TRACE: EPISODE-LOCAL TRANSITION EVIDENCE FOR ZERO-SHOT SEIZURE ONSET LOCALIZATION IN SEEG

Anonymous authors
Paper under double-blind review

ABSTRACT

In SEEG-based automated seizure analysis, accurate onset localization is important both for modeling seizure evolution and for supporting downstream clinical analysis. Existing methods face two key challenges: (1) detecting seizures reliably and localizing onset without priors such as patient identity, seizure type, or prior knowledge of seizure presence; and (2) generalizing across patients and centers despite substantial variation in background activity, seizure morphology, and acquisition conditions, without target-domain annotations, retraining, or parameter tuning. We propose TRACE, a framework based on episode-local transition evidence. TRACE encodes background-contrast features from each input clip into a shared channel–time evidence field that captures seizure-related changes relative to the episode-local background. A clip-level branch then summarizes that shared representation for seizure detection. When a clip is predicted as seizure-positive, the onset-localization branch selects high-evidence channels dynamically at each time point and constructs a global evidence trajectory by cross-channel pooling. A self-calibrated decoder reads that trajectory, identifies sustained seizure evidence, and backtracks to estimate the onset time. We evaluate our approach on SWEC-ETHZ and HUP under within-domain cross-patient settings. External zero-shot evaluation uses the patient-disjoint Pennsieve-530 cohort. On SWEC-ETHZ and HUP, our approach achieves oracle onset MAEs of 4.52 s and 4.92 s, reducing error by 37.7% and 34.0% relative to the strongest baseline in each cohort. On the spontaneous seizures of Pennsieve-530 (44 seizure and 132 non-seizure clips), our approach achieves a detection AUROC of 0.913 and an F1 score of 0.650, compared with 0.829 and 0.537 for the strongest baselines, while achieving an oracle onset MAE of 4.80 s, a 46.0% reduction relative to the strongest onset-localization baseline (8.89 s). These results show that constructing episode-local transition evidence and sparsely aggregating it provides an effective basis for cross-patient SEEG onset localization without target-domain adaptation.

1 INTRODUCTION

During pre-surgical evaluation of drug-resistant epilepsy, stereo-EEG (SEEG) must be reviewed for seizures. The time at which each seizure begins anchors the analysis that follows: which channels lead, how the activity spreads, and where the onset zone lies. Hand marking of onsets is slow, and even expert consensus on onset channels is imperfect (rater-to-rater $\kappa = 0.64$) (Ojemann et al., 2026). Automatic detection and onset localization could remove this bottleneck (Arcot Desai et al., 2023), and supervised onset estimation already reaches errors of a few seconds within a single implanted-device platform (Arcot Desai et al., 2023; Afzal et al., 2026).

A model trained on some patients must analyze patients it has never seen, recorded with other amplifiers, reference schemes, sampling rates, and electrode counts, and it receives no labels from the new patient. Within-patient performance rarely survives this deployment condition. Blind evaluation in the SzCORE 2025 challenge put the best event-based F1 at 32%, and self-reported scores overestimated independently evaluated ones (Dan et al., 2025). Learned EEG encoders can also absorb dataset identity (Zare, 2026). Absolute amplitude and spectrum of SEEG differ across patients and systems, whereas the change from a recording's own background at seizure onset is more stable. This is why referencing a signal to its own baseline is an old idea (Osorio et al., 1998; Raghu et al., 2020; He & Wu, 2020; Li et al., 2026). What has not been measured is whether a reference

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

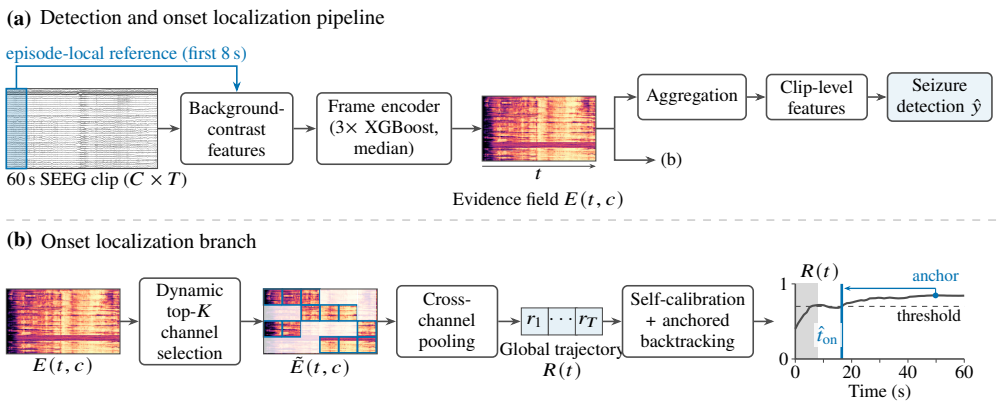


Figure 1: Overview of TRACE. (a) Each channel of a 60 s clip is referenced to the clip's own first 8 s (shaded); a three-classifier frame encoder maps the background-contrast descriptors to a channel–time evidence field $E(t, c)$, which a clip-level branch aggregates for detection. (b) The onset branch selects the top-K channels at each frame, pools them into one trajectory $R(t)$, calibrates it per clip, and backtracks from the most sustained evidence to the onset (schematic trajectory).

taken from the same episode is what makes onset localization transfer, under controls that change the reference and nothing else.

Our design starts from one observation: seizure onset is a transition away from the ongoing activity of the same episode, and the recordings before and after it differ in amplitude and spectral balance, with fast activity replacing the background (Bartolomei et al., 2008; Kumar et al., 2026). Localizing the onset then means measuring how far each moment departs from the episode's own background, rather than matching absolute ictal patterns. TRACE, the framework we present for seizure detection and onset localization, turns this observation into episode-local transition evidence (Figure 1). It references every channel of a 60 s clip to the clip's own first 8 s and turns the resulting background-contrast descriptors into a channel–time evidence field. It then aggregates the field for detection and reads the onset from a pooled evidence trajectory with a self-calibrated backtracking decoder.

We trained TRACE on two public cohorts and applied it unchanged to the spontaneous seizures of a third, consensus-annotated SEEG cohort. TRACE localized onsets with a mean absolute error (MAE) of 4.80 s, compared with 8.89 s for the strongest of eight baselines. When each missed seizure is charged the full clip duration, the error rises to 9.61 s, against 17.45 s for the strongest baseline under this metric. Across patients, the average gain over the strongest baseline was 3.65 s. TRACE also attained the highest detection AUROC, improving on the strongest baseline from 0.829 to 0.913. Controlled comparisons show what the reference contributes: without it TRACE reaches 7.54 s, with it 4.80 s. The same descriptors raised the onset error in all three transfer settings once we switched them from the episode-local reference to absolute scales. Ten alternative referencing schemes localized worse. References drawn from another episode, another patient, or another site degraded in that order.

This paper contributes the following.

- **TRACE**, which detects seizures and localizes onset by referencing each episode to its own background and needs no target-patient labels, retraining, or threshold adjustment (Section 3).
- **A zero-shot evaluation with patient-level inference** covering three public cohorts and eight reimplemented baselines, with frozen patient-level splits, frozen thresholds, and released per-clip predictions (Section 4).
- **Controlled evidence for the mechanism**: a matched reference switch, ten referencing controls, and a reference-identity ladder show that the episode-local reference, not normalization in general, supplies most of the oracle localization gain, and ablations show that the learned field and sparse selection also matter (Section 4.2.3).

2 Related work

Patient-independent seizure detection and onset localization. Single-patient detectors perform well on that patient (Burrello et al., 2020). Unseen-patient detection is much harder, and cross-cohort evaluation harder still (Dan et al., 2025). Patient-independent detectors include a CNN–Transformer assessed at six centers on scalp EEG and iEEG (Peh et al., 2023); for iEEG, a lightweight CNN (Si et al., 2023), a multi-task Transformer (Sun et al., 2025), and self-supervised anomaly detection with zero-shot transfer from HUP to SWEC-ETHZ (Yu, 2026). Hyperdimensional ensembles address the SWEC-ETHZ short-term recordings (Burrello et al., 2021). Spectral features with a regression readout (Chan et al., 2008), vision transformers on spectrograms (Afzal et al., 2026), partitioning of long-term EEG (Chen et al., 2025), and changepoint segmentation of ictal transitions (Kumar et al., 2026) all estimate the onset time. We reimplement six of these, plus a 1D-CNN (Wang et al., 2021) and an anomaly detector (Martini et al., 2021), and evaluate all eight zero-shot on a consensus-annotated cohort.

Referencing a signal to its own background. A quarter century ago, Osorio et al.'s foreground-to-background power ratio anchored implanted detectors (Osorio et al., 1998). Adaptive median baseline correction improves leave-one-database-out scalp detection (Raghu et al., 2020). Euclidean alignment is standard in cross-subject BCI (He & Wu, 2020). A pretrained model for patient-independent SEEG detection subtracts each channel's background to reduce domain shift (Yuan et al., 2023). A recent survey organizes such methods around one hypothesis: absolute signal characteristics vary across subjects and systems, while relative change is stable (Li et al., 2026). We make no claim to the idea. TRACE differs in where the reference comes from and in what we measure. The reference is the episode's own first seconds, not a dataset, patient, or running statistic. We test it with a matched switch, ten alternative schemes, and references taken from other episodes, patients, and sites.

Learned aggregation under shift. Whether a learned layer on top of fixed evidence transfers is contested: stacking can improve transportability across clinical databases (Yang et al., 2023), whereas EEG foundation models often fail to beat compact decoders when data are scarce (Yang et al., 2026), tree ensembles remain strong on tabular features (Grinsztajn et al., 2022), and estimated combination weights lose to simple averages (Chan & Pauwels, 2018). Appendix N takes up this question for an earlier configuration of the same evidence design.

3 Method

3.1 Seizure Detection and Onset Localization Problem Formulation

Seizure detection asks whether an intracranial EEG clip contains a seizure. Onset localization, on clips that are seizure-positive, estimates the onset time. Following the 60 s clip protocol of a prior study (Tang et al., 2022), we study both tasks on fixed multichannel clips in an offline setting. The framework takes a clip $\mathbf{X} \in \mathbb{R}^{C\times L}$, where C and L denote the numbers of channels and signal samples, respectively. It predicts a seizure label $\hat{y} \in \{0, 1\}$ and reports an onset estimate $\hat{t}_{\text{on}} \in [0, 60)$ s when $\hat{y} = 1$. At inference, the input is only the SEEG clip. Patient identity, seizure type, and seizure-presence labels are unavailable. We also consider zero-shot generalization to unseen patients and cohorts without target-domain annotations, retraining, or target-specific parameter tuning.

3.2 Background-Referenced Evidence Field Construction

Seizure activity spreads across recording channels (Craley et al., 2022). Deviations from channel-specific background have been used to characterize intracranial seizure activity (Gascoigne et al., 2024). TRACE builds a background-referenced evidence field that keeps channel resolution and temporal resolution. The encoder produces local evidence, and downstream aggregation and decoding set the clip-level decision and the onset time (Fig. 1a).

Episode-local reference. Each clip is divided into $T = 120$ non-overlapping frames of duration $\Delta = 0.5$ s, with frame index $t \in \{1, \ldots, T\}$. To reference every clip to its own pre-ictal background without labels or per-patient adjustment, we take the episode-local reference as the initial 8 s, corresponding to $\mathcal{B} = \{1, \ldots, 16\}$, and apply the same fixed interval to every clip. Appendix J audits the length, identity, position, and purity of this reference at test time.

The vector $\mathbf{b}_c$ holds the reference samples from channel c. We compute location and scale from this vector and normalize the signal as

$$m_c = \mathrm{median}(\mathbf{b}_c), \qquad s_c = \frac{Q_{75}(\mathbf{b}_c) - Q_{25}(\mathbf{b}_c)}{1.349}, \\ \tilde{x}_c[n] = \mathrm{clip}\left(\frac{x_c[n] - m_c}{s_c}, -8, 8\right), \tag{1}$$

where Q_p denotes the p-th percentile. We set $\tilde{x}_c[n] = 0$ when $s_c \leq 10^{-12}$, and the normalized signal supplies the background-scaled morphology descriptors.

Background-contrast feature extraction. Seizure onset in SEEG appears in amplitude and signal-complexity changes, in a redistribution of power toward fast activity, and in the persistence of these changes (Osorio et al., 1998; Burrello et al., 2020; Chan et al., 2008; Bartolomei et al., 2008). Each descriptor targets one of these signatures as a contrast to the episode-local reference (three persistence statistics threshold it at its whole-clip median), measuring change, not absolute level. We compute the log-domain deviation from the reference median for a nonnegative magnitude descriptor $d(t, c)$. We use a linear-domain contrast for a spectral-shape descriptor $q(t, c)$:

$$\phi_d^{\log}(t, c) = \log(1 + d(t, c)) - \mathrm{median}_{u \in \mathcal{B}} \log(1 + d(u, c)), \\ \phi_q^{\Delta}(t, c) = q(t, c) - \mathrm{median}_{u \in \mathcal{B}}\, q(u, c). \tag{2}$$

The magnitude descriptors include root-mean-square amplitude, line length, peak-to-peak amplitude, and 13–200 Hz power. The spectral descriptors include relative band powers, normalized spectral entropy, and a combined high-frequency relative-power measure. We also form temporal summaries of selected descriptor trajectories and line length on $\tilde{x}_c$.

The vector $\boldsymbol{\phi}(t, c) \in \mathbb{R}^{23}$ that results combines magnitude contrasts, background-scaled morphology, spectral redistribution, and temporal persistence. A clip-context entropy-persistence statistic is included as well: the fraction of frames in an 8 s trailing window whose entropy exceeds the channel's whole-clip median. Appendix C gives the complete descriptor definitions.

Frame-level encoding. We implement a shared encoder with three XGBoost classifiers (Chen & Guestrin, 2016) that use different random seeds and class-weighted binary logistic loss and that each map the descriptor vector $\boldsymbol{\phi}(t, c)$ to a post-onset probability. We derive training labels from the annotated onset and apply them to all channels. A frame whose center time s_t satisfies $s_t \geq t_{\mathrm{on}} + 1\,\mathrm{s}$ gets $z_t = 1$, and a frame with $s_t < t_{\mathrm{on}} - 2\,\mathrm{s}$ gets $z_t = 0$. Training drops the intervening interval. Seizure-negative clips supply background examples.

The encoder takes the median of the three predicted probabilities:

$$E(t, c) = f_\theta(\boldsymbol{\phi}(t, c)) = \underset{j \in \{1,2,3\}}{\mathrm{median}}\, p_{\theta_j}(z = 1 \mid \boldsymbol{\phi}(t, c)). \tag{3}$$

The parameters of classifier j are θ_j. The collected scores form $\mathbf{E} \in [0, 1]^{T \times C}$. When $E(t, c)$ is high, the channel's background-referenced descriptor matches the learned post-onset pattern. Detection and localization share this field. Localization keeps the temporal structure throughout decoding.

3.3 Seizure Detection by Clip-Level Aggregation

The detection branch decides whether an onset estimate should be reported. It summarizes $\mathbf{E}$ with three clip-level features: the late-minus-early change of the channel-averaged evidence, the same change restricted to the channels that change most, and the spread of the late evidence. It also adds a one-sided Page CUSUM (Page, 1954) computed on the line-length and high-frequency descriptor trajectories before encoding. Appendix D defines the four features z_{diff}, z_{top}, z_{cu}, and z_{var}.

A shallow XGBoost head, after training-set feature standardization, maps the vector $\mathbf{z}_{\mathrm{det}} = [z_{\mathrm{diff}}, z_{\mathrm{top}}, z_{\mathrm{cu}}, z_{\mathrm{var}}]^\top$ to a seizure probability:

$$p_{\mathrm{sz}} = g_\psi(\mathbf{z}_{\mathrm{det}}), \qquad \hat{y} = \mathbb{I}[p_{\mathrm{sz}} \geq \tau_{\mathrm{det}}]. \tag{4}$$

We train the head on patient-level out-of-fold encoder predictions. We select the operating threshold from out-of-fold detection probabilities with a composite objective that combines F1, sensitivity, specificity, and an onset score that penalizes missed seizures. Appendix B details the objective and classifier settings.

3.4 Dynamic Channel Aggregation for Onset Localization

Onset localization needs a trajectory that keeps temporal evidence while the contributing channels change. We therefore stabilize each channel's evidence over a local window before selecting and pooling the strongest channels, so that channels are picked from temporally supported evidence rather than isolated frame responses (Fig. 1b).

Dynamic Top-K Channel Selection. The forward 8 s window that starts at frame t is $\mathcal{W}_t$, and a full window has 16 frames. Stabilized channel scores are computed as

$$\begin{aligned} M(t,c) &= \mathrm{median}_{u\in\mathcal{W}_t}\, E(u,c), \\ \mathcal{S}_t &= \mathrm{TopK}_c\, M(t,c), \qquad K = \min(5, C). \end{aligned} \tag{5}$$

The operator TopK returns the indices of the K largest scores. Recomputing $\mathcal{S}_t$ at every frame lets different channels support the trajectory as the evidence evolves. The selection is local in time, unlike the once-per-clip selection in detection (Appendix D).

The selected view in Fig. 1b, $\widetilde{E}(t,c) = M(t,c)\,\mathbb{I}[c \in \mathcal{S}_t]$, is a deterministic transformation of $\mathbf{E}$, not an additional learned representation.

Cross-Channel Pooling. Averaging the selected scores yields the global evidence trajectory:

$$R(t) = \frac{1}{K}\sum_{c\in\mathcal{S}_t} M(t,c) = \frac{1}{K}\sum_{c=1}^{C} \widetilde{E}(t,c). \tag{6}$$

Pooling drops the channel dimension and keeps time. Each value summarizes support from multiple locally informative channels and does not require a fixed channel subset throughout the clip. When $C < 5$, the method uses all available channels, and it applies no additional smoothing. The onset decoder then reads $R(t)$.

3.5 Adaptive Onset Decoding

Strong evidence marks an established seizure but does not name its onset. The decoder therefore normalizes the trajectory within each clip, picks a sustained high-evidence anchor, and backtracks to the onset of its connected evidence region, keeping confirmation of the seizure separate from estimation of when it began.

Per-Clip Self-Calibration. Let Q_p be the p-th percentile of the complete trajectory $\{R(t)\}_{t=1}^{T}$. We set the lower and upper reference levels by

$$(l,h) = \begin{cases} (Q_{20}, Q_{80}), & Q_{80} - Q_{20} \geq 0.05, \\ (\min_t R(t), \max_t R(t)), & \text{otherwise}, \end{cases} \tag{7}$$

and compute

$$R_{\mathrm{cal}}(t) = \mathrm{clip}\left(\frac{R(t) - l}{\max(h - l, 10^{-6})}, 0, 1\right). \tag{8}$$

The rule is fixed: it rescales the trajectory relative to the clip and uses no labels or model updates.

A forward moving minimum then measures sustained evidence:

$$R_{\mathrm{sus}}(t) = \min_{u\in\mathcal{H}_t} R_{\mathrm{cal}}(u), \qquad \mathcal{H}_t = \{t, \ldots, \min(t+5, T)\}. \tag{9}$$

A complete window lasts 3 s, and the clip boundary truncates it. A frame meets the sustained-evidence criterion when $R_{\mathrm{sus}}(t) \geq \eta$, with $\eta = 0.5$.

Anchored Backtracking. When $\max_t R_{\mathrm{sus}}(t) \geq \eta$, the anchor is the maximum of the sustained trajectory:

$$a = \arg\max_t R_{\mathrm{sus}}(t). \tag{10}$$

We start at a and walk R_{sus} backward. A frame at or above η resets the sub-threshold gap count. The search tolerates up to four consecutive sub-threshold frames (2 s) and stops at the fifth consecutive

Table 1: Cohorts (60 s clips). SWEC-ETHZ uses strip, grid, and depth electrodes; HUP and Pennsieve-530 use SEEG. Splits are by patient; parentheses give patient counts. Pennsieve-530 is used zero-shot only, and its 34 stimulation-induced seizures are excluded.

Cohort	Patients	Seizure clips	Non-seizure clips	Train/val/test (patients)	Test seizures (patients)
SWEC-ETHZ	16	100	100	10 / 3 / 3	12 (3)
HUP	36	132	284	25 / 6 / 5	20 (5)
Pennsieve-530	29	44	132	zero-shot only	44 (22)

sub-threshold frame or the beginning of the clip. The earliest frame with $R_{\mathrm{sus}}(t) \geq \eta$ that the search reaches before it stops is $\hat{j}$. The onset estimate is the center time of that frame,

$$\hat{t}_{\mathrm{on}} = (\hat{j} - \tfrac{1}{2})\Delta. \tag{11}$$

The decoder returns a fixed fallback estimate of 25 s if no frame meets the sustained-evidence criterion.

The decoder uses the full clip and forward-looking windows, consistent with offline onset localization. It reports an onset estimate only when $\hat{y} = 1$. The learned encoder, detection head, and decision rules stay fixed across patients and centers. Only the reference and trajectory statistics are recomputed inside each input clip.

4 Experiments

4.1 Experimental Setup

Datasets and splits. We study *spontaneous seizures only*; stimulation-induced events are outside the scope of this work and excluded from all analyses. Table 1 summarizes the three cohorts: SWEC-ETHZ (Burrello et al., 2020) and HUP (Bernabei et al., 2023) for development and within-cohort evaluation, and the Multi-Expert Seizure Annotation dataset (Pennsieve-530) (Ojemann et al., 2026) for external evaluation only. Development cohorts are partitioned by patient into training, validation, and test sets with no patient in more than one split; clip construction and per-cohort accounting are detailed in Appendix A.

Data preprocessing. Each recording is notch-filtered and resampled to a common rate and cut or padded into fixed-length multichannel clips by a deterministic, patient-agnostic chain; Appendix A specifies every setting. Seizure onsets are positioned randomly within each clip rather than at its midpoint, so a positional prior is weak: a constant 25 s guess scores 8.89 s on Pennsieve-530.

For zero-shot evaluation, TRACE trained on SWEC-ETHZ+HUP was applied unchanged to Pennsieve-530, with configuration, thresholds, and decoder rules frozen before its first output on this cohort and no target-domain training or tuning; an earlier configuration had been evaluated on the cohort (Appendix M). External patients are disjoint from the development cohorts, but 19 of 29 were recorded at HUP's institution; we term this *zero-shot cross-cohort generalization*.

Baselines. We compare with eight reimplemented published approaches: the probability-based deep detectors 1D-CNN (Wang et al., 2021), LightWeightCNN (Si et al., 2023), CNN–Transformer (Peh et al., 2023), and MT-Transformer (Sun et al., 2025); the anomaly-based Martini2021 (Martini et al., 2021) and the partition-based SODor (Chen et al., 2025); the spectrogram vision transformer ViT-SOD (Afzal et al., 2026); and the spectral-regression SVM-Reg (Chan et al., 2008). All methods share the patient-level splits, source-only model selection, and frozen thresholds (TRACE takes its thresholds from training-fold out-of-fold predictions, the baselines from validation patients), and the four probability-based detectors additionally share one post-hoc onset readout whose constants are selected on source validation patients only. Appendix E tabulates each baseline's readout, threshold provenance, and no-valid-onset behavior, and Appendix M records the remaining shared protocol (input scaling, seeds, epoch budgets).

Table 2: Cross-patient onset localization on the frozen patient-disjoint test sets (seconds; oracle and deployed MAE as in Section 4.1; baselines seed 42).

	Oracle MAE ↓		Deployed MAE ↓	
Method	SWEC-ETHZ	HUP	SWEC-ETHZ	HUP
1D-CNN	16.40	15.34	34.19	32.92
LightWeightCNN	12.24	12.85	12.24	27.99
CNN–Transformer	19.61	16.43	21.44	29.18
MT-Transformer	7.26	22.02	7.26	35.69
Martini2021	20.68	25.04	20.68	42.66
SODor	23.46	21.69	28.65	29.93
ViT-SOD	7.54	8.78	7.54	32.24
SVM-Reg	9.86	7.45	17.94	40.32
TRACE	**4.52**	**4.92**	**4.52**	**14.41**

Evaluation protocol. Localization is evaluated by two co-primary metrics: an oracle onset MAE over all seizure clips, using the decoder output whether or not the clip is detected, and a deployed MAE, which charges a missed seizure the full 60 s clip length and therefore couples localization to detection. Detection is evaluated by the auxiliary metrics AUROC and F1 at source-frozen thresholds. Statistical comparisons use patients as the sampling unit: clip errors are first averaged within patient, and paired 95% confidence intervals are obtained from patient-level bootstrap resampling.

Implementation details. All TRACE results use one frozen configuration (8 s episode-local reference, 0.5 s frames, top-5 mean pooling), with thresholds and decoder constants selected on source data only; within-cohort results use models fit on that cohort's training patients. Appendix B gives the ensemble construction, tree settings, threshold objective, seeds, and hardware.

4.2 Experimental Results

4.2.1 Cross-Patient Onset Localization

Table 2 evaluates onset localization on held-out patients from the two development cohorts. TRACE achieves an oracle onset MAE of 4.52 s on SWEC-ETHZ and 4.92 s on HUP, reducing error by 37.7% and 34.0% relative to the lowest-MAE published baseline in each cohort. Under the deployed metric the advantage is unchanged on SWEC-ETHZ, where TRACE misses no seizure, and grows on HUP: 14.41 s against 27.99 s for LightWeightCNN, a 48.5% reduction. The oracle result is stable across training seeds: the five-seed MAEs are 4.46 ± 0.07 s and 4.84 ± 0.08 s.

4.2.2 Zero-Shot Cross-Cohort Generalization

We next evaluate the frozen system on spontaneous seizures from Pennsieve-530 (Table 3). TRACE achieves an oracle onset MAE of 4.80 s and a deployed MAE of 9.61 s. The lowest seed-averaged baseline oracle MAE is 8.89 s (ViT-SOD), no better than the constant 25 s guess; only TRACE is below that error. The deployed metric reorders the baselines: ViT-SOD degrades to 24.12 s and MT-Transformer gives the lowest value (17.45 s), so TRACE's 9.61 s is a 44.9% reduction under this metric as well. On the auxiliary detection metrics, the same shared representation attains AUROC 0.913 and F1 0.650; the highest seed-averaged baseline values are 0.829 and 0.537, respectively. The learned head is not the best ranker of this evidence: a fixed fusion of the same signals ranks better under transfer between the development cohorts ($+0.05$ AUROC) and ties it on Pennsieve-530 (Appendix N). Figure 2 walks through the frozen pipeline on one of these seizures, from the evidence field to the decoded onset.

Several baseline implementations signal a failed onset with a sentinel timestamp. The matched-fallback analysis (Appendix F) gives every such readout the 25 s default that the TRACE decoder uses, which TRACE never triggers on these seizures. TRACE stays at 4.80 s and the lowest baseline MAE stays at 8.89 s (ViT-SOD, which never emits the sentinel). After Bonferroni correction the patient-level gain is significant against seven of eight baselines; against ViT-SOD it is 3.65 s (uncorrected 95% CI 0.80–6.21) and not significant.

Table 3: Zero-shot cross-cohort evaluation on Pennsieve-530 using 44 spontaneous seizure clips and 132 non-seizure clips. Baselines report mean±SD over three seeds; TRACE reports the designated frozen configuration.

Method	Oracle MAE (s) ↓	Deployed MAE (s) ↓	AUROC ↑	F1 ↑
ViT-SOD	8.89 ± 0.50	24.12 ± 4.90	0.660 ± 0.027	0.453 ± 0.032
SVM-Reg	10.28 ± 0.00	18.65 ± 0.00	0.756 ± 0.000	0.471 ± 0.000
CNN–Transformer	13.76 ± 1.35	26.33 ± 4.98	0.779 ± 0.024	0.537 ± 0.041
LightWeightCNN	14.79 ± 0.89	20.74 ± 4.86	0.783 ± 0.048	0.531 ± 0.044
1D-CNN	14.95 ± 1.97	41.92 ± 4.11	0.742 ± 0.024	0.442 ± 0.051
MT-Transformer	16.26 ± 2.67	17.45 ± 2.86	0.829 ± 0.045	0.500 ± 0.035
Martini2021	22.52 ± 0.55	27.43 ± 2.34	0.447 ± 0.026	0.380 ± 0.011
SODor	23.03 ± 1.56	34.56 ± 7.70	0.702 ± 0.012	0.471 ± 0.015
TRACE	**4.80**	**9.61**	**0.913**	**0.650**

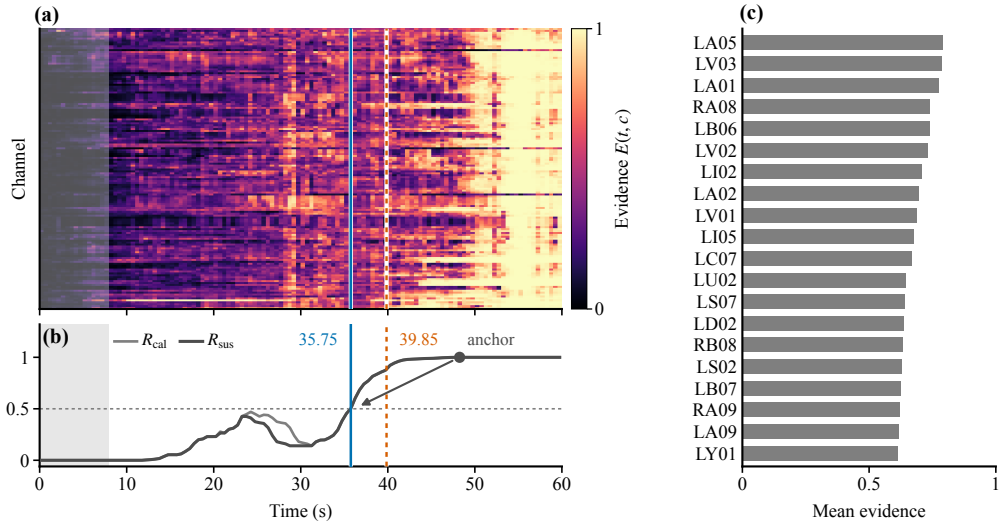


Figure 2: TRACE on a zero-shot Pennsieve-530 seizure with a typical error (clip psv_000203, spontaneous, 146 channels; absolute error 4.10 s against a median of 3.97 s over the 44 seizures). (a) Evidence field $E(t, c)$ over all channels; the grey band is the 8 s episode-local reference. (b) Calibrated trajectory $R_{\text{cal}}(t)$ and sustained trajectory $R_{\text{sus}}(t)$; anchored backtracking from the anchor (48.25 s) returns the decoded onset at 35.75 s (solid) against the consensus onset at 39.85 s (dashed). (c) The 20 channels with the highest mean evidence in the 8 s after the decoded onset.

4.2.3 Background-Referenced Evidence Improves Transfer

Referencing descriptors to the episode's own background lowers onset error in every setting tested (Figure 3a). The matched control keeps the encoder family, training budget, aggregation, and decoder and removes only the episode-local contrast.

On Pennsieve-530, removing the reference raises clip-level MAE from 4.80 to 7.54 s; averaged within patient first, the paired improvement is 3.62 s (95% CI 1.32–6.04). Across five retrained seeds the episode-local representation yields 4.54–4.65 s against 7.17–7.73 s without it; the designated run reported throughout (4.80 s) is the least favorable, so the headline figures are conservative. Nine further normalization and rereferencing controls all localize worse, each with a patient-level interval above zero (Figure 3b); the closest reaches 6.08 s, 1.70 s behind TRACE at the patient level.

We further intervene on the identity of the reference while keeping the trained model fixed. Taking the reference from another seizure of the same patient, same-patient interictal activity, another patient, or another site raises MAE step by step (6.70, 8.04, 9.76, 12.28 s). The useful reference is therefore not a patient-level normalization statistic but the background of the current episode. At test time

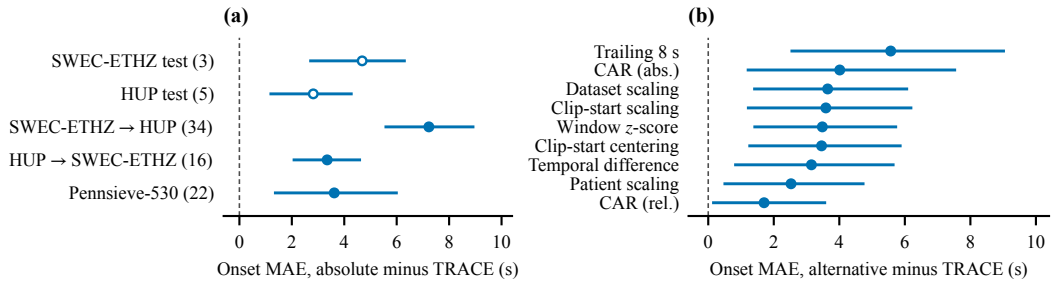


Figure 3: The episode-local reference localizes better than absolute descriptors in all five settings (a) and better than each of nine alternative references on Pennsieve-530 (b; 22 patients; CAR: common-average reference). Oracle MAE difference averaged within patient; bars are patient-level paired bootstrap 95% CIs; open points rest on 3 or 5 patients. Values in Appendix I.

(Appendix J), 8 and 16 s references perform comparably on the development test sets and 4 s degrades; on Pennsieve-530, windows starting at 16 s or later raise MAE to 8.02–8.68 s, and ictal mixing is tolerated up to 30% (9.00 s at 50%).

4.2.4 Which Components Carry the Transfer

Beyond the reference, the learned field, its channel–time structure, and sparse selection matter, while dynamic re-selection adds little (Appendix H); every variant is built from development data only and applied unchanged to Pennsieve-530 (Table 7). Predicting onset directly without an evidence field reaches 6.87 s at best (clip-level XGBoost regression), and replacing the learned field with the unlearned signal track $u_c(t)$ raises MAE to 13.08 s. Collapsing the channel dimension gives 8.54 s, and collapsing the temporal dimension reduces the decoder to its constant fallback (8.89 s). Removing sparse channel selection costs 3.13 s (7.93 s), whereas a static top-5 set in place of dynamic re-selection costs 0.20 s (5.00 s). On out-of-fold training predictions, no other pool size or pooling rule improves significantly on the frozen $K{=}5$ mean (Appendix G). In the decoder (Figure 2b; Appendix K), replacing anchored backtracking with first crossing raises MAE from 4.80 to 5.71 s, and the two rules agree on 42 of 44 seizures. No ablation in Table 7 improves on the full frozen framework (4.80 s).

5 Conclusion

We presented TRACE, which detects seizures and localizes their onset from evidence referenced to each episode's own background. Trained on two public cohorts and applied unchanged to the spontaneous seizures of a third, it localized onsets with an error of 4.80 s against 8.89 s for the strongest of eight baselines. A matched reference switch, ten alternative referencing schemes, and references drawn from other episodes, patients, and sites show that, under oracle MAE, the episode-local reference supplies most of this gain, and single-pass ablations indicate that the learned evidence field and sparse channel selection also matter.

The zero-shot result rests on 44 spontaneous seizures from 22 patients. Nineteen of its 29 patients come from HUP's institution; an earlier configuration had been evaluated on the cohort before TRACE was frozen, and the reference-switch onset contrast was specified after those results (Appendix M). Under the deployed metric the reference-switch gain on Pennsieve-530 shrinks to 3.08 s, with an interval that includes zero (-0.21 to 6.44). Stimulation-induced seizures violate the assumption of an endogenous transition: on the 34 excluded clips TRACE reaches 11.96 s against 7.94 s for ViT-SOD. Between the training cohorts the advantage over baselines holds in one direction (Appendix L). The baselines were reimplemented under fixed compute budgets, and TRACE reads offline 60 s clips whose annotated onset lies between 10 and 40 s.

Reproducibility statement

All cohorts are public. The supplementary material contains the cohort manifests and the per-clip predictions, decisions, and frozen thresholds of TRACE, of every referencing control and reference intervention, and of all baselines and seeds. Its script, `reproduce.py`, recomputes from these files Table 3, Figure 3 with Tables 9 and 8, the reference interventions of Appendix J, the constant-onset comparison, and the deployed-metric reference switch, and checks each value against the one printed in this paper; the evidence-field ablation and the reference-length comparison are not covered. We reran the SWEC-ETHZ clip constructor from the raw source data and confirmed bit-exact signal checksums against the released clips; the HUP split is produced by a frozen-manifest constructor that inherits split and onset assignment row by row under per-row assertions. The analysis of Appendix N was run under a nested protocol whose contamination gates (patient-disjointness assertions, frozen fold maps, cohort-size checks) are enforced in code. The clip constructors, the training code of TRACE and the baselines, and the dated protocol of Appendix N will be released upon publication.

Ethics statement

This study uses only de-identified, publicly released human iEEG (OpenNeuro ds004100 and ds003029; SWEC-ETHZ; and a publicly released consensus-annotated SEEG dataset (Ojemann et al., 2026)). No new human data were collected. Seizure detection models are assistive research artifacts and must not inform clinical decisions without prospective validation under regulatory oversight.

AI use statement

In this work, we used generative AI tools for four purposes. (1) To aid and polish writing: editing the text for readability and formatting references. (2) For retrieval and discovery: searching for, identifying, and summarizing related literature, and checking bibliographic entries against their sources. (3) For research ideation and execution: proposing and refining hypotheses; providing feedback on the research methodology, including the nested cross-fitting protocol and its dated pre-registration amendments; implementing experimental code (the cohort-construction, feature-extraction, training, and analysis scripts); and interpreting results. (4) To draft sections of the paper: drafting parts of the manuscript text and creating or modifying scientific figures. We did not use generative AI tools to generate synthetic datasets or to formulate or prove mathematical claims. We have reviewed all AI-assisted work: experimental code was checked against independently recomputed statistics before any number was trusted, every citation was checked against its source, and all methodological decisions and their justifications were made by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

References

Muhammad Furqan Afzal, Sharanya Arcot Desai, Wade Barry, Jonathan Kuo, Shawna W. Benard, Christopher B. Traner, Erik J. Kobylarz, Thomas K. Tcheng, David Greene, Cairn Seale, and Martha J. Morrell. AI-driven electrographic seizure classification and seizure onset detection using image- and time-series-based approaches. *IEEE Transactions on Biomedical Engineering*, 73(2): 872–880, 2026. doi: 10.1109/TBME.2025.3594592.

Sharanya Arcot Desai, Muhammad Furqan Afzal, Wade Barry, Jonathan Kuo, Shawna Benard, Christopher Traner, Thomas Tcheng, Cairn Seale, and Martha Morrell. Expert and deep learning model identification of iEEG seizures and seizure onset times. *Frontiers in Neuroscience*, 17: 1156838, 2023. doi: 10.3389/fnins.2023.1156838.

Fabrice Bartolomei, Patrick Chauvel, and Fabrice Wendling. Epileptogenicity of brain structures in human temporal lobe epilepsy: a quantified study from intracerebral EEG. *Brain*, 131(7): 1818–1830, 2008. doi: 10.1093/brain/awn111.

John M. Bernabei, Adam Li, Andrew Y. Revell, Rachel J. Smith, Kristin M. Gunnarsdottir, Ian Z. Ong, Kathryn A. Davis, Nishant Sinha, Sridevi V. Sarma, and Brian Litt. HUP iEEG Epilepsy Dataset. OpenNeuro, ds004100, `https://doi.org/10.18112/openneuro.ds004100.v1.1.3`, 2023.

Alessio Burrello, Kaspar Schindler, Luca Benini, and Abbas Rahimi. Hyperdimensional computing with local binary patterns: One-shot learning of seizure onset and identification of ictogenic brain regions using short-time iEEG recordings. *IEEE Transactions on Biomedical Engineering*, 67(2): 601–613, 2020. doi: 10.1109/TBME.2019.2919137.

Alessio Burrello, Simone Benatti, Kaspar Schindler, Luca Benini, and Abbas Rahimi. An ensemble of hyperdimensional classifiers: Hardware-friendly short-latency seizure detection with automatic iEEG electrode selection. *IEEE Journal of Biomedical and Health Informatics*, 25(4):935–946, 2021. doi: 10.1109/JBHI.2020.3022211.

Alexander M. Chan, Felice T. Sun, Erem H. Boto, and Brett M. Wingeier. Automated seizure onset detection for accurate onset time determination in intracranial EEG. *Clinical Neurophysiology*, 119(12):2687–2696, 2008. doi: 10.1016/j.clinph.2008.08.025.

Felix Chan and Laurent L. Pauwels. Some theoretical results on forecast combinations. *International Journal of Forecasting*, 34(1):64–74, 2018. doi: 10.1016/j.ijforecast.2017.08.005.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016. doi: 10.1145/2939672.2939785.

Zheng Chen, Yasuko Matsubara, Yasushi Sakurai, and Jimeng Sun. Long-term EEG partitioning for seizure onset detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 14221–14229, 2025. doi: 10.1609/aaai.v39i13.33557.

Jeff Craley, Christophe Jouny, Emily Johnson, David Hsu, Raheel Ahmed, and Archana Venkataraman. Automated seizure activity tracking and onset zone localization from scalp EEG using deep neural networks. *PLoS ONE*, 17(2):e0264537, 2022. doi: 10.1371/journal.pone.0264537.

Jonathan Dan, Amirhossein Shahbazinia, Christodoulos Kechris, and David Atienza. Quantifying the generalization gap in seizure detection: A large-scale empirical benchmark via the SzCORE challenge. *arXiv preprint arXiv:2505.18191*, 2025.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Sarah J. Gascoigne, Nathan Evans, Gerard Hall, Csaba Kozma, Mariella Panagiotopoulou, Gabrielle M. Schroeder, Callum Simpson, Christopher Thornton, Frances Turner, Heather Woodhouse, Jess Blickwedel, Fahmida A. Chowdhury, Beate Diehl, John S. Duncan, Ryan Faulder, Rhys H. Thomas, Kevin Wilson, Peter N. Taylor, and Yujiang Wang. Incomplete resection of the intracranial electroencephalographic seizure onset zone is not associated with postsurgical outcomes. *Epilepsia*, 65(9):e163–e169, 2024. doi: 10.1111/epi.18061.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 35, 2022.

He He and Dongrui Wu. Transfer learning for brain–computer interfaces: A Euclidean space data alignment approach. *IEEE Transactions on Biomedical Engineering*, 67(2):399–410, 2020. doi: 10.1109/TBME.2019.2913914.

Himanshu Kumar, N. P. Guhan Seshadri, David Martinez, Imad Najm, Andreas Alexopoulos, Juan C. Bulacio, Demitre Serletis, and Balu Krishnan. Three-phase seizure segmentation in stereotactic EEG using envelope-based multivariate changepoint analysis. *Annals of Biomedical Engineering*, 2026. doi: 10.1007/s10439-026-04097-7.

Adam Li, Chester Huynh, Zachary Fitzgerald, Iahn Cajigas, Damian Brusko, Jonathan Jagid, Angel O. Claudio, Andrés M. Kanner, Jennifer L. Hopp, Stephanie Chen, Jennifer Haagensen, Emily L. Johnson, William S. Anderson, Nathan E. Crone, Sara K. Inati, Kareem A. Zaghloul, Juan Bulacio, Jorge González-Martínez, and Sridevi V. Sarma. Neural fragility as an EEG marker of the seizure onset zone. *Nature Neuroscience*, 24(10):1465–1474, 2021. doi: 10.1038/s41593-021-00901-w.

Taida Li, Yujun Yan, Fei Dou, Wenzhan Song, and Xiang Zhang. Cross-subject generalization for EEG decoding: a survey of deep learning methods. *Progress in Biomedical Engineering*, 8(2): 022013, 2026. doi: 10.1088/2516-1091/ae65f0.

Michael L. Martini, Aly A. Valliani, Claire Sun, Anthony B. Costa, Shan Zhao, Fedor Panov, Saadi Ghatan, Kanaka Rajan, and Eric Karl Oermann. Deep anomaly detection of seizures with paired stereoelectroencephalography and video recordings. *Scientific Reports*, 11(1):7482, 2021. doi: 10.1038/s41598-021-86891-y.

William K. S. Ojemann, Daniel J. Zhou, Catherine V. Kulick-Soper, Jacob Korzun, Joshua J. LaRocque, Caren Armstrong, Benjamin C. Kennedy, Sudha K. Kessler, Eric D. Marsh, Nishant Sinha, Brian Litt, and Erin C. Conrad. Multi-expert consensus annotations of spontaneous and stimulation-induced seizures in stereotactic EEG. *medRxiv preprint*, 2026. doi: 10.64898/2026.01.15.26344025.

Ivan Osorio, Mark G. Frei, and Steven B. Wilkinson. Real-time automated detection and quantitative analysis of seizures and short-term prediction of clinical onset. *Epilepsia*, 39(6):615–627, 1998. doi: 10.1111/j.1528-1157.1998.tb01430.x.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954. doi: 10.1093/biomet/41.1-2.100.

Wei Yan Peh, Prasanth Thangavel, Yuanyuan Yao, John Thomas, Yee-Leng Tan, and Justin Dauwels. Six-center assessment of CNN-Transformer with belief matching loss for patient-independent seizure detection in EEG. *International Journal of Neural Systems*, 33(3):2350012, 2023. doi: 10.1142/S0129065723500120.

John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pp. 61–74. MIT Press, 1999.

S. Raghu, Natarajan Sriraam, Erik D. Gommer, Danny M. W. Hilkman, Yasin Temel, Shyam Vasudeva Rao, Alangar Satyaranjandas Hegde, and Pieter L. Kubben. Cross-database evaluation of EEG based epileptic seizures detection driven by adaptive median feature baseline correction. *Clinical Neurophysiology*, 131(7):1567–1578, 2020. doi: 10.1016/j.clinph.2020.03.033.

Xiaopeng Si, Zhuobin Yang, Xingjian Zhang, Yulin Sun, Weipeng Jin, Le Wang, Shaoya Yin, and Dong Ming. Patient-independent seizure detection based on long-term iEEG and a novel lightweight CNN. *Journal of Neural Engineering*, 20(1):016037, 2023. doi: 10.1088/1741-2552/acb1d9.

Yulin Sun, Longlong Cheng, Xiaopeng Si, Runnan He, Tania Pereira, Meijun Pang, Kuo Zhang, Xin Song, Dong Ming, and Xiuyun Liu. Multi-task transformer network for subject-independent iEEG seizure detection. *Expert Systems with Applications*, 268:126282, 2025. doi: 10.1016/j.eswa.2024.126282.

Siyi Tang, Jared Dunnmon, Khaled Kamal Saab, Xuan Zhang, Qianying Huang, Florian Dubost, Daniel Rubin, and Christopher Lee-Messer. Self-supervised graph neural networks for improved electroencephalographic seizure analysis. In *International Conference on Learning Representations (ICLR)*, 2022. URL `https://openreview.net/forum?id=k9bx1EfHI_-`.

Xiaoshuang Wang, Xiulin Wang, Wenya Liu, Zheng Chang, Tommi Kärkkäinen, and Fengyu Cong. One dimensional convolutional neural networks for seizure onset detection using long-term scalp and intracranial EEG. *Neurocomputing*, 459:212–222, 2021. doi: 10.1016/j.neucom.2021.06.048.

Cynthia Yang, Egill A. Fridgeirsson, Jan A. Kors, Jenna M. Reps, Peter R. Rijnbeek, Jenna Wong, and Ross D. Williams. Does using a stacking ensemble method to combine multiple base learners within a database improve model transportability? In *Studies in Health Technology and Informatics*, volume 302, pp. 129–130, 2023. doi: 10.3233/SHTI230080.

Liuyin Yang, Qiang Sun, Ang Li, and Marc M. Van Hulle. Are EEG foundation models worth it? Comparative evaluation with traditional decoders in diverse BCI tasks. In *International Conference on Learning Representations (ICLR)*, 2026.

Zhikai Yu. Zero-annotation seizure detection and onset zone localization in intracranial EEG via self-supervised learning and Mahalanobis distance modeling. *Research Square preprint*, 2026. doi: 10.21203/rs.3.rs-9568998/v1.

Zhizhang Yuan, Daoze Zhang, Yang Yang, Junru Chen, and Yafeng Li. PPi: Pretraining brain signal model for patient-independent seizure detection. In *Advances in Neural Information Processing Systems*, volume 36, pp. 69586–69604, 2023. doi: 10.52202/075280-3047.

Marzieh Zare. A negative-control protocol for clinical EEG foundation-model benchmarks: dataset identity and external-cohort stress testing. *arXiv preprint arXiv:2607.24519*, 2026.

A DATA PREPROCESSING

This appendix gives the full preprocessing chain from the public source recordings to the 60 s clips evaluated in the main text. Every step is patient- and cohort-agnostic: no step uses a label, a patient identity, or a statistic computed outside the clip being constructed, so the same frozen constructor serves the training, validation, test, and zero-shot cohorts alike.

Signal conditioning. The three cohorts are acquired at different native sampling rates and with different hardware (Table 1). Each recording is notch-filtered at 50, 60, and 150 Hz (the two mains frequencies of the recording sites and a low harmonic) and then resampled to a common 512 Hz with a polyphase filter, in that order; a 60 s channel therefore holds 30,720 samples. No montage change, re-referencing, channel selection, or artifact-channel rejection is applied: channels are kept as recorded, and the channel count C varies freely across patients and cohorts (the zero-shot example of Figure 2 carries 146 channels). The episode-local referencing at the heart of TRACE is deliberately *not* part of this chain — it happens inside feature construction (Eq. equation 1), clip by clip — so no quantity computed outside the incoming clip ever enters the signal path.

Clip construction. Seizure-positive clips are 60 s windows placed so that the annotated onset falls in $[10, 40]$ s, with the position drawn at random rather than fixed at the clip midpoint; no positional prior is therefore available, and localization must be evidence-driven. The $[10, 40]$ s placement guarantees at least 10 s of pre-onset and 20 s of post-onset activity in every positive clip; in particular, the 8 s episode-local reference $\mathcal{B} = \{1, \ldots, 16\}$ is pre-ictal by construction and never overlaps the onset it must help localize. Windows reaching past a recording boundary are cut or padded to the full 60 s. Non-seizure clips are 60 s windows containing no annotated seizure; their placement on Pennsieve-530 follows a pre-registered negative-window rule, frozen before any clip of that cohort was built (Appendix Q). After conditioning, each clip is divided into $T = 120$ non-overlapping 0.5 s frames per channel on which the 23 background-contrast descriptors of Appendix C are computed.

Cohort-specific construction and provenance. SWEC-ETHZ contributes strip, grid, and depth recordings, and HUP and Pennsieve-530 SEEG recordings (Table 1). The SWEC-ETHZ clip constructor was rerun from the raw source data and confirmed against the released clips with bit-exact signal checksums. The HUP split is produced by a frozen-manifest constructor that inherits the split and onset assignment row by row under per-row assertions. Pennsieve-530 onsets are multi-expert consensus marks; its clips were assembled as 210 windows — 78 positive (44 spontaneous and 34 stimulation-induced) and 132 negative — before any TRACE output existed, and the full evaluation ledger for the cohort is recorded in Appendix M. The 34 stimulation-induced seizures are outside the scope of this work and are excluded from all seizure analyses of the main text.

Splits and integrity gates. The development cohorts are partitioned by patient into train, validation, and test sets (SWEC-ETHZ 10/3/3 patients; HUP 25/6/5), and no patient appears in more than one split (Table 1). Before any evaluation runs, the constructors assert unique clip identifiers across cohorts, patient-disjoint splits, and the expected cohort sizes (Appendix P). Pennsieve-530 patients are disjoint from both development cohorts, although institutions partially overlap, and the cohort is used zero-shot only: no model configuration, threshold, or decoder rule of TRACE was selected on it. For accounting, the pooled SWEC-ETHZ and HUP training set comprises 405 clips (149 seizures); per-patient clip counts are carried by the released split lists and clip manifests.

B IMPLEMENTATION DETAILS

Recordings were notch-filtered at 50, 60, and 150 Hz, resampled to 512 Hz with a polyphase filter, and cut or padded to 60 s, which yields 120 frames of 0.5 s per channel. Frame-level evidence was learned from the 23 descriptors of Appendix C, with the frame labels of Section 3.2. Each frame classifier is a gradient-boosted tree ensemble (XGBoost, 300 trees, maximum depth 6, learning rate 0.1, row and column subsampling 0.8, histogram splits) with the positive class weighted by the negative-to-positive frame ratio. The decoder constants of Section 3.5 were fixed before any evaluation reported here.

Detection head and operating threshold. The segment-level head is an XGBoost classifier with the same booster settings as the frame encoder, except that the maximum depth ($\{2, 3\}$) and the number of trees ($\{100, 200\}$) are selected inside each training fold by patient-grouped inner three-fold cross-validation. The four detection features are standardized with the mean and standard deviation of the training folds. Frame-evidence predictions are always patient-level out-of-fold over five folds, so the head is never fitted on in-fold encoder outputs. The operating threshold τ_{det} maximizes, on the same out-of-fold probabilities, a composite of detection metrics and a deployed-onset score,

$$\begin{aligned} \mathrm{TS} &= \tfrac{1}{15}\big(5\,\mathrm{AUROC} + 3\,\mathrm{F1} + 2\,\mathrm{Sens} + 2\,\mathrm{Spec} + 3\,\mathrm{Ons}\big), \\ \mathrm{Ons} &= \max\!\left(0,\, 1 - \frac{\text{deployed MAE}}{60}\right). \end{aligned} \tag{12}$$

with missed seizures charged the full 60 s in the deployed MAE. The AUROC term is constant across candidate thresholds and does not move the argmax, so the effective objective trades F1, sensitivity, and specificity against an onset score that penalizes missed seizures.

Hardware and runtime. All models train on CPU (Python 3.13, XGBoost 3.3; the deep baselines use CPU-only PyTorch 2.13). TRACE trains end to end in approximately 380 s on the combined SWEC-ETHZ and HUP training set. The CUSUM recurrence behind the z_{cu} feature is given in Appendix D.

C FRAME-LEVEL DESCRIPTOR DEFINITIONS

Each 0.5 s frame of each channel is described by the 23 quantities in Table 4, all computed relative to the episode-local background $\mathcal{B}$ (the first 8 s, 16 frames). Amplitude descriptors use the log-domain contrast to the background median, and spectral-shape descriptors use the plain difference to the background median, both as defined in the Method section. Band shares are fractions of the 0.5–256 Hz power in the delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), low-gamma (30–80 Hz), and high-frequency (80–200 Hz) bands; H is the normalized spectral entropy; LL_z is the line length of the background-normalized signal $\tilde{x}_c$. Trailing statistics use the frames preceding t. Indices are columns of the released feature code; columns 11–13 replicate columns 7–9 in the released implementation and are kept because the frozen models were trained on this layout.

D DETECTION FEATURES

The detection branch determines whether an onset estimate should be reported. It combines three summaries of $\mathbf{E}$ with a complementary feature-level cumulative-change signal. Let $\mathcal{P} = \{1, \ldots, 48\}$ and $\mathcal{Q} = \{73, \ldots, 120\}$ denote fixed early and late windows, corresponding to 0–24 s and 36–60 s. Define $\mu(t) = C^{-1} \sum_c E(t, c)$ and the channel-wise contrast $d_c = |\mathcal{Q}|^{-1} \sum_{t \in \mathcal{Q}} E(t, c) - |\mathcal{P}|^{-1} \sum_{t \in \mathcal{P}} E(t, c)$. The four detection features are

$$\begin{aligned} z_{\mathrm{diff}} &= Q_{75}(\{\mu(t) : t \in \mathcal{Q}\}) - Q_{75}(\{\mu(t) : t \in \mathcal{P}\}), \\ z_{\mathrm{top}} &= \frac{1}{|\mathcal{S}_{\mathrm{det}}|} \sum_{c \in \mathcal{S}_{\mathrm{det}}} d_c, \\ z_{\mathrm{cu}} &= \max_t \sum_c G_c(t), \\ z_{\mathrm{var}} &= \mathrm{std}(\{E(t, c) : t \in \mathcal{Q},\ c = 1, \ldots, C\}). \end{aligned} \tag{13}$$

Table 4: The 23 frame-level descriptors. Indices are columns of the released feature code.

Index	Descriptor	Definition
7, 12	Line-length contrast	$\log(1 + \mathrm{LL})$ minus its median over $\mathcal{B}$; LL is the mean absolute first difference within the frame
8, 11	RMS contrast	the same contrast for the root-mean-square amplitude
9, 13	Peak-to-peak contrast	the same contrast for the peak-to-peak amplitude
16, 20	RMS contrast, trailing means	mean over the preceding 2 s and 4 s
17	LL_z, trailing mean	mean over the preceding 2 s
25	LL_z, first difference	$\mathrm{LL}_z(t) - \mathrm{LL}_z(t-1)$
28, 29, 31	Supra-median persistence of the RMS contrast, LL_z, and H	fraction of the preceding 8 s in which the quantity exceeds its median over the clip
33	High-frequency power contrast	$\log(1 + P_{13\text{–}200\,\mathrm{Hz}})$ minus its median over $\mathcal{B}$
34	High-frequency power contrast, trailing mean	mean of descriptor 33 over the preceding 4 s
35–40	Band-share changes	share of each of the six bands minus its median over $\mathcal{B}$
41	Spectral-entropy change	H minus its median over $\mathcal{B}$
42	Fast-activity share change	summed beta, low-gamma, and high-frequency share minus its median over $\mathcal{B}$

The set $\mathcal{S}_{\mathrm{det}}$ contains the channels with the largest $K_{\mathrm{det}} = \max(3, \lfloor C/10 \rfloor)$ contrasts, selected once per clip. Here $G_c(t)$ is a one-sided Page CUSUM (Page, 1954) computed from the line-length and high-frequency feature trajectories before encoding, defined as follows. Let $u_c(t)$ be the sum of the z-domain line-length and high-frequency descriptor values of channel c at frame t (feature columns 14 and 15 of the released implementation). Each channel is detrended within the clip, $\tilde{u}_c(t) = u_c(t) - \mathrm{median}_t\, u_c$, and the allowance is set adaptively per channel, $k_c = 0.5\ \mathrm{median}_t\, |\tilde{u}_c(t)| + 10^{-6}$. The recurrence is

$$G_c(0) = 0, \qquad G_c(t) = \max(0,\ G_c(t-1) + \tilde{u}_c(t) - k_c)\,, \tag{14}$$

and $z_{\mathrm{cu}} = \max_t \sum_c G_c(t)$ as in Eq. equation 13. Thus, the CUSUM feature supplements the evidence-field summaries rather than being derived from $\mathbf{E}$.

E Baseline onset readouts

The four probability-based detectors (1D-CNN, LightWeightCNN, CNN–Transformer, MT-Transformer) share one post-hoc readout: a three-window moving average of the clip-level seizure probability, then the first sustained threshold crossing. The onset threshold and the minimum duration are selected on source validation patients and frozen. Martini2021, SODor, ViT-SOD, and SVM-Reg use the published method's own readout as implemented. Several implementations encode failure to return a valid onset as timestamp 0.0 s, which cannot occur for a true onset in $[10, 40]$ s. TRACE uses a 25 s fallback when no frame meets the sustained-evidence criterion; that fallback is never triggered on the 44 spontaneous seizures.

Detection thresholds for all eight baselines follow one rule — maximum F1 over $\{0.5\}$ and 19 quantiles on the source validation patients, frozen for every target — and inputs are z-scored per clip with the median and interquartile range of the full 60 s; seeds, bounded epoch budgets, and the remaining shared protocol are recorded in Appendix M. Table 5 states the conventions for each baseline.

F Matched-fallback sensitivity

Replacing the 0.0 s sentinel with TRACE's 25 s fallback changes six of the eight baselines. Counts of sentinel outputs over three seeds and 44 spontaneous clips: 1D-CNN 20, LightWeightCNN 18, CNN–Transformer 29, MT-Transformer 17, Martini2021 76, SODor 99. ViT-SOD and SVM-Reg emit none. Seed-averaged MAE under this equal treatment: 1D-CNN 12.57 s, LightWeightCNN 12.05 s, CNN–Transformer 10.53 s, MT-Transformer 13.82 s, Martini2021 13.19 s, SODor 11.93 s, ViT-SOD 8.89 s, SVM-Reg 10.28 s. TRACE remains 4.80 s. The lowest baseline is therefore still ViT-SOD at 8.89 s. Patient-level paired bootstrap, Bonferroni-corrected over eight comparisons,

Table 5: Per-baseline onset readout and no-valid-onset behavior. Sentinel counts are pooled over three seeds and the 44 spontaneous Pennsieve-530 clips (Appendix F); a 0.0 s timestamp cannot arise for a true onset, which lies in $[10, 40]$ s. TRACE's own 25 s fallback is never triggered on those clips.

Method	Onset readout	No valid onset
1D-CNN	shared post-hoc[s]	0.0 s sentinel (20)
LightWeightCNN	shared post-hoc[s]	0.0 s sentinel (18)
CNN–Transformer	shared post-hoc[s]	0.0 s sentinel (29)
MT-Transformer	shared post-hoc[s]	0.0 s sentinel (17)
Martini2021	own published readout	0.0 s sentinel (76)
SODor	own published readout	0.0 s sentinel (99)
ViT-SOD	own published readout	none
SVM-Reg	own published readout	none

[s]Three-window moving average of the clip-level seizure probability, then the first sustained threshold crossing; threshold and minimum duration selected on source validation patients and frozen.

leaves seven intervals excluding zero. The comparison that does not survive is TRACE versus ViT-SOD (uncorrected $+3.65$ $[0.80, 6.21]$; Bonferroni lower bound -0.48).

G CHANNEL-AGGREGATION FACTORIAL

Table 6 is the full $K\times$pooling audit on the only data that can legitimately inform the choice: patient-out-of-fold predictions from the combined SWEC-ETHZ and HUP training set (405 clips, 149 seizures). The $K=1$ mean and median arms are identical, leaving nine unique configurations. No sparse alternative significantly improves on the frozen $K=5$ mean arm. Dense all-channel pooling is worse.

Table 6: Full channel-aggregation factorial on the legitimate selection set (SWEC-ETHZ+HUP train OOF; 405 clips, 149 seizures). The frozen TRACE configuration is bolded. p is the nominal patient-level paired Wilcoxon test versus the frozen arm.

K	Pooling	MAE (s) ↓	Hit@5s ↑	p
1	mean/median	7.14	57.0%	0.331
3	mean	7.11	59.7%	0.699
3	median	7.05	57.7%	0.973
5	**mean (frozen)**	**7.08**	**61.7%**	—
5	median	<u>6.67</u>	**61.7%**	0.325
10	mean	7.44	61.1%	0.058
10	median	7.41	**61.7%**	0.174
all	mean	9.23	46.3%	0.008
all	median	9.98	44.3%	0.001

As a post hoc robustness check, the same frozen grid was evaluated without re-selection on the held-out development splits and on spontaneous Pennsieve-530 seizures. On the external cohort, $K=10$ with mean pooling is numerically lowest (4.13 s) compared with 4.80 s for the frozen arm; this external readout cannot legitimately be used to change the model. At a pre-specified practical tolerance of $\delta=1$ s, an equivalence test finds the frozen arm equivalent to the nominal best arm on the external spontaneous-seizure set (95% CI $[-0.02, +0.96]$).

H WHAT ENABLES ZERO-SHOT TRANSFER?

This appendix examines which parts of TRACE are responsible for the cross-cohort behavior of §4.2.2. All variants are constructed using the development cohorts only and are evaluated on Pennsieve-530 without target-domain training, tuning, or parameter re-selection. Table 7 progressively removes or replaces the representation and aggregation stages of TRACE. The primary endpoint is oracle onset MAE on the 44 spontaneous seizures. The two aggregation arms complement the selection-set factorial of Appendix G, which audits all $K\times$pooling combinations on out-of-fold development data.

Table 7: Zero-shot ablation of the TRACE evidence-field formulation on Pennsieve-530. Values are oracle onset MAE in seconds (↓) over 44 spontaneous seizures. Δ denotes the increase in MAE relative to the full frozen TRACE framework. All variants are defined using source-cohort data only and applied unchanged to the external cohort.

Configuration	**MAE (s) ↓**	**Δ MAE**
Direct onset prediction without an evidence field		
Clip-level XGBoost regression (69-d)	6.87	+2.07
CNN regression, episode-local features	8.54	+3.74
CNN regression, absolute features	12.11	+7.31
Replacing the learned evidence field		
$E(t, c) \leftarrow u_c(t)$ (unlearned signal track)	13.08	+8.28
Removing evidence-field structure		
Temporal mean of $E(t, c)$	8.89[†]	+4.09
Channel mean of $E(t, c)$	8.54	+3.74
Ablating sparse and dynamic channel aggregation		
All-channel mean (no sparse selection)	7.93	+3.13
Static top-5 (no dynamic re-selection)	5.00	+0.20
Non-informative onset baseline		
Constant onset (25 s)	8.89	+4.09
TRACE (ours; full frozen framework)	**4.80**	—

[†]Temporal averaging removes all temporal variation from the evidence trajectory. The frozen decoder therefore finds no sustained anchor (fallback rate $= 1.000$) and returns the fixed 25 s fallback for every clip. This row is a structural sanity check rather than an independently learned alternative.

Variant definitions. *Clip-level XGBoost regression* aggregates each of the 23 frame-level descriptors to three clip-level statistics—the early–late window contrast (windows $\mathcal{P}$, $\mathcal{Q}$ of Appendix D), the maximum, and the standard deviation—and regresses onset time directly from these 69 features. Each *CNN regressor* takes the clip's descriptor matrix channel-averaged to 23×120 (23 descriptors $\times$ 120 frames; episode-local or absolute variant), passes it through $\mathrm{Conv1d}(23{\to}64, k{=}5)$–BN–ReLU and $\mathrm{Conv1d}(64{\to}64, k{=}5)$–ReLU, and reads the onset from a global-average-pooling linear head; training is MSE on onset$/60$ for a fixed 300 epochs without early stopping, learning rate 10^{-3}, weight decay 10^{-4}, three seeds $\{42, 43, 44\}$ (reported values average the three seeds), on the 149 source training positives supplemented with validation positives (fewer than 200 in total, per the task protocol), on CPU, and the trained regressor is forwarded to each target cohort exactly once. The *unlearned signal track* replaces $\mathbf{E}$ with the $u_c(t)$ trajectory of Appendix D, min–max normalized per clip, with every downstream stage frozen. *Temporal mean* replaces each channel's evidence trajectory by its clip-level mean; *channel mean* decodes the channel-averaged trajectory $\mu(t)$. *All-channel mean* pools all channels without sparse selection ($K{=}\mathrm{all}$). *Static top-5* ranks channels by the temporal mean of the frozen evidence $E(t, c)$ over the clip's 120 frames, fixes the top five channels for the entire clip, and feeds the per-frame mean of those five channels' evidence to the unchanged decoder.

The evidence field cannot be replaced by direct onset prediction. We first bypass the channel–time evidence field and predict onset directly from source-trained features. A clip-level XGBoost regressor reaches 6.87 s MAE, 2.07 s worse than TRACE. Direct CNN regression degrades further: episode-local features yield 8.54 s, whereas using absolute features increases the error to 12.11 s. The difference between the two CNN variants shows that episode-local referencing remains beneficial even without the evidence field, but the substantially lower error of TRACE indicates that reference-conditioned features alone do not explain the zero-shot result. Organizing them into a temporally resolved evidence representation provides an additional transfer advantage.

Transfer depends on a learned, structured channel–time field. Replacing the learned evidence $E(t, c)$ with the unlearned signal trajectory $u_c(t)$ increases MAE to 13.08 s, the largest degradation among the non-degenerate ablations. Thus, preserving a time-varying signal alone is not sufficient; the learned transformation into seizure evidence is important for transfer. The structure of this field also

matters. Collapsing the channel dimension increases MAE to 8.54 s, showing that channel-specific evidence contains information that is lost after global averaging. Collapsing the temporal dimension causes the frozen onset decoder to return its constant fallback for every seizure. This latter experiment should be interpreted as a structural sanity check: temporal variation is required by construction for localizing a transition boundary.

Sparse channel selection matters more than dynamic re-selection. The channel dimension is most useful when TRACE suppresses uninformative channels before pooling. Averaging over all channels increases zero-shot MAE from 4.80 to 7.93 s. In contrast, replacing dynamic channel re-selection with a static top-5 set produces only a small change, from 4.80 to 5.00 s. The ablation therefore distinguishes two roles that would otherwise be conflated: *sparse selection* provides a substantial contribution, whereas *dynamic re-selection* is a comparatively small refinement.

Summary. Taken together, these experiments identify a hierarchy within TRACE rather than suggesting that every component contributes equally. Episode-local features already transfer better than absolute features, but their advantage is substantially strengthened when they are converted into a learned channel–time evidence field. Preserving channel-specific structure and sparsely aggregating informative channels provides further gains, while dynamic re-selection contributes only modestly. The full frozen TRACE framework achieves the lowest zero-shot onset error, supporting the view that cross-cohort generalization arises from the reference-conditioned evidence formulation as a whole rather than from a direct regressor or a single downstream decoding heuristic.

I REFERENCING CONTROLS

Each control replaces the episode-local reference of the descriptors with another normalization or rereferencing scheme; encoder family, training budget, aggregation, and decoder are unchanged, and each is trained on SWEC-ETHZ and HUP and applied once to the 44 spontaneous Pennsieve-530 seizures. Table 9 gives the matched switch in all five settings, and Table 8 lists all ten controls on Pennsieve-530, with the patient-level paired gain of TRACE (22 patients; percentile bootstrap, 10,000 draws).

Table 8: Referencing controls on the 44 spontaneous Pennsieve-530 seizures. Oracle onset MAE (s); gain is control minus TRACE, averaged within patient, with 95% bootstrap CI. TRACE: 4.80 s.

Control	MAE (s)	Gain of TRACE
Absolute descriptors (matched switch)	7.54	3.62 [1.32, 6.04]
Clip-start centering	7.25	3.46 [1.22, 5.90]
Clip-start scaling	7.89	3.59 [1.18, 6.23]
Window z-score	7.44	3.48 [1.38, 5.77]
Common-average reference, absolute	8.10	4.01 [1.17, 7.57]
Common-average reference, relative blocks	6.08	1.70 [0.12, 3.60]
Dataset robust scaling	7.69	3.64 [1.37, 6.10]
Patient robust scaling	6.92	2.53 [0.46, 4.77]
Trailing 8 s background	9.26	5.57 [2.51, 9.06]
Temporal difference	7.60	3.15 [0.79, 5.69]

Table 9: Matched reference switch. Absolute: the same descriptors without the episode-local reference, retrained. Oracle MAE (s); patient-level gain with 95% bootstrap CI (descriptive for the 3- and 5-patient test sets).

Setting	Absolute	TRACE	Patient-level improvement
SWEC-ETHZ test	8.99	**4.52**	4.68 [2.66, 6.35]
HUP test	7.32	**4.92**	2.82 [1.15, 4.32]
SWEC-ETHZ → HUP	15.40	**8.62**	7.23 [5.53, 8.97]
HUP → SWEC-ETHZ	10.34	**6.21**	3.35 [2.03, 4.64]
Pennsieve-530	7.54	**4.80**	3.62 [1.32, 6.04]

Table 10: Robustness of the frozen episode-local reference to test-time interventions. Every intervention modifies the reference only at test time; model parameters, thresholds, and decoder rules remain unchanged. Endpoints are onset MAE.

Axis	Intervention	Main observation
Length	Reference of 4, 8, or 16 s	4 s degrades development-test MAE, while 8 and 16 s are comparable; the frozen 8 s choice is not narrowly tuned.
Identity	Own episode → same-patient seizure → same-patient interictal → other patient → other site	Spontaneous-seizure MAE increases monotonically: $4.80 \rightarrow 6.70 \rightarrow 8.04 \rightarrow 9.76 \rightarrow 12.28$ s.
Position	Shift the 8 s window away from or toward the onset	Nearby clean pre-onset background performs best; windows that approach or cross the onset degrade sharply (Table 11).
Purity	Mix 0–50% ictal frames into the reference	Stable through moderate contamination, deteriorating strongly at 50% (Table 11).

J REFERENCE ROBUSTNESS

This appendix audits the frozen episode-local reference of §3.2 along four axes: length, identity, position relative to the onset, and purity. All interventions modify the reference at test time only; trained model parameters, thresholds, and decoder rules are unchanged. The length axis is evaluated on the frozen development test sets; the identity, position, and purity axes are evaluated on the 44 spontaneous Pennsieve-530 seizures, the identity ladder restating that of §4.2.3 and the position and purity controls being reported in Table 11.

Table 11: Reference-window position and purity controls on the 44 spontaneous Pennsieve-530 seizures (frozen model and decoder, no retraining; onset MAE in seconds). *Position*: the 8 s reference window slides within the clip. *Oracle nearest-clean*: the window ends g s before the annotated onset; clips without a feasible window are skipped, so the sample shrinks as the gap grows ($n = 44, 38, 28, 21, 18$ clips at gaps of $2, 5, 10, 15, 20$ s). *Purity*: the reference is mixed with $[\text{onset}+10,\ \text{onset}+18)$ s activity at rate α.

Setting	MAE (s) ↓
Position: window $[d,\ d+8)$ *s*	
$[0, 8)$ (frozen)	4.80
$[4, 12)$	4.65
$[8, 16)$	4.67
$[12, 20)$	5.44
$[16, 24)$	8.02
$[20, 28)$	8.68
Oracle nearest-clean: window ends at onset$-g$	
gap 2 s	3.93
gap 5 s	3.66
gap 10 s	3.82
gap 15 s	5.10
gap 20 s	6.05
gap 25 s	5.23
Purity: ictal mixing rate α	
$\alpha = 0$ (frozen)	4.80
$\alpha = 10\%$	4.77
$\alpha = 20\%$	4.05
$\alpha = 30\%$	4.52
$\alpha = 50\%$	9.00

Position. Sliding the 8 s window through the clip leaves the frozen choice essentially interchangeable with its neighbors: starting the reference at 4 or 8 s changes MAE by at most 0.15 s (4.65 and

4.67 s against the frozen 4.80 s). Degradation begins once the window can reach the onset — 5.44 s for $[12, 20)$, then 8.02 and 8.68 s for $[16, 24)$ and $[20, 28)$ — and grows monotonically with the window's start time, tracking the fraction of clips whose reference overlaps ictal activity (onsets lie in $[10, 40]$ s, so the frozen $[0, 8)$ window is pre-ictal by construction; Appendix A). The oracle rows sharpen the same reading from the other side: given the annotated onset, a clean window ending 2–10 s before it localizes better than the frozen clip-head reference (3.66–3.93 s), while windows ending ≥ 15 s before the onset are worse (5.10–6.05 s). Two caveats bound those rows: they are oracle — they use the label the deployed system must estimate — and their feasible sample shrinks with the gap, so their absolute values are not directly comparable across rows or to the frozen row. What transfers to the deployed setting is the direction: the useful reference is the most recent clean pre-onset background, and the frozen $[0, 8)$ s window forgoes about 1.1 s of MAE against the best oracle-clean placement while using no onset information.

Purity. Mixing ictal activity from $[\text{onset}+10,\ \text{onset}+18)$ s into the reference at rates of 10–30% leaves onset MAE at 4.05–4.77 s, at or slightly below the frozen 4.80 s; contamination of 50% raises it to 9.00 s. The break at the majority boundary is consistent with how the reference enters the descriptors: the robust standardization of Eq. equation 1 and the contrasts of Eq. equation 2 are taken against medians over the 16 reference frames, and a median is stable only while the contaminated frames remain a minority. We read the 10–30% rows as tolerance, not improvement; they are single-pass point values, and we attach no interval or claim to them.

K Decoder ablations

The decoder separates detection of established seizure evidence from recovery of the transition boundary. On the 44 spontaneous Pennsieve-530 seizures, replacing anchored backtracking with first crossing increases MAE from 4.80 to 5.71 s. The two rules agree on 42 of 44 seizures; in the two differing cases, an early transient crosses the threshold near 12 s, while backtracking from the sustained-evidence anchor returns an onset close to the clinical annotation. Anchored backtracking therefore protects against occasional early transient crossings. Source-side ablations that isolate per-clip self-calibration are not regenerated from the released per-clip files and are omitted here.

L Transfer between the two training cohorts

Transfer between the two training cohorts, which come from different centers, gave a mixed picture. From HUP to SWEC-ETHZ, TRACE localized onsets with an MAE of 6.21 s against 8.45 s for the vision transformer, a patient-level gain of 2.79 s (CI 1.25 to 4.36) that held for each of the three baseline seeds. From SWEC-ETHZ to HUP the two systems could not be separated. TRACE reached 8.62 s, whereas the vision transformer reached 8.16 s and 7.60 s with two seeds and 17.25 s with the third, so the average gap of 2.09 s (CI 0.13 to 3.99) rests on one unstable baseline run. Detection ranking under this transfer favored TRACE in one direction (AUROC 0.899 against 0.866 for the multi-task Transformer) and the CNN–Transformer in the other (0.714 against 0.665). We therefore claim an onset advantage on Pennsieve-530 and in the HUP-to-SWEC-ETHZ direction only.

M TRACE and the earlier configuration: differences, prior evaluation of Pennsieve-530, and baseline protocol

Two configurations of the same evidence design appear in this paper: TRACE, the configuration of the main text, and an earlier configuration used for the nested-protocol analysis of Appendix N. Table 13 lists the differences. No number mixes the two, and every TRACE number was recomputed from the released per-clip outputs with independently written code.

Prior evaluation of Pennsieve-530. The cohort was evaluated twice, first with the earlier configuration and then with TRACE, and we record every step:

Three facts follow. The reference window, the reference-dependent descriptors, and the decoder constants predate both evaluations at the byte level. The absolute-descriptor control, the paired-stability definition, and the pre-registered strata and threshold rules of TRACE were created after

Table 12: Zero-shot transfer between the two development cohorts, all eight baselines in both directions. Baselines report mean±SD over three seeds; TRACE reports the designated frozen configuration. Onset MAE is in seconds on the target cohort; AUROC is computed on all target clips. The three SWEC-ETHZ→HUP seeds of ViT-SOD (8.16, 7.60, 17.25 s) straddle TRACE's 8.62 s, which is why the text claims an onset advantage in the HUP→SWEC-ETHZ direction only.

	SWEC-ETHZ → HUP		HUP → SWEC-ETHZ	
Method	MAE (s) ↓	AUROC ↑	MAE (s) ↓	AUROC ↑
ViT-SOD	11.00 ± 5.42	0.577 ± 0.041	8.45 ± 0.32	0.585 ± 0.020
SVM-Reg	10.30 ± 0.00	0.528 ± 0.000	10.84 ± 0.00	0.654 ± 0.000
CNN–Transformer	24.23 ± 0.28	$\mathbf{0.714 \pm 0.020}$	14.84 ± 0.32	0.822 ± 0.007
LightWeightCNN	19.92 ± 0.62	0.624 ± 0.026	18.09 ± 0.26	0.789 ± 0.053
1D-CNN	19.98 ± 1.48	0.587 ± 0.039	18.15 ± 1.37	0.596 ± 0.039
MT-Transformer	20.73 ± 2.54	0.632 ± 0.031	18.57 ± 4.54	0.866 ± 0.034
Martini2021	23.16 ± 0.28	0.477 ± 0.024	21.36 ± 0.33	0.482 ± 0.021
SODor	24.95 ± 0.00	0.571 ± 0.036	22.78 ± 1.62	0.669 ± 0.060
TRACE	**8.62**	0.665	**6.21**	**0.899**

Table 13: The two configurations. Every contrast in this paper is internal to one column.

	Earlier configuration	TRACE
Features per (frame, channel)	33; frozen 13-dim stable subset	43; frozen 23-feature stable subset
Reference window	first 8 s of the clip	first 8 s of the clip (identical)
Frame classifier	XGBoost, 3-seed ensemble	XGBoost, 3-seed median ensemble
Training subjects	B: 25, nested 5-fold	A+B training patients (405 clips)
Detection head	shallow XGBoost (depth 2–3)	XGBoost (its TRACE head)
Head thresholds	max-F1 on nested out-of-fold	training-fold out-of-fold (disclosed)
Cohort E construction	156 clips (78 positive)	210 clips (78 positive, wider negatives)
Onset decoder	anchored backtracking (P2)	identical, byte-identical constants

the first outputs of the earlier configuration existed. And those earlier results were known before the TRACE pre-registration of 2026-09-11 was frozen, so the TRACE analyses of Pennsieve-530 are a locked re-analysis of a previously evaluated cohort, never a first look.

Baseline protocol. The eight baselines share one threshold rule: maximum F1 over $\{0.5\}$ and 19 quantiles on the source validation patients, frozen for the target. Checkpoint selection and the onset-readout search use the same patients, whereas TRACE takes its thresholds from training-fold out-of-fold predictions, a difference we state and do not smooth over. Baseline inputs are z-scored per clip with the median and interquartile range of the full 60 s. Hyperparameters that the original papers do not report were set to documented assumptions. All baselines carry three seeds (42, 3407, and 20260805); seeds 3407 and 20260805 were run in a sweep completed on 2026-09-16, after the pre-registration and after the single-seed results were available. Two epoch budgets are bounded by compute (30 for the spectrogram vision transformer, 20 for the CNN–Transformer). Conclusions drawn from these baselines are scoped to the evaluated implementations.

N Aggregation-layer analysis under a nested protocol (earlier configuration)

This appendix reports the analysis behind the decision-branch ablation. It was run with an earlier configuration of the same evidence pipeline (33 frame descriptors, a frame classifier trained on HUP only, and four hand-specified per-clip signals) under a nested cross-fitting protocol on four cohorts: HUP (B), SWEC-ETHZ (A), a multi-center cohort with clinical onset annotations (D) (Li et al., 2021), and the consensus cohort (E), which is Pennsieve-530. In this appendix, "the earlier configuration" and "our pipeline" both name that configuration, and TRACE is the configuration of the main text. Throughout, the cohort letters follow this mapping — A: SWEC-ETHZ, B: HUP, D: the multi-center cohort, E: Pennsieve-530 — and a letter is used only after the full cohort name at first mention in each subsection.

Table 14: Timeline of evaluations on Pennsieve-530. The last column names the configuration: the earlier one or TRACE.

Date	Event	Where
2026-09-02	design fixed; reference window frozen	both
2026-09-03 00:21	cohort assembled (210 clips)	TRACE
2026-09-03 13:57	cohort construction pre-registered (156 clips)	earlier
2026-09-03 14:19	first model outputs on the cohort	earlier
2026-09-03 14:42	single-pass results written up	earlier
2026-09-10	paired-stability definition	TRACE
2026-09-11 10:29	absolute-descriptor control parameterized	TRACE
2026-09-11 15:41	TRACE pre-registration frozen	TRACE
2026-09-11 16:05	first TRACE outputs on the cohort	TRACE

Table 15: In-domain results on HUP (B) (nested; each subject scored by a system fitted without it). The factorial separates the head from the threshold policy; ECE is threshold-invariant. E0F+Platt fits a one-parameter logistic calibration (Platt, 1999) on the fixed fusion inside each outer fold; because each fold's Platt map is a different monotone transform, its pooled AUC/PR-AUC need not equal (and does not equal) E0F's raw pooled values, though both are valid. Its F1/Sens/Spec are omitted: at the nominal 0.5 threshold every calibrated score falls on one side, a degenerate operating point, not an informative one (full detail in Table 17).

Head	Policy	AUC	PR-AUC	F1	Sens	Spec	ECE
E0F (fixed)	legacy	0.751	0.718	0.586	0.629	0.765	0.524
E0F (fixed)	max-F1 scan	0.751	0.718	0.624	0.551	0.903	0.524
E1a (learned)	legacy	0.731	0.679	0.601	0.517	0.908	0.144
E1a (learned)	max-F1 scan	0.731	0.679	0.592	0.472	0.944	0.144
E0F + Platt	—	0.727	0.600	—	—	—	0.165

N.1 In-domain: an operating-point trade, not a ranking gain

Table 15 reports the nested in-domain comparison. We find no evidence that the head changes threshold-free ranking: $\Delta\mathrm{AUC} = +0.020$ with CI $[-0.056, +0.085]$ on the training folds and $+0.009$ $[-0.090, +0.115]$ on the 5-subject test split; PR-AUC agrees (0.718 vs. 0.679). What the head changes is the operating point: specificity rises from 0.765 to 0.944 and calibration improves from ECE 0.524 to 0.144, while sensitivity falls from 0.629 to 0.472.

The factorial decomposes the trade. Moving the threshold policy alone (row 1 to row 2) buys most of the specificity gain at a sensitivity cost; the learned head adds calibration and a further specificity increment. The Platt control settles attribution: a single-parameter supervised calibration layer on the *fixed* fusion reaches ECE 0.165, within 0.02 of the full head (Figure 4). The calibration benefit therefore belongs to supervised calibration as such; the head's distinctive contribution reduces to the sensitivity-for-specificity exchange. Per-fold threshold instability corroborates the calibration reading from another angle: the head's max-F1 thresholds ranged from 0.415 to 0.798 across outer folds — a two-fold spread on identically constructed folds.

One further in-domain observation matters for what follows. This benchmark is harder than the corpus the pipeline was developed on: fixed-fusion nested AUC is 0.751 here. A supervised ranking advantage that had appeared on the easier development corpus did not reproduce under the harder public benchmark — the in-domain gain was itself benchmark-dependent, which is the pattern the transfer experiments make systematic.

N.2 Zero-shot transfer: the fixed fusion ranks better

Table 16 freezes the HUP-trained (B) deployment system and applies it to SWEC-ETHZ (A), the multi-center cohort (D), and Pennsieve-530 (E) (Figure 6). The fixed fusion is the better ranker on A and D: on A, $\Delta\mathrm{AUC} = +0.050$ with CI $[+0.024, +0.083]$ — significant, and stable when the head is retrained under seeds 43 and 44 ($+0.041$ $[+0.017, +0.069]$; $+0.043$ $[+0.018, +0.071]$). On D the

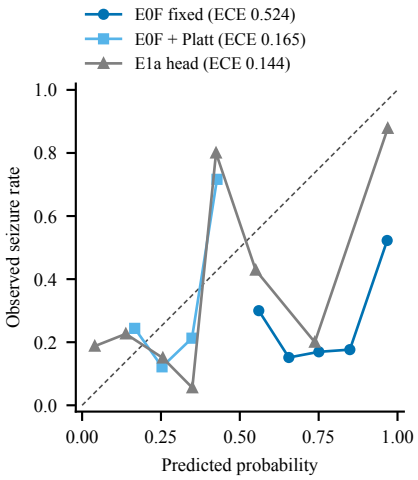


Figure 4: Reliability diagrams on the nested in-domain folds (bins with ≥ 5 clips). The fixed fusion is severely over-confident; a one-parameter Platt layer on the same fusion nearly matches the learned head's calibration.

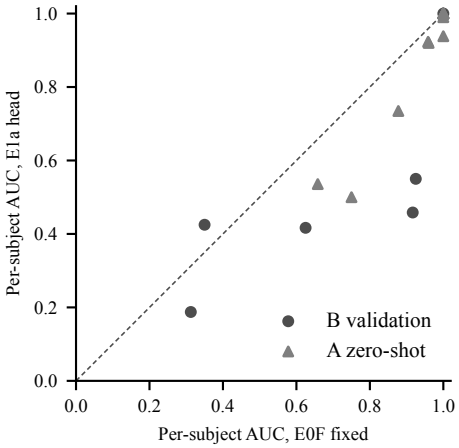


Figure 5: Per-subject AUC, fixed fusion vs. learned head (subjects with both classes), on HUP (B) validation and SWEC-ETHZ (A) zero-shot. Points below the diagonal favor the fusion; the head's failures concentrate on individual patients rather than degrading uniformly.

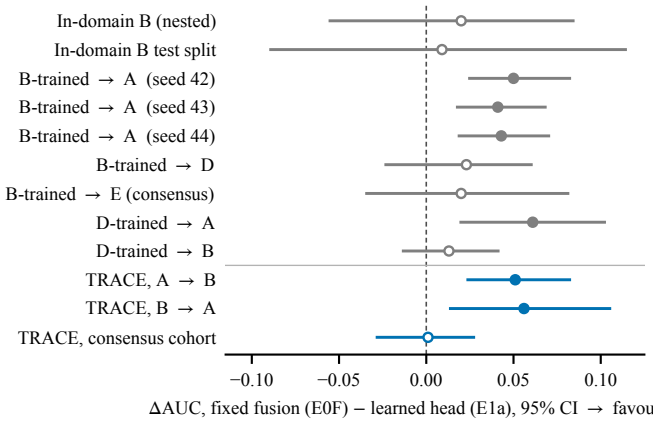


Figure 6: The pre-registered contrasts and their replications: ΔAUC (E0F $-$ E1a) with subject-clustered 95% CIs. Filled points: CI excludes zero. The in-domain contrast is a wash; the transfer advantage of the fixed fusion is significant on A under both trainers and all head seeds, and direction-consistent everywhere. The three grey rows are TRACE's fixed-fusion-minus-learned-head contrasts, recomputed from its released per-clip outputs (Appendix M).

delta is $+0.023$ $[-0.024, +0.061]$, same sign but not significant, and metric-dependent (PR-AUC favors the head, 0.682 vs. 0.621), so we make no claim on D in either direction.

The consensus cohort extends the pattern to a third zero-shot target. On E's primary block the fixed fusion again ranks higher ($\Delta\text{AUC} = +0.020$, CI $[-0.035, +0.082]$, directional), the head repeats its operating-point signature at the frozen thresholds (sensitivity $0.932 \rightarrow 0.818$ for specificity $0.727 \rightarrow 0.841$), and its calibration advantage transfers (ECE 0.092 vs. 0.420). E's peri-ictal negatives make its absolute scores incomparable to the other cohorts', but the within-cohort contrast is unaffected by that framing. (On the exploratory stimulation-induced block both deciders lose roughly 0.15 AUC and the delta is a wash; we make no claims there.)

Table 16: Zero-shot transfer of HUP-trained (B) systems to SWEC-ETHZ (A), the multi-center cohort (D), and Pennsieve-530 (E) (frozen thresholds; forward-only). Rotation: an identically constructed system trained on D and applied to A and to all of B.

Trainer → cohort	Model	AUC	PR-AUC	F1	Sens	Spec
B → A	E0F	**0.948**	**0.946**	**0.851**	**0.800**	0.920
	E1a	0.898	0.903	0.779	0.670	**0.950**
B → D	E0F	**0.767**	0.621	**0.649**	**0.878**	0.500
	E1a	0.744	**0.682**	0.596	0.714	**0.586**
B → E (spont.)	E0F	**0.898**	**0.885**	**0.845**	**0.932**	0.727
	E1a	0.879	0.862	0.828	0.818	**0.841**
D → A	E0F / E1a	0.946 / 0.885	$\Delta\mathrm{AUC} = +0.061$, CI $[+0.019, +0.103]$			
D → B	E0F / E1a	0.746 / 0.733	$\Delta\mathrm{AUC} = +0.013$, CI $[-0.014, +0.042]$			

The pre-committed rotation arm repeats the comparison with D as trainer. The direction holds on both targets and is significant again on A ($+0.061$, CI $[+0.019, +0.103]$). Across the five trainer–cohort pairs there is no reversal. Two caveats are part of the result. First, A is the easiest ranking cohort in the suite — zero-shot AUC on A exceeds in-domain AUC on B — and it is where both significant results occur; the harder multi-center shift and the consensus cohort stay directional. Second, with at most three zero-shot cohorts per trainer this is replication of a direction, not a survey.

Corroboration under TRACE (secondary). TRACE, an independently configured implementation of the same evidence design (23 relative features, a learned XGBoost head, trained jointly on SWEC and HUP), repeats the aggregation contrast with deliberately different construction choices (Appendix M); we recomputed every number below from its released per-clip outputs with our own code. Under patient-clustered bootstrap, the fixed fusion minus the learned head is $+0.051$, CI $[+0.023, +0.083]$, in the SWEC→HUP direction and $+0.056$, CI $[+0.013, +0.106]$, in HUP→SWEC: the same ordering and a similar magnitude as the B-trained earlier configuration, from a different feature set, joint training, and its own thresholds. The fixed-minus-logistic gap is not significant in any direction, matching the capacity ordering of the earlier configuration. On the consensus cohort the two deciders tie ($+0.001$, CI $[-0.029, +0.028]$). This is corroboration, not strict replication: the configuration differs by design, and the shared cohort was previously exposed to the earlier configuration (Appendix M). Under one shared threshold rule the head's specificity edge shrinks in the SWEC–HUP directions (0.850 vs. 0.930 on HUP→SWEC) but widens on the consensus cohort; we therefore claim only the ranking direction.

A capacity control (secondary). To separate "learned head" from "this head family," we repeated the full nested protocol with a logistic head on the same four signals. In-domain it matches both alternatives (nested AUC 0.745; ECE 0.084 — supervision's calibration benefit is head-family-independent). In transfer, SWEC orders by learned capacity: fixed 0.948, logistic 0.930, XGBoost head 0.898, with the fixed-vs-logistic gap no longer significant ($+0.018$, CI $[-0.011, +0.047]$); on the multi-center cohort the fixed rule beats even the logistic head ($+0.039$, CI $[+0.006, +0.074]$). Each transfer cohort thus certifies a fixed-over-learned contrast against a different head family, and the transfer penalty on SWEC grows monotonically with the head's degrees of freedom. As an analysis added after pre-registration, we label it secondary and draw no primary claims from it.

A deeper architecture, pre-registered to fail informatively (secondary). A DeepSets/attention ensemble on the same four signals, with an optional subject-adversarial objective (Ganin et al., 2016), was pre-registered against a gate of zero-shot AUC ≥ 0.93 on SWEC and ≥ 0.75 on the multi-center cohort. Neither arm cleared it (SWEC 0.90–0.91, multi-center 0.66); on the multi-center cohort the richer aggregator is certified worse than the fixed fusion ($\Delta\mathrm{AUC} = +0.107$, CI $[+0.054, +0.153]$) and the shallow head. Full protocol, the inert adversary, and per-arm tables are in Appendix R.

Where the head's deficit lives. Three diagnostics agree. The fusion's transfer is trainer-invariant — AUC 0.948 (B-trained) vs. 0.946 (D-trained) on A — while the head's varies (0.898 vs. 0.885): what transfers is carried by the shared representation, and the part that varies is what the head learned from its trainer. Per-subject analysis of the B validation split (Figure 5) finds two subjects hard for both deciders, and two more where only the head collapses (per-subject AUC 0.92 → 0.46 and

0.93 → 0.55) — the head's failures concentrate on patients, consistent with absorbed subject-level structure rather than uniform degradation. And the per-fold threshold spread of §N.1 shows the head's probability scale is unstable across subject resamples even in-domain.

N.3 The representation control: contrast features carry the transfer

If the fusion's transfer advantage rests on the shared representation, degrading the representation's transfer properties should hurt both deciders — and it should be possible to show this with in-domain information held constant. We retrain the entire pipeline on the 20 acquisition-sensitive features excluded from the stable subset, and on the full 33-feature union, under the identical protocol. In-domain, the sensitive features match the contrast features exactly ($\Delta\mathrm{AUC} = +0.001$, CI $[-0.038, +0.036]$): they are equally informative where acquisition matches. Zero-shot, they are worse everywhere, significantly so on the multi-center cohort ($+0.046$, CI $[+0.004, +0.079]$; A: $+0.025$, CI $[-0.046, +0.079]$). The pre-registered collapse threshold (≥ 0.10) was not met: the effect is real and modest, not catastrophic — resampling to a common rate already removes much acquisition variance. The union arm matches the contrast arm on every cohort: tree ensembles ignore untransferable features when transferable ones are present, so the cost appears only when a model must rely on the sensitive set. The design lesson is correspondingly modest: within-clip contrast features are what carry cross-center transfer, but excluding absolute features is unnecessary when both are available.

The reference switch, scored at onset (secondary). TRACE's matched switch holds everything but the features' reference fixed: the same clips, models, and training, with features computed against the clip's own first 8 s or on absolute scales. Patient-level oracle onset error improves under the episode-local reference in every transfer setting — by 7.23 s in SWEC→HUP and 3.35 s in HUP→SWEC — and on the consensus cohort's spontaneous, pre-registered block by 3.62 s, CI $[1.32, 6.04]$, present under all five training seeds. The contrast was selected after the onset results were visible; the pre-registered primaries remain the two detection contrasts. One structural caveat: positive onsets lie in $[10, 40]$ s by construction, so the first-8 s reference never overlaps the onset it must predict — the gain is specific to clips that begin pre-ictally. Detection is not claimed: the switch's effect is mixed across directions and strata (the relative-featured fusion ranks higher than the absolute-featured primary on the consensus cohort, 0.877 vs. 0.855; the absolute features rank higher in one transfer direction), so the certified gains are onset-scoped.

N.4 Onset localization: a capability boundary

The pipeline includes an interpretable onset locator: a self-calibrated sustainment curve over $E(t, c)$, with onset taken by anchored backtracking from the most reliable sustained region (B2), against a first-crossing rule (B0). On public labels we evaluate both against label-tuned changepoint baselines (CUSUM, Page–Hinkley, PELT, binary segmentation, Bayesian online changepoint detection, and a double-threshold rule; hyperparameters selected leave-one-subject-out on HUP (B)-train only).

What decides onset performance is the cohort and label regime, not decoder sophistication. Rule rankings flip across splits and cohorts: on the training folds a plain PELT changepoint leads (oracle MAE 4.97 s against the anchored rule's 7.66 s), while on the held-out test split the anchored rule leads (5.90 vs. 6.88 s); on the constructor-embedded cohort SWEC-ETHZ (A) the anchored rule is the best of the rule set (6.36 s, better than PELT on 13 of 16 subjects; secondary paired Wilcoxon $p = .003$); on the single-rater clinical multi-center cohort (D) the leading rules are indistinguishable (B2, PELT, binary segmentation, and Page–Hinkley within 0.3 s, Table 18); and on the consensus cohort Pennsieve-530 (E), where only the frozen deployed rules exist under the single-pass protocol, the anchored rule attains its best transfer MAE on the spontaneous block, while the exploratory stimulation-induced block shows substantially higher errors and no claims are made there. Errors are heavy-tailed everywhere — median errors run 2–3 s below means — so a subset of clips localizes well while a tail fails badly. The pre-registered demotion of onset from our claims, decided before E existed, stands; the durable, certified onset effect in this paper is the evidence-reference switch of §N.3, not any single decoder's absolute error. Supervised regression on a single homogeneous platform reports 3.12 s median error (Arcot Desai et al., 2023); whether the gap reflects labels, hardware homogeneity, or supervision is a question our released benchmark makes testable.

TRACE sharpens the onset boundary in three ways, all secondary. Its gains are oracle-estimand; under the deployed estimand, where a missed seizure is charged the full 60 s, the consensus cohort's spontaneous gain shrinks to $+3.08$ s with a CI covering zero ($[-0.21, +6.44]$) — real for localization, not yet for a deployable detector. The strata behave as TRACE's pre-registration anticipated: the spontaneous block carries the measurable effect ($+3.62$ s, CI $[1.32, 6.04]$), the stimulation-evoked block an uncertain one (-0.71 s, CI $[-3.81, +2.45]$). And every deep onset baseline localizes worse than TRACE on the pooled Pennsieve-530 positives (7.92 vs. 8.47 s oracle MAE for the best baseline's three-seed mean); those baselines were tuned under their own protocol (Appendix M), so the comparison covers the evaluated implementations, not deep detectors in general.

O Complete results tables (earlier configuration)

Table 17 extends the main-text tables to every head, cohort, and metric; Table 19 collects every subject-clustered contrast in one place with its pre-registration status; Table 18 reports all thirteen onset rules (on cohort E, only the frozen deployed rules exist under the single-pass protocol). All three tables are emitted by a script that reads only the released per-clip prediction files.

Table 17: Complete detection metrics for every head and cohort. E0F/E1a rows use the frozen thresholds of the main text (legacy for E0F, max-F1 scan for E1a); LR is the secondary logistic head; E0F+Platt rows are calibration-only cells (nominal 0.5 threshold). Daggered rows are cohort E's exploratory stimulation-induced block. All values regenerate from released per-clip predictions.

Block	n	pos	AUC	PR-AUC	F1	Sens	Spec	Brier	ECE
E0F | B-train-OOF	285	89	0.751	0.718	0.586	0.629	0.765	0.467	0.524
E1a | B-train-OOF	285	89	0.731	0.679	0.592	0.472	0.944	0.174	0.144
E0F | B-val	71	23	0.744	0.622	0.450	0.391	0.833	0.420	0.478
E1a | B-val	71	23	0.585	0.523	0.333	0.261	0.854	0.267	0.256
E0F | B-test	60	20	0.829	0.815	0.750	0.750	0.875	0.421	0.495
E1a | B-test	60	20	0.820	0.810	0.667	0.550	0.950	0.142	0.125
E0F | A-zeroshot	200	100	0.948	0.946	0.851	0.800	0.920	0.268	0.337
E1a | A-zeroshot	200	100	0.898	0.903	0.779	0.670	0.950	0.137	0.118
E0F | D-zeroshot	260	98	0.767	0.621	0.649	0.878	0.500	0.487	0.527
E1a | D-zeroshot	260	98	0.744	0.682	0.596	0.714	0.586	0.308	0.289
E0F+Platt | B-train-OOF	285	89	0.727	0.600	0.000	0.000	1.000	0.195	0.165
E0F+Platt | B-val	71	23	0.744	0.622	0.000	0.000	1.000	0.196	0.046
E0F+Platt | B-test	60	20	0.829	0.815	0.000	0.000	1.000	0.182	0.178
E0F+Platt | A-zeroshot	200	100	0.948	0.946	0.000	0.000	1.000	0.206	0.306
E0F+Platt | D-zeroshot	260	98	0.767	0.621	0.000	0.000	1.000	0.211	0.100
LR | B-train-OOF	285	89	0.745	0.720	0.584	0.584	0.811	0.158	0.084
LR | B-val	71	23	0.793	0.716	0.611	0.478	0.958	0.163	0.123
LR | B-test	60	20	0.864	0.824	0.684	0.650	0.875	0.127	0.067
LR | A-zeroshot	200	100	0.930	0.930	0.860	0.860	0.860	0.114	0.089
LR | D-zeroshot	260	98	0.728	0.594	0.621	0.878	0.426	0.264	0.219
E0F | E-spont.	88	44	0.898	0.885	0.845	0.932	0.727	0.374	0.420
E1a | E-spont.	88	44	0.879	0.862	0.828	0.818	0.841	0.146	0.092
E0F | E-stim†	68	34	0.740	0.743	0.700	0.824	0.471	0.419	0.441
E1a | E-stim†	68	34	0.753	0.657	0.633	0.559	0.794	0.212	0.161

P Protocol details (earlier configuration)

Folds and nesting. The 25 B-train subjects map to five outer folds by sorted-index modulo five; the assignment is hard-coded and runs abort on any mismatch with the loaded data. Within each outer iteration, frame models (three seeds), the head (inner subject-grouped 3-fold CV over depth $\in \{2, 3\}$, trees $\in \{100, 200\}$ for XGBoost; $C \in \{0.1, 1, 10\}$ for the logistic control), and both thresholds are fitted with no access to the held-out fold; the head's training evidence is generated by the same fold-excluded frame models. Selected inner-CV AUCs ranged 0.759–0.829 across folds; selected max-F1 thresholds ranged 0.415–0.798, the instability discussed in §N.1.

Contamination gates. Runs assert: unique clip identifiers across all cohorts; subject-disjoint B splits; expected cohort sizes (285/71/60/200/260/156); subject-registry disjointness between cohorts; and bit-level agreement of recomputed onset decisions with stored artifacts before any baseline runs (285/285 on B-train evidence). The cohort-E forward pass additionally required the re-instantiated deployment system to reproduce the frozen B-val/B-test/A/D per-clip probabilities to a maximum

Table 18: Onset oracle MAE (s) for every rule and cohort. Baseline hyperparameters are selected leave-one-subject-out on B-train positives only (unanimity 21–23/23); B2's retuned configuration changes one inherited constant (sustainment $3.0 \rightarrow 1.5$ s, selected 23/23). Cohorts are never pooled. On E (spontaneous block) only the frozen deployed rules were run under the single-pass protocol.

Rule	B-train	B-val	B-test	A	D	E (spont.)
B2 anchored (retuned)	7.66	9.04	5.90	6.36	6.91	5.14
B2 anchored (inherited)	8.09	9.04	5.90	6.57	6.91	5.14
B0 first-crossing	8.62	12.55	8.02	7.10	7.61	6.52
PELT (ℓ_2)	4.97	6.88	6.88	6.84	6.79	—
PELT anchored	4.97	6.88	6.88	6.84	6.79	—
Binary segmentation	8.31	8.55	9.43	6.66	7.03	—
CUSUM	7.78	8.53	5.98	6.43	7.32	—
Page–Hinkley	6.07	8.25	6.74	6.68	6.93	—
Double threshold	8.36	10.31	7.27	6.96	7.78	—
BOCPD	7.49	12.09	7.88	8.22	7.35	—
Half-max rise	8.73	13.34	8.02	7.17	8.03	—
Constant 25 s	7.68	7.53	7.10	7.76	7.49	7.19
Train-mean onset	7.68	7.52	7.12	7.76	7.48	—

Table 19: All subject-clustered ΔAUC contrasts (E0F − head unless noted; 2,000 resamples). Contrasts marked † are secondary analyses added after pre-registration.

Contrast	ΔAUC [95% CI]	Status
In-domain B, nested (primary 1)	$+0.020$ $[-0.056, +0.085]$	no evidence of gain
In-domain B test split	$+0.009$ $[-0.090, +0.115]$	descriptive (5 subjects)
B-trained → A, seed 42 (primary 2)	$+0.050$ $[+0.024, +0.083]$	significant
B-trained → A, seed 43	$+0.041$ $[+0.017, +0.069]$	significant
B-trained → A, seed 44	$+0.043$ $[+0.018, +0.071]$	significant
B-trained → D	$+0.023$ $[-0.024, +0.061]$	directional
B-trained → E (consensus, spont.)	$+0.020$ $[-0.035, +0.082]$	directional
B-trained → E (stim)†	-0.012 $[-0.103, +0.070]$	exploratory
D-trained → A (rotation)	$+0.061$ $[+0.019, +0.103]$	significant
D-trained → B (rotation)	$+0.013$ $[-0.014, +0.042]$	directional
B-trained → A vs. logistic head†	$+0.018$ $[-0.011, +0.047]$	directional
B-trained → D vs. logistic head†	$+0.039$ $[+0.006, +0.074]$	significant
Ablation stable−sensitive, in-domain B	$+0.001$ $[-0.038, +0.036]$	parity
Ablation stable−sensitive, A	$+0.025$ $[-0.046, +0.079]$	directional
Ablation stable−sensitive, D	$+0.046$ $[+0.004, +0.079]$	significant
Ablation stable−full (33), in-domain B	-0.020 $[-0.046, +0.003]$	parity
Onset: PELT − constant, in-domain B	-2.71 s $[-4.55, -0.73]$	significant
Onset: B2 anchored − constant, E spont.	-2.05 s $[-3.25, -0.59]$	significant
Onset: B2 anchored − constant, E stim†	$+4.73$ s $[+2.07, +7.58]$	certified-worse (exploratory)

deviation of 1.1×10^{-16} and the frozen thresholds exactly, before E features were read; any mismatch aborts with no E output.

Deployed accounting. Deployed onset error charges each missed positive the full 60 s clip length; Table 17's sensitivity column determines the deployed–oracle gap, and we report oracle errors in §N.4 because rule choice does not alter the detector.

Environments. Rebuild experiments: Linux, Python 3.12, numpy 2.5.2, scikit-learn 1.9.0, xgboost 3.4.1, ruptures 1.1.10; per-run manifests record exact versions. Head-seed replication used seeds {42, 43, 44}; frame-model ensembles use the same three seeds throughout.

Q Pre-registration and amendments (earlier configuration)

The protocol, endpoints, and decision rules were frozen in a dated document before the first public-benchmark result existed, with four dated amendments, each made before the result it governs: (1) replacement of a development-era composite objective with inner-CV AUC selection and max-F1 threshold scanning; (2) fully nested per-fold head fitting after an external code review identified selection bias in a draft evaluation; (3) permission to re-select the onset locator's constants on B-train by the same leave-one-subject-out procedure that originally produced them, with mandatory uninformative baselines in every onset table; (4) cohort E's negative-window rule, amended from the planned ≥ 90 s interictal guard when the source recordings' fixed 120 s padding made that guard

geometrically impossible, frozen before any E clip was built. Two pre-registered kill criteria executed as written: the onset contribution was demoted from the abstract and contribution list when no rule cleared the constant baseline with clustered significance on transfer cohorts, and the SWEC transfer claim retained certified language only after surviving the three-seed replication. The uninformative-predictor comparisons the protocol mandated are included in the released per-clip artifacts; following the reporting convention of the onset-localization literature, the main text evaluates rules against label-tuned changepoint and matched-representation comparators. Cohort E's results postdate the demotion decision, which stands; E's onset results are reported in §N.4 as a boundary condition, not a reinstated claim. The logistic-head control was added at review and is labeled secondary throughout.

R DEEP AGGREGATOR GATE EXPERIMENT (EARLIER CONFIGURATION)

Full protocol: a 9-model CNN ensemble (3 torch seeds $\times$ 3-fold inner selection) reading the same four-signal per-clip evidence as E0F/E1a, with an optional subject-adversarial branch (gradient-reversal layer, 20 training subjects per outer fold; site labels are not used, so we name this arm subject- rather than site-adversarial). Pre-registered before any real-data result existed; an external design review found three gate-changing defects — a doubled gradient-reversal scaling, caller-controlled rather than manifest-derived gate eligibility, and an undefined multi-fold subject probe — all fixed and re-frozen (dated amendment) before the registered run. Gate rule: zero-shot AUC ≥ 0.93 on SWEC and ≥ 0.75 on the multi-center cohort, with in-domain OOF ≥ 0.74, evaluated on the base and $\lambda_{\max} = 0.3$ arms; $\lambda_{\max} \in \{0.1, 1.0\}$ and a same-evidence logistic control are non-gating sensitivity checks.

Table 20: Deep aggregator gate results (subject-clustered bootstrap, 2,000 resamples). ΔAUC is E0F $-$ arm, matching the paper-wide convention. Neither eligible arm clears the pre-registered gate; the multi-center deficit is certified. Daggered rows are sensitivity checks, not gate-eligible.

Arm (gate-eligible)	SWEC AUC	Multi-ctr. AUC	In-dom. OOF	ΔAUC (E0F $-$ arm, multi-ctr.)	Gate
Base (yes)	0.901	0.660	0.790	$+0.107$ $[+0.054, +0.153]$	FAIL
Subj.-adv. λ=0.3 (yes)	0.912	0.662	0.782	$+0.105$ $[+0.058, +0.147]$	FAIL
Subj.-adv. λ=0.1†	0.902	0.657	0.776	$+0.110$ $[+0.054, +0.156]$	FAIL
Subj.-adv. λ=1.0†	0.905	0.663	0.786	$+0.103$ $[+0.056, +0.147]$	FAIL
LR on pooled evidence†	0.856	0.633	0.800	$+0.133$ $[+0.071, +0.187]$	FAIL
Fixed fusion E0F (floor)	**0.948**	**0.767**	0.751	—	—
XGB head E1a (paper's arm)	0.898	0.744	0.731	—	—

The negative is certified, not merely sub-threshold: the multi-center ΔAUC CI excludes zero for every aggregator arm against both E0F and E1a. Per-torch-seed AUCs are tight (multi-center 0.642–0.668 across seeds), so the deficit is not a seed artifact. In-domain, the aggregator matches or slightly beats both deciders (beats E1a by 0.059 AUC, CI $[+0.002, +0.116]$) — its capacity is real and usable in-domain; it does not transfer. The subject-adversarial objective moved nothing: 25-way subject decodability of the deployed embedding was 0.219 (base) vs. 0.213–0.220 (adversarial arms, chance 0.04), because the adversary head itself never fit well at this scale (train balanced accuracy plateaued at 0.10–0.15 even at the terminal schedule value); we read this as an underpowered objective at 25 training subjects, not a claim against identity-erasure training in general. This experiment does not appear in the main-text contribution list: it was pre-registered as a probe of whether the paper's central pattern is an artifact of shallow aggregators, and the certified answer — a materially deeper architecture reproduces and slightly sharpens the deficit — tightens rather than qualifies the paper's claim.