0.447 ± 0.200	0.284 ± 0.162	0.180 ± 0.189	0.292 ± 0.375	0.500 ± 0.471	0.542 ± 0.469	0.708 ± 0.452
LaBraM	1.082 ± 0.087	0.596 ± 0.108	0.380 ± 0.219	0.275 ± 0.184	0.333 ± 0.398	0.500 ± 0.471	0.625 ± 0.375	0.833 ± 0.356
ST-EEGFormer	1.007 ± 0.062	0.432 ± 0.271	0.270 ± 0.151	0.156 ± 0.137	0.250 ± 0.463	0.500 ± 0.535	0.500 ± 0.535	0.750 ± 0.463
EEGNet	2.178 ± 2.729	0.607 ± 0.175	0.454 ± 0.252	0.326 ± 0.297	**0.708 ± 0.375**	**0.833 ± 0.356**	**0.917 ± 0.154**	0.958 ± 0.118
EvoBrain	0.909 ± 0.076	0.373 ± 0.133	0.192 ± 0.110	0.051 ± 0.057	0.000 ± 0.000	0.167 ± 0.309	0.250 ± 0.463	0.458 ± 0.469
ScatterFormer	0.978 ± 0.046	0.365 ± 0.122	0.192 ± 0.105	0.051 ± 0.064	0.000 ± 0.000	0.167 ± 0.252	0.208 ± 0.354	0.500 ± 0.471
BF-EML	1.007 ± 0.170	0.577 ± 0.250	0.385 ± 0.145	0.346 ± 0.342	0.458 ± 0.305	0.708 ± 0.415	0.750 ± 0.388	0.833 ± 0.356
CST	1.045 ± 0.351	0.504 ± 0.197	0.266 ± 0.103	0.211 ± 0.184	0.208 ± 0.248	0.375 ± 0.375	0.625 ± 0.452	0.750 ± 0.463
TSMNet	1.915 ± 2.273	0.713 ± 0.170	0.509 ± 0.314	**0.403 ± 0.300**	0.667 ± 0.471	0.792 ± 0.396	0.875 ± 0.354	**1.000 ± 0.000**
RIVER	**3.178 ± 4.402**	0.691 ± 0.202	0.505 ± 0.284	0.399 ± 0.269	0.667 ± 0.436	**0.833 ± 0.356**	0.875 ± 0.248	**1.000 ± 0.000**

Table H.3 and Figure H.2 summarize the 60s localization results. RIVER has the highest enrichment (3.178) but also the largest patient variability, while BIOT leads SOZ AUROC (0.714) and AP (0.515), and TSMNet slightly leads Recall@K (0.403). Although mean enrichment exceeds one for several models, none of the model-wise enrichment tests remains statistically significant after the joint Holm correction across 12s and 60s. Across the two window lengths, no model dominates every localization criterion, and the longer window length can amplify SOZ concentration for a

subset of methods without producing uniformly more stable localization. Several models also reach perfect Hit@10 despite weaker global ranking metrics, reinforcing that coarse SOZ inclusion is easier than precise contact ordering.

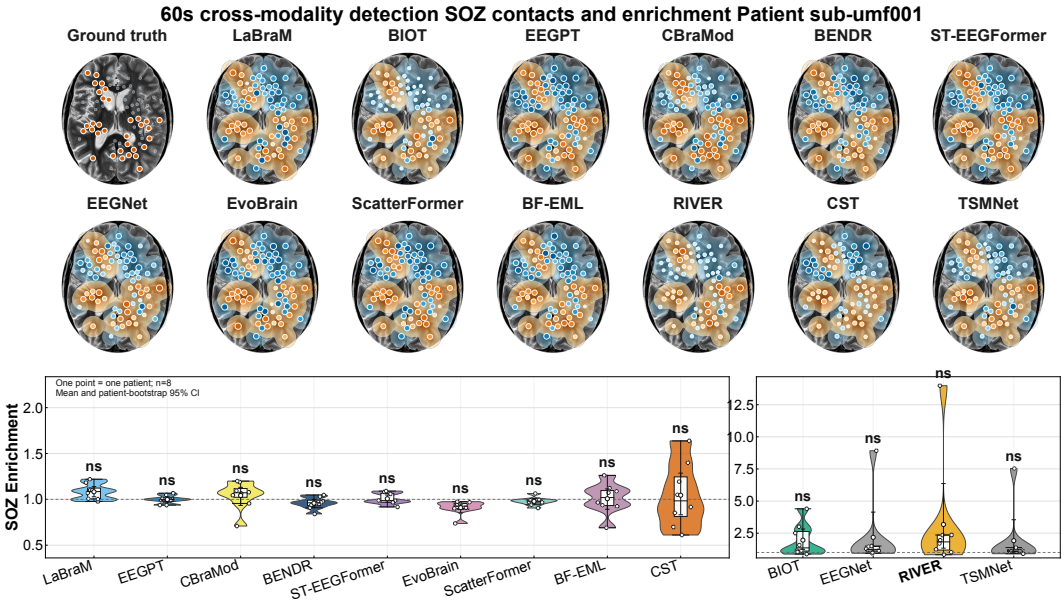

Figure H.2: SOZ evidence for 60s zero-shot cross-modality detection. For readability, models with larger enrichment ranges are shown in a separate panel with a different y-axis scale. The contact maps compare clinically annotated SOZ contacts with native contact evidence for patient sub-umf001 across neural models. The violin plots summarize SOZ enrichment across $n = 8$ patients with the mean and patient bootstrap 95% confidence interval; the dashed line at one denotes chance enrichment. Enrichment significance uses one-sided Wilcoxon tests at the patient level after seed averaging with one Holm family across both window lengths. Significance: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, and ns.