Clinically-Guided Test-Time Incorporation EEG Localization into a Fixed Segmentation Model for Focal Cortical Dysplasia

Anonymized Authors

Anonymized Affiliations
email@anonymized.com

Abstract. Medical imaging models are typically trained once and deployed statically, yet the clinical information available at inference time, including electrophysiological localization, imaging protocols, and reader annotations, continues to evolve after training. We ask whether a fixed 3D nnU-Net segmentation model for focal cortical dysplasia (FCD) can be adapted at inference time to incorporate newly available clinical context, specifically EEG-derived seizure localization, without any retraining. We present a structural-functional integration framework that layers EEG localization, at multiple anatomical granularities, onto a frozen nnU-Net's raw probability output using a ratio-weighted, distance-decayed multi-peak fusion mechanism and a soft, confidence-aware out-of-zone penalty. Across 257 evaluable cases spanning four public cohorts, this inference-time adaptation recovered 74.7% of predictions under a strict signal-above-background criterion, and raised recovery of complete false negatives from 30.8% to 48.7% under multi-peak search. We further show that recovery improves with finer localization granularity, with statistical significance against baseline reached at lobar ($p < 0.01$) and subregional ($p < 0.001$) restriction but not at coarse hemisphere-level restriction alone, and that soft out-of-zone suppression outperforms hard masking by preserving confident predictions when clinical localization is incomplete. These results suggest that clinically-guided, training-free adaptation is a practical strategy for keeping deployed medical segmentation models aligned with evolving clinical context.

Keywords: continual learning · test-time adaptation · focal cortical dysplasia · multimodal medical imaging · EEG-guided segmentation

---

This submission shares an underlying dataset, baseline segmentation model, and core localization pipeline with a concurrently submitted workshop paper at a different venue. That paper analyzes anatomical-localization granularity and out-of-zone-penalty behavior in a static evaluation setting; this paper reframes the same mechanism as a training-free, inference-time adaptation layer and emphasizes its relevance to continually evolving clinical annotations. This overlap is disclosed in the interest of transparency, consistent with double-blind review norms, without identifying authorship.

1 Introduction

Clinical practice is inherently dynamic: patient populations shift, imaging protocols evolve, and the clinical context available for a given case, including electrophysiological findings, changes over time. Most medical segmentation models, in contrast, are trained once under static assumptions and remain fixed after deployment, so any new source of clinical information that becomes available at inference time is left unused unless the model is retrained. This work asks whether such a fixed model can instead be adapted at inference time, without weight updates, to incorporate a clinically important but previously unused signal.

We note that our use of "adaptation" here refers specifically to a deterministic, hand-specified fusion rule applied identically at every inference call, not to an online-updating or streaming test-time adaptation procedure in the sense typically studied in the continual-learning literature (e.g., entropy-minimization or batch-statistic adaptation under distribution shift). We frame our contribution accordingly in Section 1, and see genuinely online adaptation to streaming EEG annotations as important future work rather than a claim of this paper.

We study this question in the context of focal cortical dysplasia (FCD), one of the most common structural causes of treatment-resistant epilepsy, where a substantial proportion of pathologically confirmed cases show no visible abnormality on conventional MRI. The 3D nnU-Net is a strong, self-configuring reference architecture for FCD segmentation, achieving an F1 score of 0.58 and a detection rate of 55% [1], yet baseline evaluations show that roughly 40% of false negatives arise from this architecture [1]. Retraining a new model each time additional clinical information becomes available, such as EEG-derived seizure localization from a neuroradiologist's report, is costly and disconnected from how clinicians actually work: experienced readers narrow their visual search using EEG findings without altering their underlying diagnostic model. We ask whether this narrowing-of-search behavior can be encoded as a lightweight, training-free adaptation layer applied directly to a frozen model's output.

We present a structural-functional integration framework that treats a trained 3D nnU-Net as fixed and adapts its use at inference time by fusing its raw, subthresholded probability map with EEG-derived anatomical localization at multiple granularities, a distance-decayed multi-peak search, and a soft, confidence-aware penalty for predictions outside the clinically indicated region. Because this adaptation happens entirely at inference time and can absorb new clinical annotations as they arrive, without retraining or access to the original training data, it is naturally compatible with continual, evolving clinical workflows, though we emphasize this compatibility is demonstrated here on a static, retrospective cohort rather than a genuinely streaming deployment. Our contributions are: (1) a training-free, EEG-guided adaptation layer for a frozen FCD segmentation model, evaluated on a static retrospective cohort as a proxy for evolving clinical annotation; (2) a systematic study of how the granularity of clinical localization, from hemisphere to lobe to subregion, affects recovery of missed lesions; and (3) a soft-penalty adaptation rule that preserves confident predictions under

Table 1. Number of cases per dataset.

Dataset	Number of cases
Bonn	78
IDEAS	32
MR-RATE	125
UCSF-PDGM	27

incomplete or evolving clinical localization, in contrast to a hard, non-adaptive exclusion rule.

2 Materials and Methods

2.1 Datasets and Baseline Model

Four publicly available cohorts were analyzed: the Bonn Open Presurgery MRI Dataset [2], the Imaging Database for Epilepsy and Surgery (IDEAS) [3], the MR-RATE dataset[1], and the UCSF Preoperative Diffuse Glioma MRI (UCSF-PDGM) dataset [4], totaling 257 cases (Table 1). Of 262 total publicly available cases across these four cohorts, 5 were excluded due to incompatible data, leaving the 257 cases analyzed below. All four cohorts are publicly available, de-identified imaging datasets; no new human subjects data were collected for this study. A 3D full-resolution nnU-Net was trained on FLAIR alone with five-fold cross-validation, a patch size of $160 \times 160 \times 96$, and a differentiable pseudo-Dice loss; 112 healthy-control scans with all-zero masks were added to improve specificity. This model is treated as a fixed, frozen component throughout; the adaptation framework described below operates only on its output probability map and never updates its weights.

2.2 Inference-Time Clinical Adaptation

SynthSeg parcellation mapped each case into DKT atlas space [5], grouped into six lobar regions per hemisphere. The EEG-indicated seizure-onset region, drawn from clinical annotation, was expanded by a 10 mm margin to define a localization zone, evaluated at three granularities: whole-hemisphere, lobar, and subregional (native DKT parcel). To quantify lesion visibility for downstream analysis, cases were also scored on a 0–10 conspicuity scale (Table 2) by consensus of two readers, with an independent neuroradiologist resolving discrepant cases.

[1] The MR-RATE dataset does not yet have an associated peer-reviewed publication at the time of writing; it is publicly available from its originating consortium and is cited here by name. We will update this citation with a formal reference upon publication.

Table 2. Focal cortical dysplasia (FCD) conspicuity scoring scale.

Score	Description
10 – Maximal conspicuity	The FCD is readily visible; even an untrained person with no medical training could identify it with minimal guidance.
5 – Moderate conspicuity	The FCD would be detected by most radiologists using a standard search strategy and is readily identifiable to most neuroradiologists and epilepsy-imaging experts.
1 – Minimal conspicuity	The FCD is extremely subtle, generally not detectable without multimodal information (e.g., EEG, PET) that restricts the search region.
0 – Non-detectable	The FCD is considered non-identifiable on MRI if it cannot be detected by two neuroradiologists, allowing substitution of one reviewer by an epilepsy-imaging specialist, even when its precise location is disclosed.

Within the localization zone, the top-k local probability maxima ($k \in \{1, 3, 5\}$) were identified and combined via a ratio-weighted, distance-decayed fusion ("Synergy Boost"): for a peak A boosted by a neighboring peak B,

$$\text{Composite}_A = P_A \times \left(1 + \frac{P_B}{P_A} \times e^{-\text{Distance}/\sigma}\right), \qquad \sigma = 15.0 \text{ mm}, \tag{1}$$

so that spatially close weak peaks reinforce one another while isolated weak peaks are largely unaffected. Each boosted peak seeded a 6-connected flood-fill to recover a physically contiguous candidate region while discarding disconnected background noise. Outside the localization zone, a *soft penalty* raised the acceptance threshold to 0.90 rather than excluding predictions entirely (a *hard penalty*), allowing the adaptation to defer to the model's own confident output when the clinical localization signal is itself incomplete or imprecise – a proxy for the adaptation needing to remain robust as available clinical annotations evolve in completeness and reliability. Recovery was scored strictly (highest peak overlaps ground truth and exceeds ten times the 99.9th-percentile background probability) and leniently (any of the top-5 peak clusters overlaps ground truth), with granularity comparisons evaluated via McNemar's exact test; given the moderate size of the 66-case granularity cohort, these results are best read as evidence of a directionally consistent, statistically detectable effect rather than a precise effect-size estimate.

3 Results

3.1 Recovery of Complete False Negatives

The baseline model produced 43 complete false negatives (16%) across 257 valid cases, of which 39 were evaluable after excluding cases with missing localization or registration failures; these cases were disproportionately small (median

Table 3. Recovery of complete false negatives under inference-time EEG adaptation ($n = 39$).

Outcome	Top-1	Top-3	Top-5
Recovered, lenient	12 (30.8%)	19 (48.7%)	19 (48.7%)
Recovered, strict	12 (30.8%)	19 (48.7%)	19 (48.7%)

Table 4. Representative false-negative and faint-signal cases demonstrating multi-peak adaptation.

Case	Lobe	Baseline Dice	Top-1 Dice	Top-5 Dice	Peak rank at recovery	Strict?
MR-RATE_subtle-00237	frontal left	0.0000	0.3250	0.7997	2	Yes
Bonn_sub-00122	temporal right	0.0000	0.1022	0.5683	1	Yes
Bonn_sub-00080	frontal left	0.0000	0.3373	0.5544	2	Yes
IDEAS_subtle_182	temporal left	0.0000	0.0251	0.5173	N/A	Yes
Bonn_sub-00133	temporal right	0.0000	0.0131	0.1851	1	Yes
Bonn_sub-00095	frontal left	0.0000	0.0000	0.2079	2	Yes
MR-RATE_subtle-00293	frontal left	0.8641	0.0002	0.0005	N/A	No
MR-RATE_subtle-00186	frontal left	0.0000	0.0000	0.0000	Not recovered	No
Bonn_sub-00043	frontal right	0.0000	0.0000	0.0000	Not recovered	No

ground-truth volume 2251 voxels) and low in conspicuity. A single-peak adaptation, restricting the frozen model's output to the EEG-indicated lobe/hemisphere, recovered 12 of 39 (30.8%) of these misses. Expanding the adapted search to top-3 or top-5 peaks raised recovery to 19 of 39 (48.7%) under both lenient and strict criteria (Table 3), with all 19 recoveries occurring within the first three peaks (12 at rank 1, 5 at rank 2, 2 at rank 3), indicating that a top-3 adaptation budget is sufficient and non-lossy relative to top-5. Representative recovered and non-recovered cases are given in Table 4: they include MR-RATE_subtle-00237, adapted from a complete miss to Dice 0.7997 at peak rank 2, and Bonn_sub-00100, a 294-voxel, conspicuity-0 lesion that remained challenging even under adaptation, illustrating the range of case difficulty in this cohort.

Across the full 257-case cohort, single-peak adaptation recovered at least partial overlap in 227 cases (88.3%, lenient) and met the strict criterion in 192 cases (74.7%), with adapted Dice exceeding baseline in 146 cases (56.8%). Strict recovery and net Dice improvement declined as baseline detection quality rose (from 100% in the 0.1–0.2 Dice bin to 40% in the 0.9–1.0 bin), while lenient recovery stayed near 100% throughout, and a small number of high-baseline cases showed a net decrease (Table 5).

Among the 12 top-1 recoveries, peak probability was low on average (mean 0.040, median 0.0099), consistent with faint, genuinely present signal. Among the 27 non-recovered cases, the top peak had a higher mean probability (mean 0.363), and 6 of these had peak probability above 0.9 in the correct lobe/hemisphere (Table 6) – confident but anatomically misplaced predictions that clinical localization alone cannot correct. A complementary view (Table 7) shows faint-signal cases (peak probability < 0.5) that were nonetheless recovered under strict criteria as the adapted search expanded from top-1 to top-5.

Table 5. Recovery per baseline Dice bin under single-peak inference-time adaptation ($n = 257$).

Dice interval	n (valid)	Recovered, lenient	Recovered, strict	Dice improved
0.0 (complete miss)	39	12 (30.8%)	12 (30.8%)	12 (30.8%)
0.0–0.1	5	4 (80.0%)	4 (80.0%)	4 (80.0%)
0.1–0.2	7	7 (100%)	7 (100%)	7 (100%)
0.2–0.3	8	8 (100%)	7 (87.5%)	6 (75.0%)
0.3–0.4	6	6 (100%)	5 (83.3%)	5 (83.3%)
0.4–0.5	12	11 (91.7%)	10 (83.3%)	9 (75.0%)
0.5–0.6	14	14 (100%)	13 (92.9%)	12 (85.7%)
0.6–0.7	30	30 (100%)	29 (96.7%)	19 (63.3%)
0.7–0.8	56	56 (100%)	48 (85.7%)	34 (60.7%)
0.8–0.9	65	64 (98.5%)	51 (78.5%)	35 (53.8%)
0.9–1.0	15	15 (100%)	6 (40.0%)	3 (20.0%)
Total	257	227 (88.3%)	192 (74.7%)	146 (56.8%)

Table 6. False-negative cases with a high-confidence but non-overlapping peak prediction.

Case	Lobe	Peak probability	GT volume (voxels)
Bonn_sub-00043	frontal right	0.982	4591
Bonn_sub-00006	frontal left	0.9999	193
MR-RATE_subtle-00186	frontal left	0.999	2251
MR-RATE_subtle-00004	frontal right	0.998	0
MR-RATE_subtle-00145	temporal left	0.960	2448
MR-RATE_subtle-00320	frontal right	0.9999	1620

3.2 Granularity of Clinical Localization

On a 66-case challenging subset, selected via a keyword filter on clinical QC notes rather than a pre-registered inclusion criterion (a limitation we return to in Section 4), any spatial restriction improved recovery over the unrestricted baseline (Table 8), from 39.4% at baseline to 48.5% at whole-hemisphere granularity, 54.5% at lobar granularity (identical under hard- and soft-penalty out-of-zone handling), and 56.1% at subregional granularity. Paired McNemar's exact tests against baseline show that statistical significance strengthens, rather than saturates, with finer granularity: the whole-hemisphere improvement was not statistically significant ($p = 0.070$), lobar restriction reached significance at the 0.01 level ($p = 0.0063$ for both hard and soft penalty), and subregional restriction reached significance at the 0.001 level ($p = 0.00098$) (Table 8). This is the opposite pattern from what a purely descriptive reading of the recovery percentages might suggest, and it revises our earlier framing: coarse, hemisphere-level localization captures a directionally consistent but not statistically reliable share of the benefit in this cohort, while the more clinically demanding finer-grained

Table 7. Faint-signal recoveries (peak probability < 0.5) compared across top-1/top-3/top-5.

Case	Lobe	Std Dice	Top-1 peak prob	Top-1 Dice	Top-1 strict?	Top-3 Dice	Top-3 strict?	Top-5 Dice
IDEAS_FCD_83	frontal left	0.7388	2.87e-04	0.0102	Yes	0.0338	Yes	0.0379
IDEAS_subtle_182	temporal left	0.0000	1.91e-03	0.0251	Yes	0.2852	Yes	0.5173
MR-RATE_subtle-00134	frontal left	0.0000	1.06e-02	0.0123	Yes	0.0247	Yes	0.0348
MR-RATE_subtle-00237	frontal left	0.0000	2.02e-03	0.3250	Yes	0.6092	Yes	0.7997
MR-RATE_subtle-00293	frontal left	0.8641	2.07e-01	0.0002	No	0.0005	No	0.0005
MR-RATE_subtle-00337	frontal left	0.0000	2.22e-02	0.0711	Yes	0.2188	Yes	0.3790
UCSF-PDGM-0091	frontal left	0.8389	4.44e-03	0.0001	No	0.0031	No	0.0031

Table 8. Recovery (Dice > 0), mean Dice, and paired McNemar's exact test vs. baseline by clinical-localization granularity (66-case cohort).

Granularity	Recovered (Dice > 0)	Mean Dice	McNemar's exact p (vs. baseline)
Unrestricted baseline	26/66 (39.4%)	0.1415	—
Whole-hemisphere	32/66 (48.5%)	0.1579	$p = 0.070$ (n.s.)
Lobar, hard penalty	36/66 (54.5%)	0.1502	$p = 0.0063$ (**)
Lobar, soft penalty	36/66 (54.5%)	0.1589	$p = 0.0063$ (**)
Subregional	37/66 (56.1%)	0.1639	$p = 0.00098$ (***)

localization is what reaches conventional significance thresholds. We report this pattern as an open question for how much localization precision a continually adapted system actually requires from clinicians, rather than a settled claim that coarse localization suffices.

3.3 Soft versus Hard Adaptation of Out-of-Zone Confidence

Comparing hard exclusion against soft, confidence-aware suppression at the lobar granularity (66-case cohort), both recovered the same number of lesions (36, 54.5%), but the soft rule achieved higher mean Dice (0.1589 vs. 0.1502; Table 9) by preserving confident predictions that extended beyond the annotated lobar boundary. In one representative case (UCSF-PDGM-0342), hard exclusion truncated the lesion at the lobar boundary (Dice 0.0337), while soft suppression preserved the extending sub-regions (Dice 0.5933), showing that a non-exclusionary adaptation rule is more robust when clinical localization is imperfect.

Table 10 further quantifies this trade-off in terms of false-positive adaptation burden. Relative to the unadapted baseline, EEG-zone restriction with a lenient out-of-zone threshold (0.5) recovers more cases but at the highest false-positive cluster rate; the strict soft-penalty threshold (0.9) recovers the same number of cases while keeping the false-positive rate close to the hard-penalty configuration, supporting the soft penalty at threshold 0.9 as the more robust adaptation rule for evolving clinical annotation quality.

Table 9. Lesion recovery under hard vs. soft out-of-zone adaptation (66-case cohort).

Metric	Baseline	Hard penalty	Soft penalty
Cases recovered (Dice > 0)	26 (39.4%)	36 (54.5%)	36 (54.5%)
Mean Dice score	0.1415	0.1502	0.1589

Table 10. False-positive cluster burden by out-of-zone adaptation strategy ($n = 67$).

Configuration	Recovered	Total FP clusters	Avg. FP clusters/case
Unadapted baseline (no zone, thr = 0.5)	26 (39.4%)	29	0.44
Adapted, lenient out-of-zone thr = 0.5	36 (54.5%)	45	0.68
Adapted, strict (soft) out-of-zone thr = 0.9	36 (54.5%)	43	0.65
Adapted, hard penalty (out-of-zone excluded)	36 (54.5%)	30	0.45

4 Discussion

These results show that a frozen FCD segmentation model can be meaningfully improved at inference time by fusing it with an evolving source of clinical context, EEG localization, without retraining. Two distinct behaviors emerge under adaptation: recovery of faint, genuinely present sub-threshold signal, and cases with confident but anatomically misplaced predictions that clinical localization alone cannot correct. The granularity results are more nuanced than a simple "coarser is sufficient" story: recovery rate rises modestly and continuously across granularities, but statistical significance against baseline strengthens rather than saturates, reaching conventional thresholds only at lobar ($p < 0.01$) and subregional ($p < 0.001$) restriction, with whole-hemisphere restriction not reaching significance in this cohort ($p = 0.070$). We therefore do not claim that coarse localization suffices; rather, the data leave open how much localization precision a continually adapted system requires, and the soft-penalty results suggest that adaptation rules should remain revisable toward the model's own confidence rather than hard-coding an exclusion boundary, a property that matters as the completeness and reliability of clinical annotations continues to change over time and across sites. Because the adaptation layer operates purely on the model's output probability map, the same approach generalizes to any voxel-wise probabilistic segmentation model and any auxiliary clinical signal that can be mapped to an anatomical or spatial prior, which we view as a step toward medical vision systems that remain aligned with evolving clinical workflows without full retraining.

Limitations. This study evaluates adaptation on a static, retrospective cohort yet provides a framework to advance to processing streaming sequences

of arriving EEG annotations. We do not compare against established test-time-adaptation baselines from the continual-learning literature (e.g., entropy-minimization or batch-normalization-statistic adaptation); such a comparison is necessary to substantiate a continual-learning contribution beyond the present static evaluation and is left to future work. The 66-case granularity cohort was selected via a keyword-based filter on QC notes rather than a pre-registered criterion, which may bias the reported effect size relative to the full cohort. Per-lobe recovery-rate comparisons are based on small case counts and should be treated as suggestive rather than conclusive.

Ethics and data statement. All four cohorts analyzed are publicly available, de-identified imaging datasets; no new human subjects data were collected for this study.

Code availability. Code for the adaptation framework will be made available upon reasonable request, and we intend to release it publicly alongside a completed comparison against established test-time-adaptation baselines.

5 Conclusion

We introduced a training-free, EEG-guided adaptation framework that substantially improves recovery of missed FCD lesions from a fixed 3D nnU-Net by combining multi-granularity clinical localization, ratio-weighted distance-decayed peak fusion, and soft confidence-aware out-of-zone suppression. The framework recovers a majority of baseline detections under lenient criteria and nearly half of complete misses; recovery gains scale with localization granularity, reaching statistical significance at lobar and subregional restriction, while coarse hemisphere-level localization alone shows a positive but not statistically significant effect in this cohort. We view this as one instance of a broader paradigm, keeping deployed medical vision models useful under continually evolving clinical inputs by adapting how their output is used rather than retraining the model itself, and future work will extend this to directly incorporate raw, continuously streaming EEG signals, compare against established test-time-adaptation baselines, and formalize the approach within existing continual-learning evaluation protocols.

Acknowledgements. Omitted for double-blind review.

References

1. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
2. Schuch, F., et al.: An open presurgery MRI dataset of people with epilepsy and focal cortical dysplasia type II. Sci. Data **10**(1), 475 (2023)

3. Taylor, P.N., et al.: The Imaging Database for Epilepsy And Surgery (IDEAS). Epilepsia **66**(2), 471–481 (2025)
4. Calabrese, E., et al.: The University of California San Francisco Preoperative Diffuse Glioma MRI Dataset. Radiol. Artif. Intell. **4**(6), e220058 (2022)
5. Klein, A., Tourville, J.: 101 labeled brain images and a consistent human cortical labeling protocol. Front. Neurosci. **6**, 171 (2012)

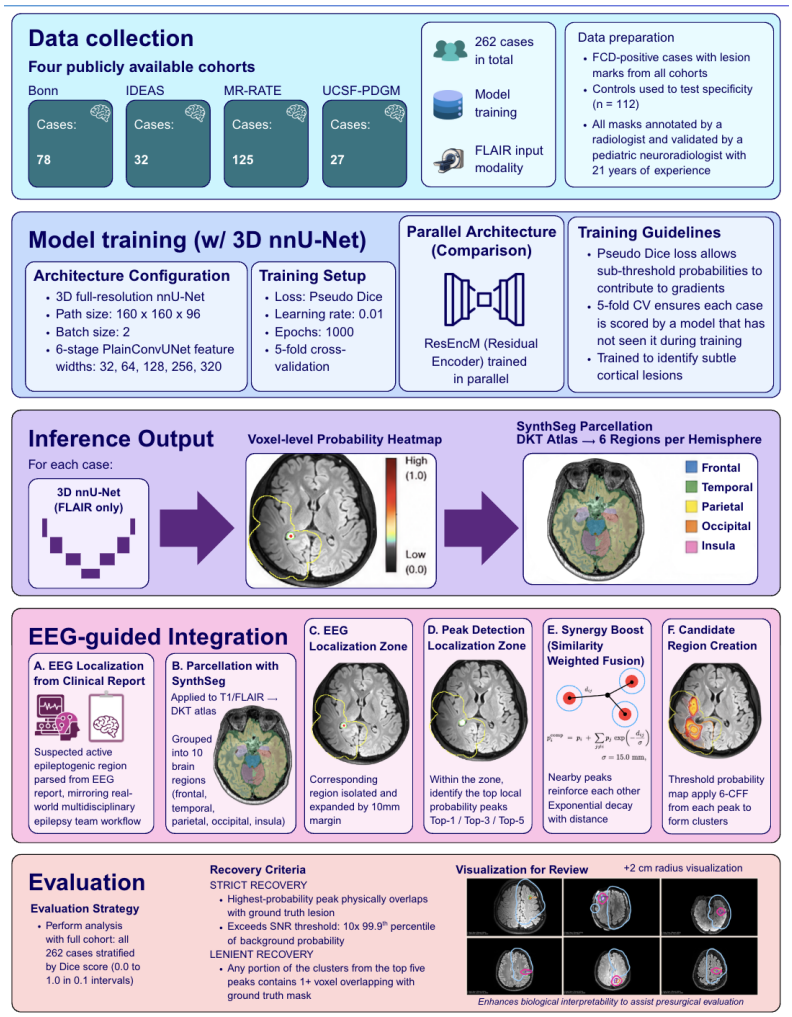

Fig. 1. Inference-time adaptation pipeline: a frozen nnU-Net probability map is fused, without retraining, with EEG-derived multi-granularity localization via ratio-weighted peak boosting, flood-fill candidate generation, and soft confidence-aware out-of-zone suppression.

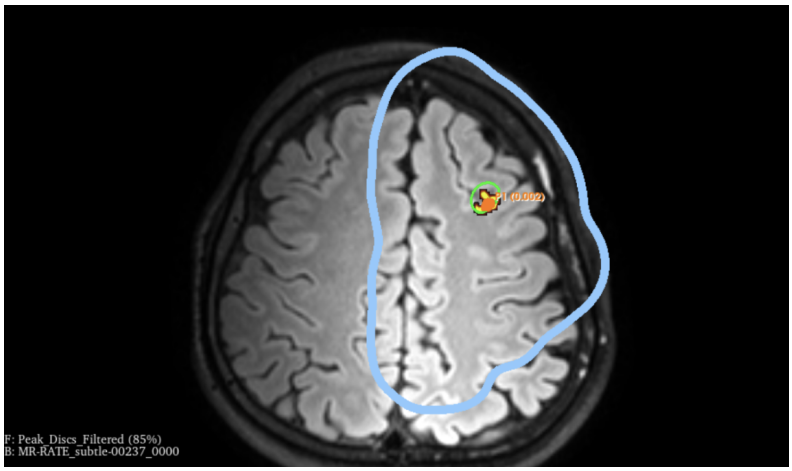

Fig. 2. Representative cases under inference-time adaptation: Bonn_sub-00100 (top), a 294-voxel conspicuity-0 lesion, and MR-RATE_subtle-00237 (bottom), recovered at peak rank 2 with the flood-fill candidate region overlapping the annotated lesion.

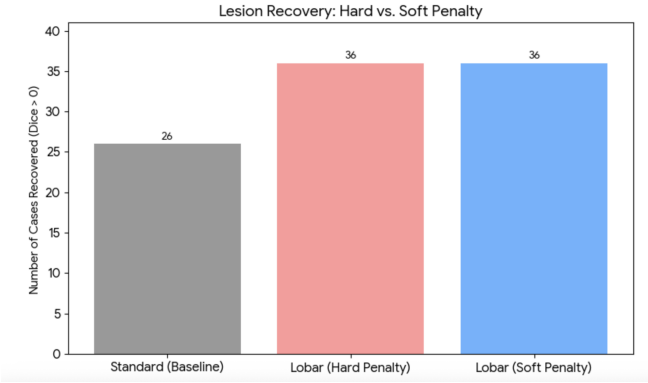

Fig. 3. Lesion recovery under hard versus soft out-of-zone adaptation (66-case cohort), corresponding to Table 9.

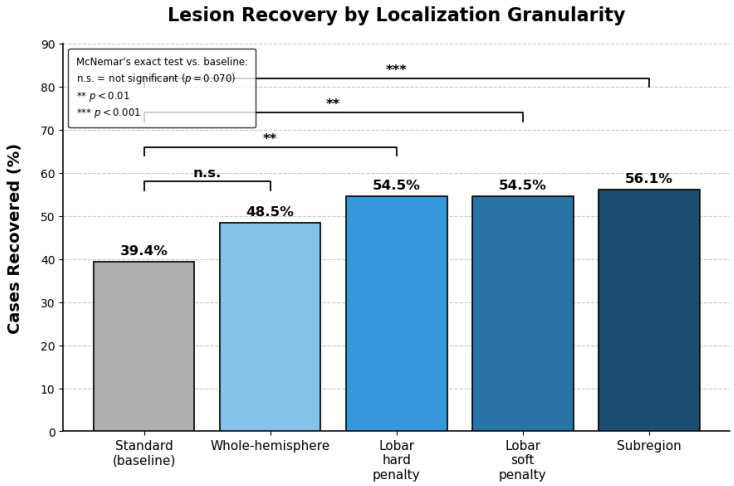

Fig. 4. Lesion recovery by localization granularity (66-case cohort), using the "Dice > 0" recovery criterion. Significance markers denote paired McNemar's exact test vs. baseline for each condition individually (n.s. = not significant, ** $p < 0.01$, *** $p < 0.001$), corresponding to Table 8.