Behavioral Latency as Weak Event-Time Supervision for EEG Reaction-Time Decoding

Anuar Aimoldin[1*]
anuar.aimoldin@mbzuai.ac.ae

Ayana Mussabayeva[1*]
ayana.mussabayeva@mbzuai.ac.ae

Yedige Mussabayev[2]
ymussabayev@hof-university.de

Xue Liu[1,3]
steve.liu@mbzuai.ac.ae

Kun Zhang[1,4]
kun.zhang@mbzuai.ac.ae

[1]Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, UAE
[2] Hof University, Hof, Germany
[3] McGill University, Montreal, QC, Canada
[4] Carnegie Mellon University, Pittsburgh, PA, USA
* Equal contribution

This is the authors' final version of the manuscript accepted for publication in Neural Computation.

Keywords: weak event-time supervision, event-time posterior modeling, EEG reaction-time decoding, behavioral latency, shifted-crop diagnostics

Abstract

Single-trial EEG analyses are often organized around events and latencies, such as stimulus onset, response execution, ERP component timing, and movement onset. Yet EEG-based reaction-time prediction is commonly posed as scalar regression on a fixed stimulus-locked window, so reaction time is treated as a window-level label rather

than timing evidence about response-relevant dynamics. Here we reformulate trial-wise reaction-time decoding as event-time posterior modeling. Instead of predicting a single RT directly, the model estimates a posterior distribution over response-relevant event times, $p(t_{\text{event}} \mid X)$, and uses the posterior mean as the RT estimate. This treats behavioral latency as a weak observation of latent response-relevant timing, making the output representation itself part of the computational hypothesis.

We evaluate this formulation on the Healthy Brain Network contrast change detection EEG task under a subject-disjoint, release-separated protocol. Across five seeds, distributional event-time supervision consistently improves held-out RT prediction relative to scalar regression and temporal-readout controls. Controlled objective comparisons isolate supervision of the event-time distribution, rather than expectation-based readout alone, as the source of this gain; architecture controls establish that the effect persists across four temporal backbones and is not explained by model scale.

Beyond point prediction, the event-time posterior serves as a diagnostic for separating temporal evidence from scalar shortcuts. Posterior geometry characterizes temporal concentration, target alignment, and interval behavior, while observation-noise calibration separates latent event-time concentration from predictive uncertainty over observed RT. We also use shifted-crop inference to probe shortcut use versus temporal localization. A matched shift-jitter intervention improves shifted-crop robustness, increases mean sensitivity, and moves predictions more often in the expected crop-relative direction. Sensitivity remains below ideal crop-relative localization, leaving a clear equivariance gap. Together, these results establish event-time posterior modeling as a probabilistic and interpretable formulation for linking single-trial EEG dynamics to behavioral timing.

1 Introduction

Behavioral timing links neural dynamics to observable behavior. In reaction-time (RT) tasks, the observed response latency reflects sensory encoding, decision formation, motor preparation, and response execution unfolding over hundreds of milliseconds. EEG analyses often exploit this temporal structure through event alignment, response locking, and the study of latency variability. These trial-to-trial timing differences can carry behaviorally meaningful information that may be lost through averaging or when models reduce temporal structure to window-level labels (Jung et al., 1998; Ouyang et al., 2017).

Supervised EEG decoding, however, often treats a fixed stimulus-locked window as an input to a class label or scalar behavioral variable. For RT decoding, this means that a delayed behavioral response latency is treated as a property of the whole EEG segment. A scalar predictor may therefore achieve low error by exploiting individual differences in typical reaction time, trial difficulty, or stimulus-locked temporal priors, without representing when response-relevant dynamics occur within the window.

We address this mismatch by treating behavioral RT as weak event-time supervision. The button press provides noisy behavioral timing evidence about response-relevant dynamics, reflecting the combined effects of sensory, decision, and motor processes rather than a direct neural event annotation. Given an EEG window X, the model estimates a posterior distribution $p(t_{\text{event}} \mid X)$ over response-relevant event times and derives the scalar RT prediction as the posterior expectation. This keeps compatibility

with standard scalar metrics such as nRMSE while making the timing semantics of RT explicit.

We study this question on the Healthy Brain Network contrast change detection EEG task (Shirazi et al., 2024), following the RT prediction setting introduced by the EEG Foundation Challenge (Aristimunha et al., 2025; EEG Challenge, 2025). Under a subject-disjoint, release-separated protocol, we compare direct scalar regression, temporal-readout regression, soft-argmax RT-loss controls, and distributionally supervised event-time models. This design separates the effects of architecture capacity, expectation-based readout, scalar timing loss with a temporal head, and distributional posterior supervision. This output-representation focus complements advances in EEG architectures, self-supervised pretraining, and foundation-style EEG models for encoding EEG across subjects, tasks, and recording setups (Wang et al., 2024a; Jiang et al., 2024; Chen et al., 2024; Wang et al., 2025; Ouahidi et al., 2025; Yue et al., 2025); representative backbones serve as architecture controls within the same evaluation.

The posterior representation also expands what can be evaluated on each trial. Posterior geometry summarizes temporal concentration, target alignment, and interval behavior beyond scalar error, while observation-noise calibration separates latent event-time concentration from predictive uncertainty over observed RT. Shifted-crop inference provides a complementary test of temporal behavior: a crop-relative timing localizer should move its prediction with the temporal frame, whereas a fixed-timing scalar shortcut should remain comparatively invariant.

Distributional event-time supervision consistently improves held-out prediction over scalar and temporal-readout controls across five seeds, and the soft-argmax ablation attributes this advantage to distributional supervision rather than expectation readout alone. Posterior geometry and shifted-crop analyses reveal temporal evidence and partial response-timing localization that scalar metrics do not capture, while shift-jitter training further improves shifted-crop robustness and expected-direction movement. The remaining sensitivity gap provides a clear target for fully equivariant temporal models. Although our experiments focus on RT prediction, the same event-time view is relevant to EEG targets with latency semantics, including ERP latency estimation, movement onset decoding, error-related responses, and other latency-varying phenomena.

Our contributions are:

1. We recast behavioral RT as weak event-time supervision and formulate EEG RT decoding as event-time posterior modeling rather than direct scalar regression, with soft-target distribution matching and EventNLL latent-event likelihoods as complementary realizations of this formulation.

2. We evaluate this formulation under a subject-disjoint, release-separated HBN-EEG protocol and controlled comparison designed to separate EEG backbone capacity, temporal readout parameterization, and distributional event-time supervision.

3. We show through controlled baselines and loss ablations, repeated across five seeds and four dense temporal backbones, that distributional event-time supervision improves over scalar regression, temporal-readout regression, and soft-argmax RT-loss controls.

4. We use the learned event-time posterior for temporally resolved evaluation: posterior geometry and observation-noise calibration separate scalar accuracy from temporal concentration, predictive uncertainty, and interval behavior, while shifted-crop diagnostics test shortcut behavior and reveal partial response-timing localization; matched shift-jitter training then strengthens crop-relative behavior.

2 Related Work

2.1 Event-Centered EEG Analysis and Latency Variability

EEG is often analyzed through events and latencies. Stimulus onset, response execution, error-related activity, movement onset, ERP component timing, and other temporally localized phenomena define not only what occurs in a trial, but also when task-relevant neural or behavioral processes unfold. This event-centered view appears in annotation systems such as Hierarchical Event Descriptors (Bigdely-Shamlo et al., 2016) and in single-trial ERP analysis, where ERP-image visualization and related methods show that sorting trials by RT can reveal performance-linked temporal structure (Jung et al., 1998). Methods for single-trial ERP analysis further show that within-subject latency variation can be estimated and linked to behavior (Ouyang et al., 2011, 2017).

EEG activity associated with responses and movements provides related examples. Lateralized readiness potentials reflect response preparation and execution (Coles, 1989). Error-related negativities are response-locked markers of error processing and performance monitoring (Gehring et al., 1993; Holroyd and Coles, 2002). Movement-related cortical potentials, including the Bereitschaftspotential, characterize neural activity around movement onset and during premovement preparation (Kornhuber and Deecke, 1965; Shibasaki and Hallett, 2006). Decision-time models make a related point at the behavioral level by treating RT variability as a consequence of latent temporal dynamics in evidence accumulation and decision formation (Ratcliff and McKoon, 2008; O'Connell et al., 2012; Kelly and O'Connell, 2013). At the EEG level, single-trial hidden multivariate pattern analyses make a parallel point by decomposing decision-task EEG into recurrent events with trial-varying latencies (Weindel et al., 2025). Rather than fitting a cognitive process model or an explicit sequence of neural events, we use observed RT as a behavioral latency that constrains an unobserved response-relevant event time.

Explicit EEG event detection is also well established when annotated onsets, offsets, or intervals are available. Sleep-event detectors, for example, estimate the timing of transient sleep EEG patterns such as sleep spindles and K-complexes rather than predicting only a window-level label (Chambon et al., 2019; Tapia-Rivas et al., 2024). In our setting, scalar RT serves as weak timing evidence rather than a manually annotated neural event time.

2.2 Fixed-Window EEG Decoding of Latency-Structured Targets

Modern supervised EEG decoding often maps a fixed stimulus-locked or task-locked EEG window to a scalar or class label. This formulation is convenient and aligns with benchmark metrics, but it can collapse temporal structure when the target is itself a

latency. RT is evaluated as one scalar per trial, yet it denotes the latency of a behavioral response following stimulus onset.

EEG-based RT estimation has commonly followed this scalar-prediction form, including regression pipelines based on Riemannian geometry features (Wu et al., 2017) and deep single-trial EEG models for visual-stimulus RT prediction (Chowdhury et al., 2020). The EEG Foundation Challenge provides a standardized scalar benchmark using normalized RMSE (Aristimunha et al., 2025; EEG Challenge, 2025). We keep this scalar evaluation protocol, but ask whether the same label can be used within the model as weak event-time supervision. This distinction matters because a scalar regressor may learn trial difficulty, subject-level response tendency, or stimulus-locked temporal priors without representing where response-relevant dynamics are expressed. A temporal soft-argmax readout alone also does not resolve the issue if the model is still trained only through scalar RT error.

2.3 Distributional and Time-to-Event Output Modeling

Latency-structured EEG targets also motivate output distributions rather than only point estimates. Label distribution learning replaces a hard label with a distribution over related labels, which is useful when neighboring labels share metric structure or when ambiguity is meaningful (Geng, 2016). For RT, neighboring labels are ordered time bins, making a smoothed event-time distribution a natural supervision target.

Time-to-event modeling provides a complementary probabilistic language for ordered event times. Classical survival analysis and neural time-to-event models estimate distributions over event times rather than only scalar predictions or risk scores (Cox, 1972; Lee et al., 2018). Distributional evaluation also makes uncertainty and calibration observable through posterior width, coverage, likelihood scores, and proper scoring rules (Chapfuwa et al., 2023; Haider et al., 2020). We adapt this perspective to finite EEG windows, rather than treating the task as a standard survival-analysis problem with long-horizon risk, censoring, or competing clinical outcomes. EventNLL instantiates this view by treating observed RT as a noisy realization of a latent response-relevant event time.

2.4 Input Representations, EEG Backbones, and Pretraining

A separate line of EEG decoding work addresses cross-subject, cross-session, and cross-dataset variability through stronger input representations. Classical approaches include Riemannian covariance-based decoding and alignment methods for reducing subject or session mismatch (Barachant et al., 2012; Jayaram et al., 2016; Zanini et al., 2018; He and Wu, 2020). Deep learning approaches pursue related goals through domain adaptation, moment matching, supervised EEG architectures, self-supervised pretraining, and foundation-style EEG backbones (Ganin et al., 2016; Sun and Saenko, 2016; Kostas et al., 2021; Wang et al., 2024a; Jiang et al., 2024; Chen et al., 2024; Wang et al., 2025; Ouahidi et al., 2025; Yue et al., 2025).

Our contribution advances an output-representation axis that is complementary to these input-representation approaches. This distinction remains important for recent EEG foundation models: they primarily address transferable input representations, whereas

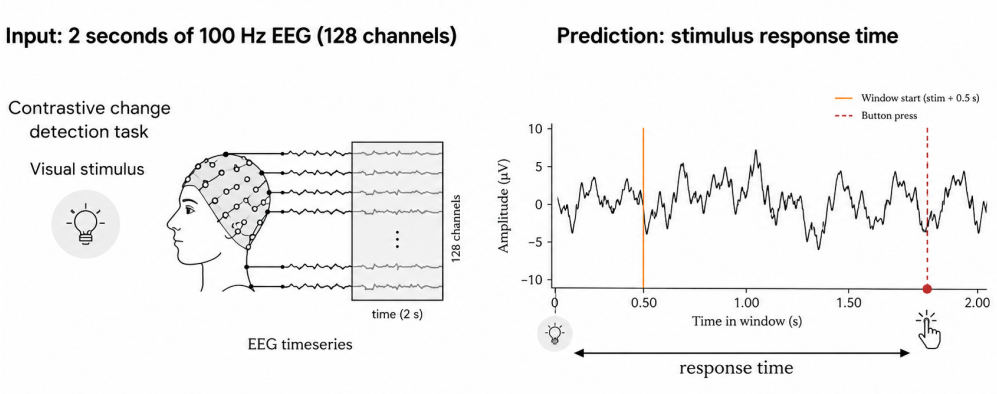

Figure 1: HBN-EEG contrast change detection reaction-time task. Each trial provides a stimulus-locked EEG segment and a trial-wise RT target from stimulus onset to button press.

our question concerns the output representation appropriate for latency-labeled EEG. Our focus is the representation and supervision of latency-structured outputs: when the target is a latency, the label is not only a scalar value to regress, but also weak evidence about when response-relevant dynamics are expressed. External EEG architectures and foundation-style models trained from scratch therefore serve as controlled architecture baselines. What remains missing in prior work is a controlled EEG RT formulation that connects scalar benchmark evaluation, temporal-readout controls, distributional event-time supervision, and posterior-level diagnostics. We fill this gap by learning a posterior over response-relevant latencies and evaluating it as temporal evidence rather than only as a source of a scalar RT estimate.

3 Dataset and Evaluation Protocol

3.1 HBN-EEG CCD Reaction-Time Task

We use the reaction-time prediction task introduced by the NeurIPS EEG Foundation Challenge (EEG Challenge, 2025; Aristimunha et al., 2025), based on curated releases of the HBN-EEG dataset (Shirazi et al., 2024). Our analysis focuses on the contrast change detection (CCD) reaction-time subset.

In the CCD task, participants viewed two flickering grating stimuli and responded after a contrast change made one grating perceptually dominant. The released target is the trial-wise reaction time from stimulus onset to button press. Each main experiment uses the same fixed 2 s stimulus-locked EEG window spanning 0.5–2.5 s after stimulus onset, sampled at 100 Hz from 128 EEG channels after excluding the reference channel. Thus each input has $T = 200$ samples. Because RT marks the response time within this post-stimulus window, the task admits an event-time interpretation in addition to a window-level regression formulation.

Table 1: Release-separated CCD protocol statistics after window construction and RT-support filtering. Prepared trials are 2 s CCD windows with stimulus and response annotations; analyzed trials additionally satisfy $0.5 \leq \text{RT} \leq 2.5$ s.

Partition	Releases	Prepared trials	Analyzed trials	Subjects
Train	R1–R8	74,576	73,030	1,221
Development	R9–R10	17,867	17,348	308
Holdout	R11	15,751	15,164	292

3.2 Shared Evaluation Protocol

We use this task as a controlled evaluation setting for separating three possible sources of RT prediction performance: EEG backbone capacity, temporal readout parameterization, and distributional event-time supervision. The benchmark therefore fixes the data split, target support, input representation, and scalar evaluation rule before varying architectures and objectives in the subsequent model and diagnostic sections.

To avoid subject leakage, all models use the same release-separated CCD split: R1–R8 for fitting, R9–R10 as the validation split used for development choices such as early stopping, checkpoint selection, readout-temperature tuning, and diagnostic settings for analyses that require them, and R11 for final holdout evaluation after all model and readout choices are fixed. R11 holdout labels are not used for model fitting, checkpoint selection, readout-temperature selection, or protocol decisions. Table 1 reports the resulting partitions; metadata checks confirmed zero subject overlap across train, development, and holdout after filtering.

The main analysis keeps trials whose behavioral RT falls inside the modeled event-time support, 0.5–2.5 s after stimulus onset. This filter removes 1,546/74,576 training trials (2.07%), 519/17,867 development trials (2.90%), and 587/15,751 holdout trials (3.73%). In the prepared CCD split, excluded trials are fast responses below 0.5 s; no prepared trial has RT above 2.5 s. The conclusions therefore apply to RTs within the modeled 0.5–2.5 s post-stimulus interval.

Following the EEG Foundation Challenge evaluation protocol (EEG Challenge, 2025), scalar performance is evaluated using normalized RMSE (nRMSE), computed as RMSE divided by the standard deviation of the true targets within each evaluated split. Main scalar comparisons use the same support-filtered R9–R10 validation set and final holdout set. Diagnostic analyses that require shifted crop starts, stricter RT support, or posterior-only summaries are introduced in the corresponding subsection or appendix and do not affect the main holdout reporting protocol.

Table 2 maps the shared protocol to the comparison blocks used below. The scalar regression controls vary compact temporal regressors and external EEG backbones under scalar RT supervision. We first fix ETS-U-Net and vary output supervision (Section 5.1), and then repeat a reduced objective set across additional backbones to test architecture robustness (Section 5.5). Posterior geometry and shifted-crop inference are post-training diagnostics, while shift-jitter training is a matched intervention; together they evaluate posterior semantics and shortcut-vs-localization behavior within the same benchmark framework.

Table 2: Controlled comparison and diagnostic blocks under the shared protocol.

Control or analysis	Readout	Supervision	Question addressed
Scalar regression controls			
External EEG backbones	Scalar RT head	Scalar RT	Does generic EEG backbone capacity explain performance?
MSP-CNN family	Segment-pooling scalar readout	Scalar RT	How far does coarse temporal pooling support scalar regression?
ETR-CNN family	Temporal expectation readout	Scalar RT	Does a learned temporal readout improve scalar regression?
Event-time objective comparison with fixed ETS-U-Net			
Soft-argmax RT-loss control	Posterior mean	Scalar RT only	Is posterior-mean readout sufficient without distributional supervision?
CE soft-target objective	Posterior mean	Soft event-time target	Does direct distributional supervision improve RT prediction?
Likelihood-based objectives	Posterior mean	Latent event-time likelihood	Does probabilistic latent-event supervision provide the same benefit?
Wasserstein control	Posterior mean	Distributional geometry loss	Is an alternative geometry-based distributional objective sufficient?
Architecture robustness controls			
ETS-TCN	Posterior mean	RT-only, CE, mixture EventNLL	Does the supervision effect persist with dilated temporal convolution?
ETS-InceptionPyramid	Posterior mean	RT-only, CE, mixture EventNLL	Does the supervision effect persist with multi-scale temporal filtering?
ETS-AttnSeg	Posterior mean	RT-only, CE, mixture EventNLL	Does the supervision effect persist with attention and local convolution?
Posterior and localization diagnostics			
Readout-temperature tuning	Temperature-adjusted posterior mean	Learned posterior	How should the scalar posterior readout be standardized across objectives?
Posterior geometry diagnostics	Posterior summaries	Learned posterior	What posterior behavior is hidden by scalar nRMSE?
Shifted-crop diagnostic	Crop-relative posterior mean	No new training	Do predictions behave as crop-relative temporal localizers?
Shift-jitter intervention	Crop-relative posterior mean	Shifted crop-relative targets	Does crop-start augmentation reduce shortcut behavior?

3.3 Shared Training, Readout Tuning, and Inference Protocol

To support the controlled comparison, we keep preprocessing, optimization, augmentation, checkpointing, and inference fixed across supervised neural models unless stated otherwise. Each 2 s trial window is normalized per channel by subtracting the channel's temporal mean within the window and dividing by its temporal standard deviation. The shared optimization and augmentation settings are summarized in Appendix C; in brief, models are trained with Adam (Kingma and Ba, 2015), early stopping on validation nRMSE, and a fixed augmentation profile combining channel dropout, temporal cutout, and Gaussian noise.

Regression models use scalar RT supervision, while event-time models represent RT either as a soft crop-relative event-time target or as an observation under a latent event-time likelihood. When temperature scaling is used for event-time posterior readout, τ is selected on R9–R10 to minimize development nRMSE of the posterior-mean readout and then applied unchanged to the holdout split. The main regression-baseline and event-time-objective comparisons are repeated over five seeds (2025–2029), and repeated-run tables report mean $\pm$ standard deviation across seeds unless otherwise stated.

4 Scalar Regression Controls

Before introducing event-time targets, we first ask how far scalar RT supervision can go under the same release-separated protocol. The regression baselines use compact task-specific temporal regressors as strong scalar controls and external EEG backbones as architecture-capacity controls.

The compact controls form a progression of window-level temporal inductive biases. The multiscale segment-statistic pooling CNN (MSP-CNN; Figure 2) explicitly accounts for temporal structure by collecting mean and max statistics from coarse temporal segments of the feature map. These segment-level summaries act as early/middle/late activation cues and are mapped to scalar RT. The event-time readout CNN (ETR-CNN) instead produces learned temporal scores over the window and predicts RT as their expectation, while training the resulting scalar prediction through RT error. ETR-CNN large keeps the same readout while increasing feature capacity.

We compare these task-specific regressors with external EEG architectures adapted to scalar RT regression. These include classical convolutional EEG baselines, Deep4Net and ShallowFBCSPNet (Schirrmeister et al., 2017); the compact EEGNet architecture (Lawhern et al., 2018); TIDNet's thinker-invariant convolutional design (Kostas and Rudzicz, 2020); the convolutional-transformer EEGConformer (Song et al., 2023); the attention temporal convolutional ATCNet architecture (Altaheri et al., 2023); EEGPT (Wang et al., 2024a); Medformer (Wang et al., 2024b); and the LaBraM foundation-style architecture (Jiang et al., 2024). All external architectures are trained from scratch; pretrained weights are not used.

ETR-CNN large is the strongest scalar baseline, with MSP-CNN and base ETR-CNN close behind. Under this protocol, the evaluated external EEG backbones remain below the task-specific temporal controls, indicating that generic backbone capacity alone does not account for the stronger scalar performance. These results establish ETR-CNN large as the scalar reference for testing whether direct event-time distribution supervision improves beyond scalar RT training.

5 Event-Time Posterior Formulation

To make response timing explicit, we reformulate RT prediction as event-time posterior modeling. Instead of predicting a scalar directly, the model produces per-time logits and a posterior distribution over response-relevant event times; scalar RT is then read out as the posterior expectation. The key distinction from temporal-readout regression is that the distributionally supervised objectives constrain the event-time representation itself, while the RT-only control uses the same posterior-mean readout with scalar supervision.

We evaluate two ways of converting scalar RT labels into event-time supervision. The first constructs an explicit soft target distribution over time, asking the model to match a smoothed target centered at the observed RT. The second treats the observed RT statistically, as a noisy observation of an unobserved response-relevant event time, and optimizes a latent event-time likelihood. These two families correspond to soft-target objectives and likelihood-based objectives, respectively.

The core event-time experiments are organized as a primary controlled comparison and a secondary architecture-robustness analysis. The regression controls above quantify

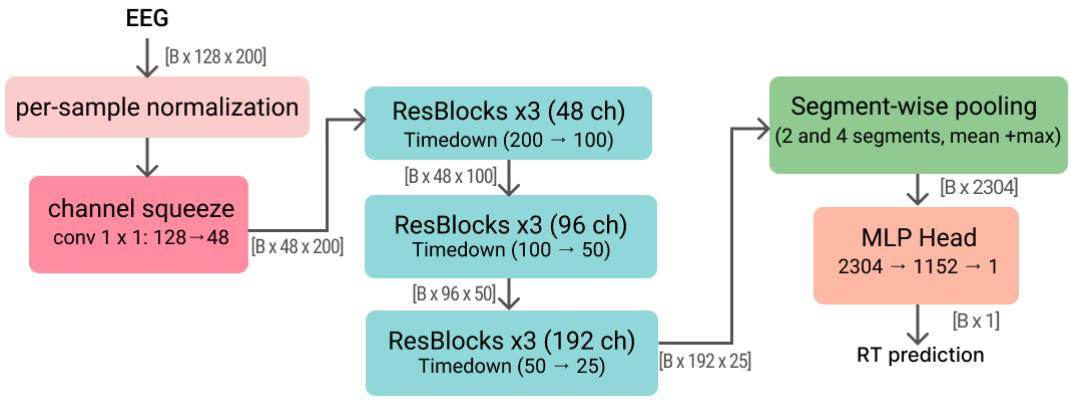

Figure 2: MSP-CNN direct RT regression baseline architecture. A lightweight 1-D temporal ConvNet maps an EEG trial window to a scalar reaction-time prediction. A learned 1×1 channel-squeeze layer (128→48) is followed by residual depthwise-separable stages with anti-aliased temporal downsampling, segment-wise pooling (2 and 4 segments; mean+max), and an MLP head.

Table 3: Direct-regression and external backbone baselines.

Model	**Role**	**Valid nRMSE ↓**	**Holdout nRMSE ↓**
Task-specific regression baselines			
MSP-CNN	segment-pooling scalar	0.9006 ± 0.0051	0.8998 ± 0.0080
ETR-CNN	temporal readout	0.9008 ± 0.0060	0.8977 ± 0.0068
ETR-CNN large	capacity ablation	$\mathbf{0.8972 \pm 0.0040}$	$\mathbf{0.8928 \pm 0.0042}$
External regression baselines			
TIDNet	thinker-invariant CNN	0.9235 ± 0.0024	$\mathbf{0.9192 \pm 0.0027}$
EEGConformer	conv-transformer	$\mathbf{0.9188 \pm 0.0026}$	0.9287 ± 0.0057
EEGNet	compact CNN	0.9350 ± 0.0054	0.9335 ± 0.0028
LaBraM	foundation-style	0.9304 ± 0.0061	0.9327 ± 0.0086
Deep4Net	deep ConvNet	0.9269 ± 0.0045	0.9260 ± 0.0044
ShallowFBCSPNet	shallow FBCSP CNN	0.9324 ± 0.0016	0.9343 ± 0.0024
ATCNet	conv/attention/TCN	0.9686 ± 0.0183	0.9666 ± 0.0151
EEGPT	foundation-style	0.9616 ± 0.0201	0.9584 ± 0.0185
Medformer	larger transformer	0.9623 ± 0.0046	0.9585 ± 0.0051

what scalar supervision achieves with task-specific temporal readouts and external EEG backbones.

We use the event-time segmentation U-Net (ETS-U-Net) as the primary backbone for the objective comparisons because its dense temporal segmentation design naturally matches the event-time output space: it maps each EEG window to per-time logits while combining local and broader temporal context through an encoder–decoder path with skip connections. Fixing this backbone allows us to isolate the effect of output supervision under a strong segmentation model. Additional dense temporal backbones then test whether the resulting supervision effects persist beyond the U-Net inductive

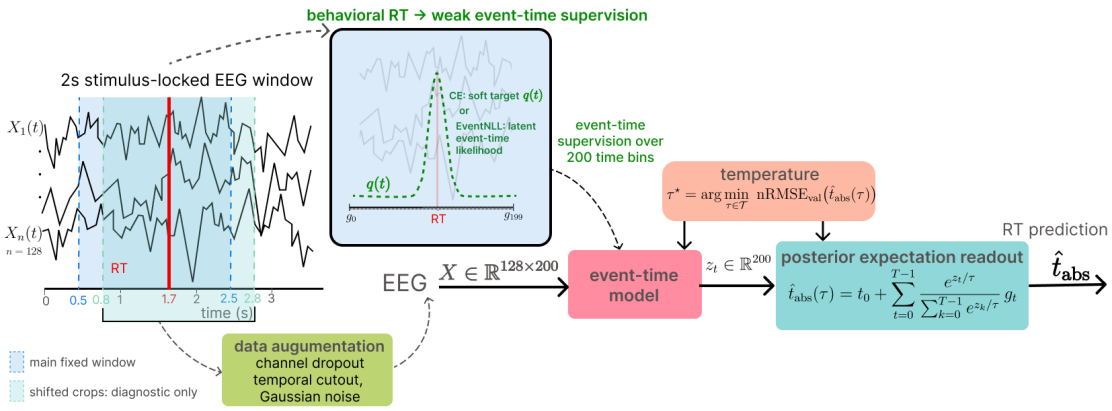

Figure 3: Event-time posterior formulation for EEG reaction-time prediction. A fixed 0.5–2.5 s stimulus-locked EEG window is mapped to time-bin logits, and behavioral RT provides weak event-time supervision either through a soft target distribution or a latent EventNLL likelihood. Scalar RT is read out as the posterior expectation after selecting a readout temperature on R9–R10. Shifted crops are used only for diagnostic and shift-jitter intervention analyses. The schematic shows the categorical posterior readout; the hazard EventNLL variant uses a hazard-derived probability mass function with the same expectation readout.

bias, as described in Section 5.5.

5.1 Primary Event-Time Segmentation Architecture

ETS-U-Net maps an EEG trial window $X \in \mathbb{R}^{C\times T}$ to per-time logits $z_{0:T-1}$ (Figure 4). It follows a 1-D U-Net style encoder-decoder with skip connections that preserve fine temporal resolution (Ronneberger et al., 2015). The encoder progressively downsamples the temporal axis to build multi-scale context, while the decoder upsamples and fuses encoder features to recover per-time outputs at the input temporal resolution (downsampling and upsampling are used only to form multi-scale features). Each scale is refined with lightweight residual blocks (ResNet-style skips) (He et al., 2016) built from depthwise-separable 1-D convolutions for computational efficiency (Howard et al., 2017; Chollet, 2017). To mitigate aliasing under temporal decimation, downsampling applies a fixed depthwise smoothing filter followed by strided averaging ("blur-pool" style) (Zhang, 2019). A dilated bottleneck further enlarges the receptive field without additional downsampling (Yu and Koltun, 2016).

5.2 Soft-Target Distribution Matching

The explicit-target formulation begins by placing each observed RT y on the model's event-time grid. Let $\Delta t = 10$ ms denote the grid spacing, and let $g_t = t\Delta t$, $t = 0,\ldots,T-1$, denote the crop-relative time assigned to bin t. We convert the observed RT from stimulus-relative to crop-relative coordinates, $y_{\mathrm{rel}} = y - t_0$, where t_0 is the crop start, and construct

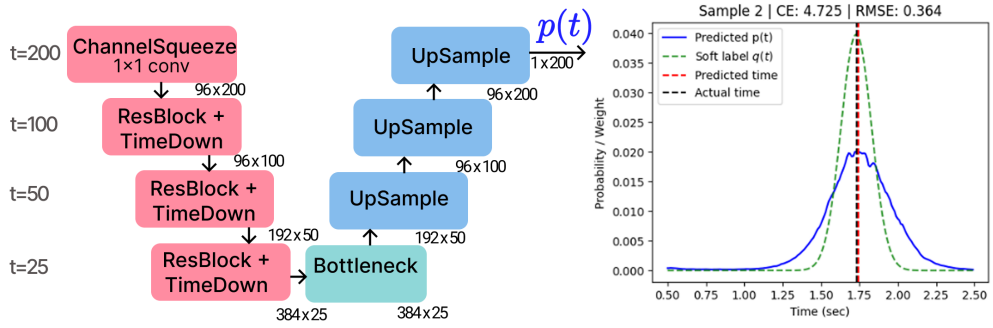

Figure 4: ETS-U-Net for RT segmentation. A lightweight 1-D U-Net maps an EEG trial window to per-time logits and an event-time distribution $p(t)$. The encoder applies a 1×1 channel-squeeze layer followed by residual blocks with temporal downsampling (`TimeDown`); a bottleneck aggregates context, and the decoder upsamples with skip connections to recover time-resolved outputs. The right panel shows an example prediction $p(t)$ and Gaussian soft target $q(t)$, together with the resulting predicted vs. observed response time.

a Gaussian soft target with bandwidth σ:

$$\tilde{q}_t(y) = e^{-\frac{(g_t - y_{\text{rel}})^2}{2\sigma^2}}, \qquad q_t(y) = \frac{\tilde{q}_t(y)}{\sum_{k=0}^{T-1} \tilde{q}_k(y)}, \qquad \sum_{t=0}^{T-1} q_t(y) = 1. \tag{1}$$

The resulting Gaussian target $q_y = (q_0(y), \ldots, q_{T-1}(y))$ provides local temporal smoothing around the observed latency; its bandwidth σ is a supervision parameter rather than an estimate of RT uncertainty.

Temporal logits $z_{\theta,t}(X)$ define the training posterior

$$p_{\theta,t}(X) = \frac{e^{z_{\theta,t}(X)/\tau_0}}{\sum_{k=0}^{T-1} e^{z_{\theta,k}(X)/\tau_0}}, \tag{2}$$

where τ_0 is fixed during training.

Distribution-matching comparison. We first ask how the predicted posterior should be matched to the constructed target. In general, this objective can be written as

$$\mathscr{L}_D(X,y) = D(q_y, p_\theta(\cdot \mid X)), \tag{3}$$

where D determines how disagreement between the two distributions is penalized. We use cross-entropy (CE) as the primary soft-target objective and the one-dimensional Wasserstein distance W_1 as a geometry-aware control:

$$D = \text{CE} \quad \Longrightarrow \quad \mathscr{L}_{\text{CE}}(X,y) = -\sum_{t=0}^{T-1} q_t(y) \log p_{\theta,t}(X). \tag{4}$$

$$D = W_1 \quad \Longrightarrow \quad \mathscr{L}_{W_1}(X,y) = \Delta t \sum_{t=0}^{T-1} \left| \text{CDF}(q_y)_t - \text{CDF}(p_\theta(\cdot \mid X))_t \right|. \tag{5}$$

Cross-entropy treats q_y as a soft categorical target, with temporal proximity represented through the Gaussian smoothing of the target itself. By contrast, W_1 explicitly uses the ordering of the temporal grid and measures how far probability mass must move to match the target.

Posterior-mean control. To test whether distributional supervision adds beyond the posterior-mean readout, we use an RT-only control:

$$\hat{t}_{\mathrm{rel}}(X) = \sum_{t=0}^{T-1} g_t \, p_{\theta,t}(X), \qquad \mathscr{L}_{\mathrm{RT}} = \mathrm{RMSE}(\hat{t}_{\mathrm{rel}}(X), y_{\mathrm{rel}}). \tag{6}$$

CE and W_1 supervise the full event-time distribution, whereas the RT-only control supervises only the scalar prediction produced by the posterior-mean readout.

5.3 Likelihood-Based Event-Time Objectives

As a probabilistic alternative to soft-target supervision, the likelihood-based formulation uses the same event-time posterior but does not construct an explicit target distribution q_y. Instead, it treats the observed RT y as a noisy observation of a latent response-relevant event at stimulus-relative time $t_0 + g_t$. Given an observation kernel K, the marginal likelihood is

$$p_\theta(y \mid X) = \sum_{t=0}^{T-1} p_{\theta,t}(X) \, K(y \mid t_0 + g_t), \qquad \mathscr{L}_{\mathrm{EventNLL}}(X, y) = -\log p_\theta(y \mid X). \tag{7}$$

Here, $p_{\theta,t}(X)$ represents the posterior over latent event times, whereas K describes how a latent event time generates the observed behavioral RT. This separation is also consistent with measurement-oriented views of learned representations as proxy measurements of latent variables (Yao et al., 2025).

Observation-kernel comparison. We first vary the observation kernel while keeping the posterior parameterization and marginal-likelihood objective fixed. The default model uses a Gaussian kernel,

$$K_{\mathrm{G}}(y \mid t) = \mathscr{N}(y; t, \sigma_y^2). \tag{8}$$

We also evaluate a two-scale Gaussian mixture,

$$K_{\mathrm{mix}}(y \mid t) = (1-\alpha)\,\mathscr{N}(y; t, \sigma_{\mathrm{narrow}}^2) + \alpha\,\mathscr{N}(y; t, \sigma_{\mathrm{wide}}^2), \tag{9}$$

which keeps most probability mass near the latent event time while allowing a wider component for occasional noisy RT observations. The observation-kernel parameters are reported in Table 11.

Posterior-parameterization comparison. Motivated by discrete-time survival modeling, we also evaluate a hazard-based construction of the event-time posterior (Gensheimer and Narasimhan, 2019). Whereas a categorical softmax assigns probability to all time bins through a single global normalization, $h_{\theta,t}(X)$ represents the conditional probability that the event occurs in bin t, given that it has not occurred earlier. The corresponding event-time PMF combines this current-bin probability with the probability of reaching bin t without an earlier event:

$$h_{\theta,t}(X) = \mathrm{sigmoid}\big(z_{\theta,t}(X)/\tau_0\big), \qquad p^{\mathrm{haz}}_{\theta,t}(X) \propto h_{\theta,t}(X) \prod_{k<t} \big(1 - h_{\theta,k}(X)\big), \tag{10}$$

The normalized $p^{\mathrm{haz}}_{\theta,t}(X)$ replaces $p_{\theta,t}(X)$ in Eq. 7.

5.4 Formulation Comparison and Robustness

We next compare the event-time objectives under a fixed ETS-U-Net backbone. The central question is whether distributional event-time supervision improves RT prediction beyond posterior-mean readout alone, and whether this benefit depends on a particular target construction, observation kernel, or posterior parameterization. Table 4 reports all variants under the same release-separated protocol. Holdout τ-nRMSE uses the readout temperature selected on R9–R10 and applied unchanged to the holdout split. After this common readout procedure, CE and the likelihood-based variants achieve the lowest held-out errors, with holdout τ-nRMSE between 0.8745 and 0.8778. These objectives outperform the strongest scalar regression baseline, ETR-CNN large (0.8928 ± 0.0042; Table 3), whereas the soft-argmax RT-loss and Wasserstein controls remain closer to the scalar-regression baselines.

ETR-CNN large provides a stringent scalar reference because it already includes a task-specific temporal expectation readout. Event-time supervision reduces held-out error beyond this scalar temporal inductive bias.

Table 4: Event-time objective variants with a fixed ETS-U-Net backbone.

Objective	Supervision signal	Valid nRMSE	Holdout nRMSE	Holdout τ-nRMSE
Soft-target objectives				
CE	Gaussian soft event-time label	0.8763 ± 0.0044	$\mathbf{0.8774 \pm 0.0044}$	0.8753 ± 0.0039
Wasserstein	CDF match to Gaussian soft event-time label	0.8997 ± 0.0035	0.8995 ± 0.0078	0.8896 ± 0.0033
Likelihood-based objectives				
EventNLL	latent event time + Gaussian RT likelihood	0.8769 ± 0.0030	0.8805 ± 0.0021	0.8772 ± 0.0018
Mixture EventNLL	Two-component core-tail RT likelihood	$\mathbf{0.8744 \pm 0.0018}$	0.8785 ± 0.0047	$\mathbf{0.8745 \pm 0.0053}$
Hazard EventNLL	hazard posterior + Gaussian RT likelihood	0.8755 ± 0.0027	0.8776 ± 0.0031	0.8778 ± 0.0041
Readout-only control				
Soft-argmax RT loss	posterior-mean RT error only; no distribution target	0.8944 ± 0.0048	0.8943 ± 0.0025	0.8917 ± 0.0046

Objective ablations. The ablations separate expectation-style temporal readout from distributional supervision. The soft-argmax RT-loss control directly optimizes posterior-mean RT error but removes distribution matching; it remains close to the strongest scalar-regression baseline and trails the CE/EventNLL group by roughly 0.014–0.017 holdout τ-nRMSE. CE and the likelihood-family variants form a tight cluster among the best-performing objectives: mixture EventNLL has the best mean temperature-tuned scalar readout, but CE, Gaussian EventNLL, mixture EventNLL, and hazard EventNLL are close relative to seed variability. The hazard result shows that the likelihood-family conclusion is not tied to a categorical softmax posterior, while the Wasserstein control shows that an alternative distributional distance alone does not match the CE/EventNLL scalar readouts. Within the fixed ETS-U-Net backbone, these results isolate the gain to the supervision objective rather than to the temporal expectation readout alone. The next subsection tests whether the same supervision pattern persists when the dense temporal backbone is changed.

Practical effect size. To express the main accuracy gain on the RT scale, we compared the strongest scalar baseline, ETR-CNN large, with the strongest event-time objectives using matched seeds and subject-level bootstrap intervals on the holdout split. Mixture EventNLL reduced τ-nRMSE from 0.8928 ± 0.0042 to 0.8745 ± 0.0053, an absolute reduction of 0.0184, or about 2.1% relative improvement. On the RT scale, this corresponds to a 6.24 ms reduction in RMSE and an 8.94 ms reduction in MAE, with subject-bootstrap 95% confidence intervals of $[4.50, 7.97]$ ms and $[7.43, 10.49]$ ms, respectively. The improvement was positive in all five matched-seed comparisons, and both bootstrap intervals exclude zero. Together, the matched-seed and subject-bootstrap results establish a consistent predictive advantage over the scalar temporal-readout ref-

erence. The event-time formulation also provides a common posterior object for the diagnostics and shifted-crop analyses in Section 6.

5.5 Architecture Robustness of the Supervision Effect

To assess whether the supervision effect is robust to backbone choice, we repeat the RT-only soft-argmax, CE, and mixture EventNLL comparison using three custom temporal backbones. These controls preserve full-resolution event-time output, closely match ETS-U-Net in capacity, and span three distinct temporal inductive biases.

ETS-TCN uses residual dilated temporal convolutions (Bai et al., 2018; Yu and Koltun, 2016). ETS-InceptionPyramid applies parallel depthwise temporal filters at multiple receptive-field scales and fuses them through residual blocks, following Inception-style design principles (Szegedy et al., 2015; Santamaría-Vázquez et al., 2020). ETS-AttnSeg uses Conformer-style blocks that combine global temporal self-attention with local gated depthwise convolution (Vaswani et al., 2017; Gulati et al., 2020). The three controls contain 3.05–3.25 million trainable parameters, compared with 3.10 million for ETS-U-Net. Table 5 reports the resulting scalar RT comparison.

Table 5: Scalar RT accuracy across event-time backbones and objectives.

Objective	**Valid nRMSE**	**Holdout nRMSE**	**Holdout τ-nRMSE**
ETS-U-Net			
RT-only soft-argmax	0.8944 ± 0.0048	0.8943 ± 0.0025	0.8917 ± 0.0046
CE	0.8763 ± 0.0044	0.8774 ± 0.0044	0.8753 ± 0.0039
Mixture EventNLL	0.8744 ± 0.0018	0.8785 ± 0.0047	$\mathbf{0.8745 \pm 0.0053}$
ETS-TCN			
RT-only soft-argmax	0.8865 ± 0.0029	0.8853 ± 0.0044	0.8842 ± 0.0045
CE	0.8717 ± 0.0025	0.8751 ± 0.0085	0.8730 ± 0.0038
Mixture EventNLL	0.8718 ± 0.0046	0.8729 ± 0.0020	$\mathbf{0.8722 \pm 0.0021}$
ETS-InceptionPyramid			
RT-only soft-argmax	0.8970 ± 0.0043	0.8894 ± 0.0054	0.8886 ± 0.0044
CE	0.8710 ± 0.0037	0.8717 ± 0.0017	$\mathbf{0.8717 \pm 0.0036}$
Mixture EventNLL	0.8703 ± 0.0018	0.8780 ± 0.0026	0.8746 ± 0.0028
ETS-AttnSeg			
RT-only soft-argmax	0.9114 ± 0.0350	0.9105 ± 0.0294	0.9052 ± 0.0300
CE	0.8742 ± 0.0036	0.8811 ± 0.0097	0.8780 ± 0.0082
Mixture EventNLL	0.8719 ± 0.0082	0.8823 ± 0.0079	$\mathbf{0.8752 \pm 0.0050}$

Across all four backbones, both CE and mixture EventNLL yield lower mean holdout τ-nRMSE than the RT-only soft-argmax control. Although the leading distributional objective varies between CE and mixture EventNLL, this supervision advantage replicates across U-Net, dilated-convolution, multi-scale convolution, and attention-convolution designs. Complementary shifted-crop and posterior-geometry analyses are reported in Appendix B.

6 Posterior Readout and Diagnostics

6.1 Scalar Readout and Temperature Tuning

Posterior-mean scalar readout. Event-time models output a posterior over response-relevant time bins, whereas the benchmark evaluates scalar RT. We therefore use the posterior expectation as the scalar readout:

$$\hat{t}_{\mathrm{rel}}(\tau) = \sum_{t=0}^{T-1} p_t(\tau)\, g_t, \qquad \hat{t}_{\mathrm{abs}}(\tau) = t_0 + \hat{t}_{\mathrm{rel}}(\tau), \tag{11}$$

where $g_t = t\,\Delta t$ denotes the discrete time grid. The relative prediction is converted to absolute RT by adding the fixed window start t_0, keeping event-time models comparable to scalar regression baselines under nRMSE.

Readout-temperature tuning. Because different objectives can produce posteriors with different sharpness, we select a readout temperature on the R9–R10 development split to minimize posterior-mean nRMSE and apply it unchanged to the holdout set. The resulting holdout τ-nRMSE is reported in Table 4. This is a scalar-readout choice, not probabilistic calibration; likelihood, coverage, and posterior concentration are evaluated separately below.

6.2 Posterior Geometry Diagnostics

Once the scalar readout has been fixed, the event-time posterior enables evaluation beyond point prediction because similar nRMSE values can correspond to substantially different posterior geometry. We therefore evaluate Width80 (central 80% interval width), Mass $\pm$150 ms (probability assigned within 150 ms of observed RT), and mode–mean disagreement (absolute distance between posterior mode and mean), measuring temporal concentration, target alignment, and agreement with the expectation used for scalar prediction. Width80 and mode–mean disagreement are summarized by their within-seed holdout medians, and target-aligned mass by its mean.

In Figure 5, CE and Wasserstein place broad probability mass across the response window, whereas EventNLL-family objectives produce sharper bands that more closely track the observed RT curve.

Figure 6 summarizes these posterior-geometry differences, while Table 6 reports scalar accuracy, shared-kernel RT NLL, width, mass, and coverage. Shared-kernel RT NLL provides a common likelihood-based score across objectives with different training scales; its parameters are reported in Table 11. Coverage80 is the fraction of observed RTs inside the central 80% posterior interval, and Coverage MAE averages the absolute nominal–empirical coverage difference over central interval levels of 50%, 60%, 70%, 80%, and 90%. Together, these metrics separate scalar accuracy, distributional scoring, temporal concentration, target alignment, and interval behavior.

These diagnostics show that losses with similar scalar error can induce different posterior semantics. CE and mixture EventNLL provide the strongest scalar readouts. EventNLL-family objectives are sharper and more target-concentrated, but their latent-posterior intervals have low empirical coverage. The soft-argmax RT-loss control

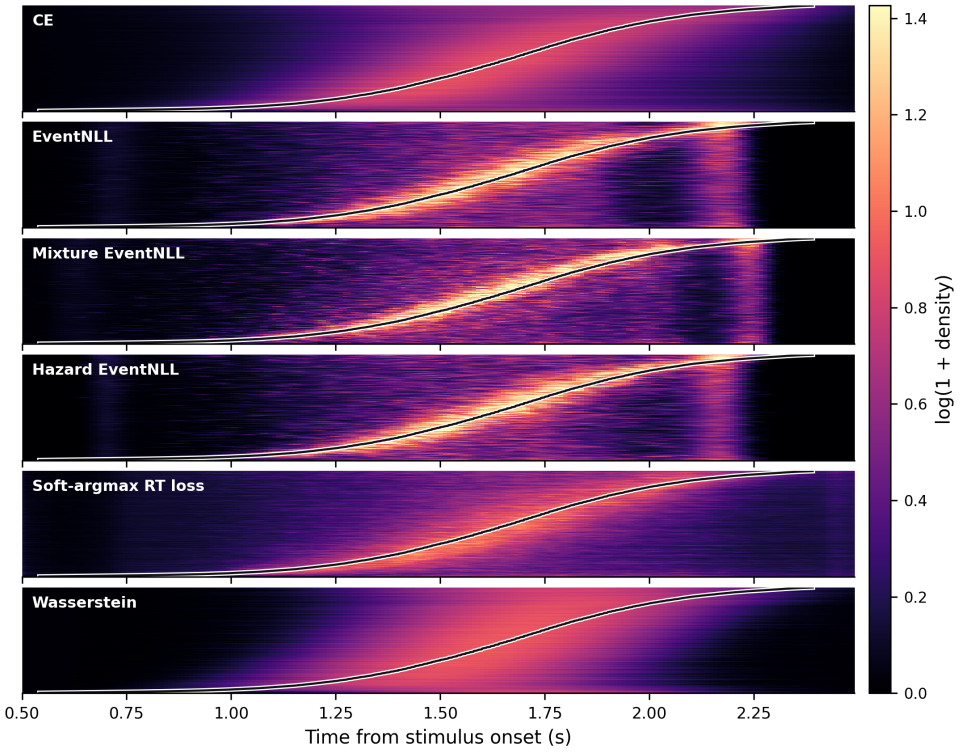

Figure 5: Trial-sorted event-time posterior maps on the holdout split for matched segmentation losses. Rows are quantile bins of support-filtered holdout trials sorted by observed RT; the x-axis is time from stimulus onset; color shows log-transformed posterior density averaged within each bin. The overlaid curve marks the mean observed RT in each bin.

Table 6: Holdout posterior-geometry metrics across event-time segmentation losses.

Objective	nRMSE ↓	Shared-kernel RT NLL ↓	Width80 ms ↓	Mass ±150 ms ↑	Coverage80	Coverage MAE ↓
CE	0.875 ± 0.004	0.08 ± 0.01	766 ± 36	0.334 ± 0.008	0.883 ± 0.013	0.101 ± 0.014
EventNLL	0.877 ± 0.002	-0.06 ± 0.01	502 ± 15	0.490 ± 0.008	0.500 ± 0.025	0.273 ± 0.021
Mixture EventNLL	**0.874 ± 0.005**	**-0.08 ± 0.01**	528 ± 22	0.471 ± 0.008	0.573 ± 0.020	0.207 ± 0.017
Hazard EventNLL	0.878 ± 0.004	-0.06 ± 0.01	**445 ± 14**	**0.502 ± 0.005**	0.468 ± 0.007	0.302 ± 0.006
Soft-argmax RT loss	0.892 ± 0.005	0.11 ± 0.03	844 ± 96	0.357 ± 0.015	0.843 ± 0.029	**0.040 ± 0.024**
Wasserstein	0.890 ± 0.003	0.20 ± 0.04	632 ± 101	0.334 ± 0.028	**0.774 ± 0.061**	0.059 ± 0.021

combines the lowest Coverage MAE with the broadest posterior, but it is weaker as a scalar predictor and lacks distributional event-time supervision. Wasserstein is closest to nominal 80% coverage, but it has weaker scalar accuracy and lower near-target mass. Thus, posterior geometry reveals trade-offs among scalar accuracy, target concentration, distributional scoring, and interval behavior that scalar nRMSE alone would hide.

Accordingly, CE and mixture EventNLL are strongest for scalar RT prediction in this study, whereas EventNLL-family objectives provide sharper, more target-concentrated latent posteriors. Wasserstein and the soft-argmax control yield latent-posterior intervals closer to nominal coverage, but these intervals are not calibrated predictive RT uncertainty; the following analysis evaluates that distinction through the observation-noise model.

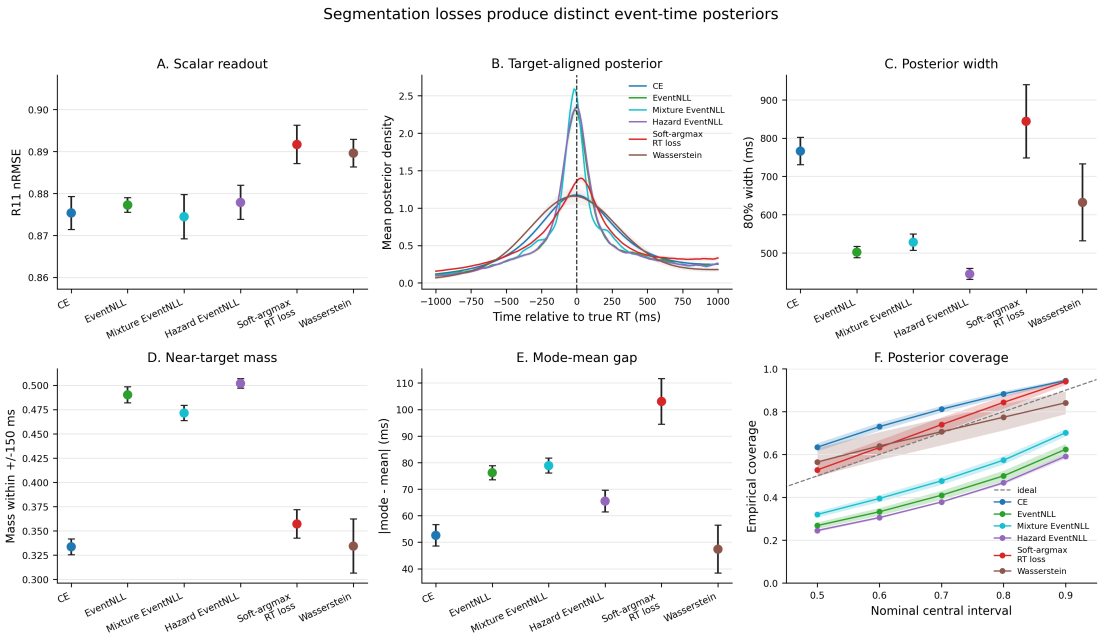

Figure 6: Posterior geometry differs across event-time losses even when scalar nRMSE is similar. Panel A reports holdout scalar readout after selecting the objective-specific readout temperature on R9–R10 to minimize nRMSE. Panel B aligns predicted posteriors to the observed RT. Panels C–E summarize posterior width, near-target mass, and mode-mean disagreement. Panel F compares empirical coverage of central posterior intervals; the diagonal is ideal interval coverage.

Latent Event-Time Posterior and Predictive RT Calibration. The coverage columns in Table 6 report how often observed RT falls inside central intervals of the latent event-time posterior, $p(t_{\text{event}} \mid X)$. For EventNLL-family models, the predictive distribution over observed RT is instead obtained by combining this posterior with an observation kernel:

$$p(y_{\text{RT}} \mid X) = \sum_t p(t_{\text{event}} = t \mid X) K(y_{\text{RT}} \mid t).$$

This separates uncertainty over the latent response-relevant event time from predictive uncertainty over observed RT. Appendix Table 8 evaluates predictive calibration while keeping the trained EEG model, event-time posterior, and posterior-mean prediction fixed. A single multiplicative scale for the observation kernel is selected on R9–R10 by Coverage MAE over central intervals from 50% to 90% and applied unchanged to holdout. This reduces predictive RT Coverage MAE from 0.040–0.059 to 0.005–0.008, with holdout Coverage80 near nominal coverage (0.790–0.795).

The latent posterior therefore describes event-time concentration, whereas the observation model supplies calibrated predictive RT intervals without altering point predictions. We next test whether posterior predictions move in a crop-relative way when the temporal frame is shifted.

Table 7: Shifted-crop diagnostic and shift-jitter intervention on the 5 s holdout dataset. Values are seed means; full mean ± standard deviation values are reported in Appendix Table 9. Arrows indicate the preferred direction.

Objective	Holdout τ-nRMSE ↓		Shifted rel. nRMSE ↓		Sensitivity ↑		Direction ↑	
	Fixed	Jitter	Fixed	Jitter	Fixed	Jitter	Fixed	Jitter
CE	0.8753	0.8749	0.8679	**0.8569**	0.5810	0.5831	0.7780	0.7924
Mixture EventNLL	**0.8745**	**0.8734**	**0.8663**	0.8576	0.6021	0.6249	0.7730	0.7925
EventNLL	0.8772	0.8771	0.8685	0.8593	0.5842	0.6174	0.7739	0.7937
Hazard EventNLL	0.8778	0.8806	0.8692	0.8609	0.5618	0.5891	0.7694	0.7925
Soft-argmax RT loss	0.8917	0.8836	0.8857	0.8613	0.5381	0.5775	0.7592	0.7951
Wasserstein	0.8896	0.8922	0.8932	0.8774	**0.6684**	**0.6852**	**0.7830**	**0.8026**

6.3 Shifted-Crop Shortcut-vs-Localization Diagnostic

Posterior geometry characterizes probability mass within the standard stimulus-locked window, but it does not test whether predictions behave as crop-relative timing localizers when the temporal frame changes. A fixed-window model may instead rely on trial difficulty, subject-level response tendency, or stimulus-locked timing structure.

We therefore use shifted-crop inference as a shortcut-vs-localization diagnostic on the holdout split. For each trial, the trained model is evaluated on 2 s crops from the same 5 s EEG segment, starting at $s \in \{0.2, 0.3, \ldots, 0.8\}$ s after stimulus onset. The canonical window starts at $s = 0.5$ s, and the target for crop s is $\mathrm{RT} - s$. A crop-relative localizer should adjust its prediction when the crop start changes, whereas a model tied to stimulus-locked scalar timing should be less sensitive.

Table 7 summarizes this diagnostic and the matched shift-jitter intervention. Shifted relative nRMSE measures crop-relative accuracy over pooled shifted-crop examples for which the response remains inside the evaluated 2 s window. Shift sensitivity compares the magnitude of the prediction change from the canonical crop with the imposed crop shift; 1 indicates ideal crop-relative responsiveness and 0 indicates crop-invariant prediction. Direction agreement is the fraction of examples for which prediction changes opposite to the crop shift. Sensitivity and direction are computed on the common trial subset whose responses remain inside every evaluated crop, ensuring comparisons over the same trials and crop starts.

We then trained matched shift-jitter variants by sampling the 2 s training window from the same crop-start range and expressing the target relative to the sampled window. For CE, mixture EventNLL, and Gaussian EventNLL, the intervention preserves standard fixed-window holdout accuracy. Shifted relative nRMSE improves for every evaluated objective. Mean direction agreement and sensitivity also increase for every objective, showing that predictions move more consistently in the expected localizer direction. Sensitivity remains below 1, quantifying the residual equivariance gap.

Across objectives, Wasserstein shows the strongest crop-relative sensitivity and direction agreement but weaker scalar and shifted-crop accuracy, illustrating a trade-off between crop-relative movement and point prediction.

7 Discussion

7.1 Latency Targets as Output Representations

For latency-labeled EEG, the supervised target representation is itself a modeling choice rather than a neutral readout convention. The label is a behavioral latency within a post-stimulus interval, and models benefit when that temporal structure is explicit in the output space and in the loss. CE and the EventNLL-family objectives form the strongest scalar-readout group, whereas the soft-argmax RT-loss control is weaker despite using the same posterior-mean readout. This supports the interpretation that the gain comes from distributional event-time supervision, not merely from replacing a scalar head with a differentiable expectation.

Objective choice therefore shapes not only scalar accuracy but also the structure of the learned posterior. CE-style soft targets provide strong scalar readouts and broad posterior support. EventNLL reaches a similar scalar-error regime through a latent-variable likelihood, marginalizing an unobserved response-relevant event time under an observation model for RT. Wasserstein behaves differently: it is more localizer-like by shifted-crop sensitivity and direction, but weaker as a point predictor. Thus, the choice of objective controls a trade-off among scalar accuracy, posterior concentration, interval behavior, and crop-relative movement.

7.2 Posterior Diagnostics Beyond nRMSE

The event-time posterior contains more information than its expectation. Posterior width, target-aligned mass, mode-mean disagreement, shared-kernel RT NLL, and interval coverage expose temporal output behavior that scalar nRMSE discards. This is important because similar point errors can correspond to different posterior semantics: broad conservative distributions, sharper target-concentrated distributions, or better interval coverage with weaker scalar prediction.

Temperature tuning serves a specific role in this protocol: it standardizes posterior-mean predictions for scalar nRMSE comparison. Distributional diagnostics remain separate: likelihood, coverage, posterior concentration, and temporal-perturbation behavior answer different questions about the learned event-time posterior. The observation-noise calibration analysis further separates three goals that are easy to conflate: scalar RT accuracy, latent event-time concentration, and calibrated predictive uncertainty over observed behavioral RT.

7.3 Temporal Perturbation and Shortcut Behavior

The shifted-crop diagnostic asks whether posterior predictions behave as temporal localizers when the crop frame changes. If models rely on fixed-window, stimulus-locked shortcuts, crop shifts should expose this by separating scalar accuracy from crop-relative movement. Shift-jitter training addresses this behavior by varying the crop start during training and expressing the target relative to the sampled crop. Across every evaluated objective, shift-jitter improves shifted-crop accuracy and expected-direction movement while preserving canonical-window accuracy for the primary event-time models. Mean sensitivity and direction increase across all objectives, demonstrating that

crop-relative behavior can be strengthened through target-aligned temporal augmentation. Sensitivity below 1 shows that localization remains partial and provides a quantitative target for fully equivariant models.

7.4 Scope, Limitations, and Generality

The external architecture baselines place the formulation result in context: under the same preprocessing and normalization protocol, compact task-specific temporal models outperform the evaluated standardized EEG backbones and foundation-style architectures trained from scratch. Across matched dense temporal backbones, distributionally supervised objectives also outperform RT-only posterior-mean training. Thus, within this benchmark, these complementary controls identify formulation and objective design as a consistent source of predictive improvement beyond architecture capacity alone.

The empirical scope is the HBN-EEG contrast change detection task under a release-separated protocol. Broader generality remains to be tested across tasks, acquisition settings, montages, and latency targets. The shifted-crop analysis provides response-timing localization evidence and also shows that full crop-relative localization remains an open target for objectives, architectures, or augmentations designed for temporal equivariance.

8 Conclusion

Reaction time is usually evaluated as a scalar, but it also specifies when a behavioral response occurs. We therefore treated RT as weak event-time supervision and learned a posterior over response-relevant timing from single-trial EEG. Under the subject-disjoint, release-separated HBN-EEG protocol, distributional event-time supervision consistently improved held-out RT prediction over direct scalar regression and an RT-only posterior-mean control. Five-seed comparisons and replication across four dense temporal backbones show that this advantage is attributable to the supervised output formulation rather than to model scale, expectation readout, or a particular architecture.

The posterior representation also changed what could be evaluated. CE and the EventNLL-family objectives achieved the strongest scalar readouts but induced different posterior geometries; observation-noise calibration separated latent event-time concentration from predictive uncertainty over observed RT; and shifted-crop analyses revealed partial crop-relative localization that scalar nRMSE alone cannot diagnose. Shift-jitter strengthened shifted-crop robustness and expected-direction movement, while the remaining sensitivity gap defines a clear target for temporally equivariant modeling.

The broader implication is that latency labels need not be reduced to scalar endpoints. They can serve as structured supervision that preserves temporal meaning while remaining compatible with standard behavioral metrics. Extending this view to other tasks and latency-defined EEG phenomena could support models that predict behavior accurately, expose the temporal organization of their outputs, and make their reliance on temporal evidence directly testable.

Table 8: Post-hoc RT observation-noise calibration for EventNLL-family posteriors. Coverage MAE is averaged over central interval levels 50%, 60%, 70%, 80%, and 90%.

Distribution evaluated	**Scale** c	**Coverage MAE** $\downarrow$	**Coverage80**	**Width80 ms**	**Predictive NLL** $\downarrow$
EventNLL					
Latent event-time posterior	–	0.273 ± 0.021	0.500 ± 0.025	502 ± 15	–
Predictive RT, original K	1.00	0.059 ± 0.004	0.850 ± 0.004	699 ± 13	-0.044 ± 0.010
Predictive RT, calibrated K_c	0.70 ± 0.00	0.006 ± 0.003	0.793 ± 0.004	632 ± 14	-0.054 ± 0.016
Mixture EventNLL					
Latent event-time posterior	–	0.207 ± 0.017	0.573 ± 0.020	528 ± 22	–
Predictive RT, original K	1.00	0.040 ± 0.010	0.835 ± 0.009	671 ± 17	-0.098 ± 0.012
Predictive RT, calibrated K_c	0.73 ± 0.04	0.005 ± 0.004	0.795 ± 0.003	622 ± 12	-0.111 ± 0.013
Hazard EventNLL					
Latent event-time posterior	–	0.303 ± 0.006	0.467 ± 0.007	440 ± 17	–
Predictive RT, original K	1.00	0.044 ± 0.006	0.834 ± 0.006	639 ± 15	-0.044 ± 0.006
Predictive RT, calibrated K_c	0.78 ± 0.04	0.008 ± 0.002	0.790 ± 0.005	587 ± 17	-0.059 ± 0.005

A Additional Readout, Calibration, and Shifted-Crop Details

This appendix records three protocol details that are secondary to the main results but important for reproducing the posterior diagnostics. First, post-hoc observation-noise calibration separates latent event-time posterior coverage from behavioral-RT predictive coverage. Second, the posterior readout temperature is selected only on the R9–R10 development split and then applied unchanged to the holdout split. Third, the shifted-crop table in the main text reports seed means for readability; the full seed-variability summary is given here.

Observation-noise calibration. Table 8 evaluates a post-hoc RT observation-noise calibration for EventNLL-family models. The trained EEG model, event-time posterior, and posterior-mean scalar readout are fixed. The latent posterior rows evaluate central intervals of $p(t_{\text{event}} \mid X)$ directly, whereas the predictive RT rows convolve the same posterior with either the original observation kernel or a calibrated kernel scale selected on R9–R10 by Coverage MAE across central intervals 50–90%. This calibration changes only the RT observation-noise layer used for probabilistic prediction; it does not change trained weights, posterior-mean RT predictions, or τ-nRMSE.

Readout-temperature selection. Figure 7 shows the development-split nRMSE curves used to select the posterior-readout temperature for each event-time objective. The minima differ across losses, confirming that objectives induce different posterior confidence scales. We therefore treat temperature selection as part of the scalar readout protocol before comparing posterior-mean predictions and posterior geometry on the holdout split.

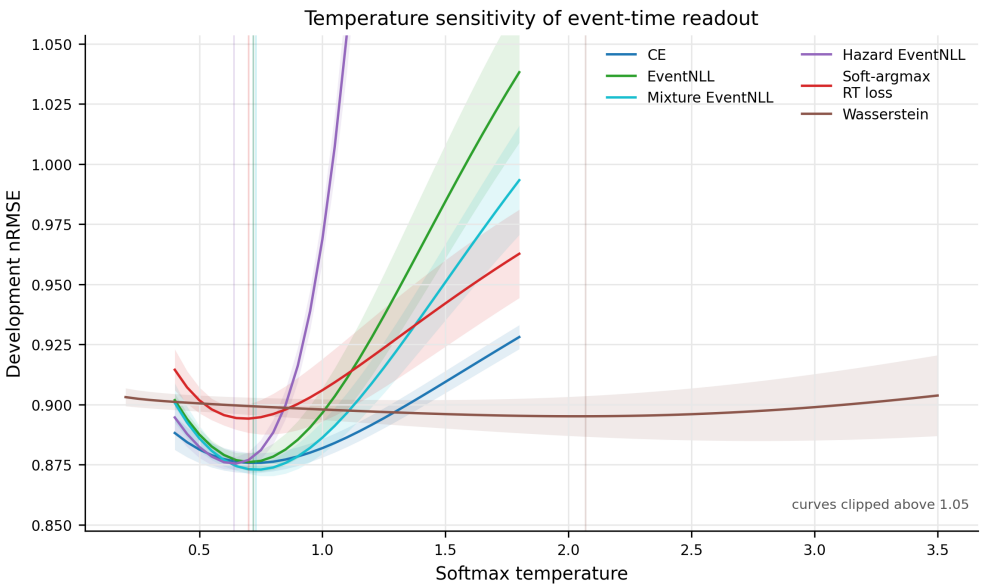

Figure 7: Readout-temperature sensitivity on the R9–R10 development split. Curves show posterior-mean nRMSE as a function of softmax temperature, and vertical lines mark selected temperatures. The selected temperature is then applied unchanged to holdout predictions and posterior diagnostics.

Table 9: Full shifted-crop diagnostic and shift-jitter intervention summary with seed variability. Values are mean ± standard deviation across seeds; the corresponding main-text table reports means only for readability.

Objective	Holdout τ-nRMSE ↓		Shifted rel. nRMSE ↓		Sensitivity ↑		Direction ↑	
	Fixed	Jitter	Fixed	Jitter	Fixed	Jitter	Fixed	Jitter
CE	0.8753 ± 0.0039	0.8749 ± 0.0060	0.8679 ± 0.0024	0.8569 ± 0.0060	0.5810 ± 0.0351	0.5831 ± 0.0406	0.7780 ± 0.0041	0.7924 ± 0.0056
Mixture EventNLL	0.8745 ± 0.0053	0.8734 ± 0.0033	0.8663 ± 0.0037	0.8576 ± 0.0041	0.6021 ± 0.0484	0.6249 ± 0.0313	0.7730 ± 0.0062	0.7925 ± 0.0076
EventNLL	0.8772 ± 0.0018	0.8771 ± 0.0040	0.8685 ± 0.0023	0.8593 ± 0.0028	0.5842 ± 0.0287	0.6174 ± 0.0222	0.7739 ± 0.0059	0.7937 ± 0.0068
Hazard EventNLL	0.8778 ± 0.0041	0.8806 ± 0.0034	0.8692 ± 0.0028	0.8609 ± 0.0045	0.5618 ± 0.0376	0.5891 ± 0.0259	0.7694 ± 0.0090	0.7925 ± 0.0074
Soft-argmax RT loss	0.8917 ± 0.0046	0.8836 ± 0.0035	0.8857 ± 0.0069	0.8613 ± 0.0032	0.5381 ± 0.0586	0.5775 ± 0.0362	0.7592 ± 0.0183	0.7951 ± 0.0052
Wasserstein	0.8896 ± 0.0033	0.8922 ± 0.0066	0.8932 ± 0.0033	0.8774 ± 0.0099	0.6684 ± 0.0711	0.6852 ± 0.0416	0.7830 ± 0.0197	0.8026 ± 0.0118

Shifted-crop seed variability. Table 9 expands the main shifted-crop diagnostic table by reporting mean ± standard deviation across seeds. The seed-level results reproduce the main-text pattern: shift-jitter lowers mean shifted-crop relative nRMSE and increases both mean sensitivity and direction for every evaluated objective. These consistent changes show that shift-jitter moves predictions toward crop-relative temporal localization across loss families. Absolute sensitivity remains below ideal equivariance, quantifying the localization gap that remains to be addressed.

B Architecture-Control Details

Architecture specifications. The architecture-control hyperparameters and parameter counts are summarized in Table 11.

Diagnostic robustness. Table 10 complements the fixed-window accuracy comparison in Table 5 by testing whether shifted-crop behavior and objective-dependent posterior geometry also persist across backbones. The posterior columns retain a common distributional score, target-aligned mass, and interval coverage error.

Table 10: Shifted-crop and posterior-geometry diagnostics across dense temporal backbones. Shifted-crop metrics follow the evaluation convention in Section 6.3; posterior metrics are evaluated on the fixed holdout window after development-set readout-temperature selection. Values are mean ± standard deviation across five seeds.

Objective	Shifted-crop diagnostics			Posterior geometry		
	Rel. nRMSE ↓	Sensitivity ↑	Direction ↑	Shared-kernel RT NLL ↓	Mass ±150 ms ↑	Coverage MAE ↓
ETS-U-Net						
RT-only soft-argmax	0.886 ± 0.007	0.538 ± 0.059	0.759 ± 0.018	0.1070 ± 0.0303	0.357 ± 0.015	0.040 ± 0.024
CE	0.868 ± 0.002	0.581 ± 0.035	0.778 ± 0.004	0.0770 ± 0.0144	0.334 ± 0.008	0.101 ± 0.014
Mixture EventNLL	0.866 ± 0.004	0.602 ± 0.048	0.773 ± 0.006	-0.0820 ± 0.0122	0.471 ± 0.008	0.207 ± 0.017
ETS-TCN						
RT-only soft-argmax	0.881 ± 0.006	0.519 ± 0.042	0.758 ± 0.012	0.0268 ± 0.0114	0.410 ± 0.012	0.080 ± 0.031
CE	0.870 ± 0.008	0.551 ± 0.025	0.774 ± 0.006	0.0710 ± 0.0243	0.336 ± 0.013	0.099 ± 0.016
Mixture EventNLL	0.867 ± 0.005	0.544 ± 0.036	0.761 ± 0.010	-0.0835 ± 0.0039	0.471 ± 0.007	0.187 ± 0.017
ETS-InceptionPyramid						
RT-only soft-argmax	0.882 ± 0.005	0.506 ± 0.035	0.758 ± 0.007	0.1634 ± 0.0195	0.318 ± 0.007	0.103 ± 0.020
CE	0.863 ± 0.004	0.533 ± 0.032	0.781 ± 0.012	0.0605 ± 0.0173	0.340 ± 0.012	0.097 ± 0.017
Mixture EventNLL	0.866 ± 0.002	0.589 ± 0.037	0.774 ± 0.005	-0.0925 ± 0.0049	0.473 ± 0.008	0.185 ± 0.019
ETS-AttnSeg						
RT-only soft-argmax	0.899 ± 0.033	0.479 ± 0.156	0.749 ± 0.028	0.2977 ± 0.1669	0.268 ± 0.048	0.156 ± 0.046
CE	0.868 ± 0.009	0.566 ± 0.057	0.771 ± 0.005	0.0771 ± 0.0276	0.334 ± 0.016	0.103 ± 0.019
Mixture EventNLL	0.867 ± 0.008	0.639 ± 0.034	0.767 ± 0.004	-0.0937 ± 0.0083	0.477 ± 0.007	0.173 ± 0.023

Across every backbone, CE and mixture EventNLL consistently achieve lower mean shifted-crop error than RT-only posterior-mean training and produce more crop-responsive predictions, as reflected in higher sensitivity and direction scores. These improvements replicate across all four temporal architectures. Absolute sensitivity remains below 1 and therefore quantifies partial rather than fully crop-relative localization. Mixture EventNLL shows an equally consistent posterior-geometry pattern: within every backbone, it yields the lowest shared-kernel RT NLL and the greatest target-aligned mass, together with higher coverage error. The replication of both patterns across four distinct dense temporal backbones establishes the architecture robustness of the supervision effect. These experiments serve as controlled robustness checks rather than a ranking of general EEG architectures.

C Reproducibility Details

YAML configuration files at `https://github.com/sneddy/neurosned/tree/main/benchmarks/configs` specify all reported experiments. Training summaries, prediction files, selected temperatures, and model summaries are written to `benchmarks/experiments/`. Table 11 summarizes the shared settings needed to reproduce the controlled comparisons: data splits, input representation, optimization, augmentation, model selection, readout-temperature tuning, and the main architecture

hyperparameters. Settings not listed in the table are objective-specific loss choices described in the main text.

Table 11: Reproducibility summary for the main controlled comparisons.

Group	Setting	Value
Data and splits		
Input window	EEG segment	0.5–2.5 s after stimulus onset; 2 s, 200 samples, 100 Hz, 128 channels.
Target support	Main analyzed set	$0.5 \leq \text{RT} \leq 2.5$ s.
Splits	Release-separated protocol	R1–R8 train, R9–R10 development, R11 final holdout.
Seeds	Repeated runs	2025–2029.
Optimization and selection		
Optimizer	Neural models	Adam, learning rate 10^{-3}, weight decay 0.
Training loop	Epochs and batch size	Maximum 100 epochs; train batch size 128.
Evaluation loader	Batch size	Development and holdout batch size 256.
Checkpoint selection	Model selection	Best development nRMSE; early stopping patience 20.
Holdout use	Final evaluation	R11 is evaluated after checkpoint and readout choices are fixed.
Confidence intervals	Holdout summaries	Subject bootstrap over R11 subjects; 1000 samples for per-run intervals and 5000 samples for the scalar effect-size analysis.
Training augmentations		
Channel dropout	Shared training augmentation	Applied with probability 0.25; dropped-channel ratio sampled uniformly from 0 to 0.3.
Temporal cutout	Shared training augmentation	Applied with probability 0.25; contiguous zeroed segment length 10–50 samples.
Gaussian noise	Shared training augmentation	Applied with probability 0.30; noise standard deviation $0.01 + 0.01\lvert\mathcal{N}(0,1)\rvert$.
Shift-jitter	Intervention experiments only	Training crop start sampled from 0.2–0.8 s, crop length 2 s, target expressed relative to crop start.

External EEG architectures are instantiated from the Braindecode constructors (Schirrmeister et al., 2017) named in the YAML configuration files and trained from scratch under the same data split, target support, optimizer, augmentation, checkpoint-selection, and holdout-evaluation protocol. Seed-level summaries, selected readout temperatures, prediction files, and reported aggregate tables are stored under `benchmarks/experiments/`.

Upon publication, we will release the benchmark materials supporting this study, including YAML configurations, data-preparation scripts, training and evaluation runners, model and loss code, selected readout temperatures, seed-level summaries, prediction files, shifted-crop diagnostics, aggregate tables, and figure-generation scripts. When raw HBN-EEG redistribution is restricted by source dataset terms, the release will provide the preparation code and expected split structure needed to reconstruct the CCD benchmark from the released data.

Table 11: Reproducibility summary for the main controlled comparisons (continued).

Group	Setting	Value
Event-time readout		
Event grid	Temporal support	200 bins over 2 s; 10 ms per bin.
Soft target	Gaussian event-time label	$\sigma = 0.15$ s.
Likelihood-based objectives	Observation kernels	$\sigma_y = 0.15$ s: $\sigma_{\text{narrow}} = 0.10$ s, $\sigma_{\text{wide}} = 0.35$ s, $\alpha = 0.10$.
Posterior diagnostics	Shared-kernel RT NLL	$\sigma_{\text{score}} = 0.12$ s across objectives.
Posterior readout	Scalar prediction	Posterior mean / soft-argmax.
Base temperature	Training/evaluation default	$\tau = 0.65$ before post-hoc readout-temperature selection.
Temperature selection	Readout tuning	Grid search on R9–R10 minimizing posterior-mean nRMSE; selected τ applied unchanged to R11.
Temperature grid	Objective-specific grid	0.2–3.5 with step 0.05.
Model architectures		
ETS-U-Net	Primary event-time backbone	$c_0 = 96$, widen 2, depth per stage 5, kernel size 15, dropout 0.2, one output channel; 3.10M parameters.
ETS-TCN	Dilated-convolution control	$c_0 = 384$, depth 10, dilation cycle $1,2,4,8,16$ repeated twice, kernel size 15, dropout 0.2, one output channel; 3.13M parameters.
ETS-InceptionPyramid	Multi-scale temporal control	$c_0 = 288$, branch width 96, depth 6, temporal scales 0.05/0.10/0.25/0.50 s, stem kernel 7, refinement kernel 11, dropout 0.2, one output channel; 3.25M parameters.
ETS-AttnSeg	Attention-convolution control	$c_0 = 128$, depth 10, eight attention heads, convolution kernel 15, feed-forward and convolution expansion 2, dropout 0.2, attention dropout 0.1, one output channel; 3.05M parameters.
ETR-CNN large	Strongest scalar baseline	$c_0 = 96$, widen 3, kernel size 15, dropout 0.2, high-resolution depth 12, low-resolution depth 10, three low-resolution stacks, two downsampling stages; 12.87M parameters.
ETR-CNN	Scalar temporal-readout baseline	Same family as ETR-CNN large with $c_0 = 64$; 5.91M parameters.
MSP-CNN	Segment-pooling scalar baseline	$c_0 = 48$, widen 2, depth per stage 3, kernel size 15, dropout 0.05; 3.01M parameters.

Acknowledgments

We would like to acknowledge support from NSF Award No. 2229881 for the AI Institute for Societal Decision Making (AI-SDM), the National Institutes of Health (NIH) under Contract R01HL159805, the MBZUAI-WIS Joint Program, the MBZUAI Startup Fund, and the Al Deira Causal Education project.

References

Altaheri, H., Muhammad, G., & Alsulaiman, M. (2023). Physics-informed attention temporal convolutional network for EEG-based motor imagery classification. *IEEE Transactions on Industrial Informatics*, *19*(2), 2249–2258. https://doi.org/10.1109/TII.2022.3197419

Aristimunha, B., Truong, D., Guetschel, P., Shirazi, S. Y., Guyon, I., Franco, A. R., Milham, M. P., Dotan, A., Makeig, S., Gramfort, A., King, J.-R., Corsi, M.-C., Valdés-Sosa, P. A., Majumdar, A., Evans, A., Sejnowski, T. J., Shriki, O., Chevallier, S., & Delorme, A. (2025). EEG Foundation Challenge: From cross-task to cross-subject EEG decoding [Preprint]. arXiv. https://arxiv.org/abs/2506.19141

Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*. https://arxiv.org/abs/1803.01271

Barachant, A., Bonnet, S., Congedo, M., & Jutten, C. (2012). Multiclass brain–computer interface classification by Riemannian geometry. *IEEE Transactions on Biomedical Engineering*, *59*(4), 920–928. https://doi.org/10.1109/TBME.2011.2172210

Bigdely-Shamlo, N., Cockfield, J., Makeig, S., Rognon, T., La Valle, C., Miyakoshi, M., & Robbins, K. A. (2016). Hierarchical event descriptors (HED): Semi-structured tagging for real-world events in large-scale EEG. *Frontiers in Neuroinformatics*, *10*, 42. https://doi.org/10.3389/fninf.2016.00042

Chambon, S., Thorey, V., Arnal, P. J., Mignot, E., & Gramfort, A. (2019). DOSED: A deep learning approach to detect multiple sleep micro-events in EEG signal. *Journal of Neuroscience Methods*, *321*, 64–78. https://doi.org/10.1016/j.jneumeth.2019.03.017

Chapfuwa, P., Tao, C., Li, C., Khan, I., Chandross, K. J., Pencina, M. J., Carin, L., & Henao, R. (2023). Calibration and uncertainty in neural time-to-event modeling. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(4), 1666–1680. https://doi.org/10.1109/TNNLS.2020.3029631

Chen, Y., Ren, K., Song, K., Wang, Y., Wang, Y., Li, D., & Qiu, L. (2024). EEGFormer: Towards transferable and interpretable large-scale EEG foundation model [Preprint]. arXiv. https://arxiv.org/abs/2401.10278

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*

(CVPR) (pp. 1251–1258). https://openaccess.thecvf.com/content_cvpr_2017/html/Chollet_Xception_Deep_Learning_CVPR_2017_paper.html

Chowdhury, M. S. N., Dutta, A., Robison, M. K., Blais, C., Brewer, G. A., & Bliss, D. W. (2020). Deep neural network for visual stimulus-based reaction time estimation using the periodogram of single-trial EEG. *Sensors*, *20*(21), 6090. https://doi.org/10.3390/s20216090

Coles, M. G. H. (1989). Modern mind-brain reading: Psychophysiology, physiology, and cognition. *Psychophysiology*, *26*(3), 251–269. https://doi.org/10.1111/j.1469-8986.1989.tb01916.x

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187–202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

EEG Challenge (2025). NeurIPS EEG Foundation Challenge. `https://eeg2025.github.io/`. Accessed 2026-02-04.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., & Lempitsky, V. (2016). Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, *17*, 1–35. arXiv:1505.07818. https://doi.org/10.48550/arXiv.1505.07818

Gehring, W. J., Goss, B., Coles, M. G. H., Meyer, D. E., & Donchin, E. (1993). A neural system for error detection and compensation. *Psychological Science*, *4*(6), 385–390. https://doi.org/10.1111/j.1467-9280.1993.tb00586.x

Geng, X. (2016). Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, *28*(7), 1734–1748. https://doi.org/10.1109/TKDE.2016.2545658

Gensheimer, M. F., & Narasimhan, B. (2019). A scalable discrete-time survival model for neural networks. *PeerJ*, *7*, e6257. https://doi.org/10.7717/peerj.6257

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., & Pang, R. (2020). Conformer: Convolution-augmented transformer for speech recognition. In *Proceedings of Interspeech* (pp. 5036–5040). https://doi.org/10.21437/Interspeech.2020-3015

Haider, H., Hoehn, B., Davis, S., & Greiner, R. (2020). Effective ways to build and evaluate individual survival distributions. *Journal of Machine Learning Research*, *21*(85), 1–63. https://jmlr.org/papers/v21/18-772.html

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

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778). https://doi.org/10.1109/CVPR.2016.90

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, *109*(4), 679–709. https://doi.org/10.1037/0033-295X.109.4.679

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, *abs/1704.04861*. https://arxiv.org/abs/1704.04861

Jayaram, V., Alamgir, M., Altun, Y., Schölkopf, B., & Grosse-Wentrup, M. (2016). Transfer learning in brain–computer interfaces. *IEEE Computational Intelligence Magazine*, *11*(1), 20–31. https://doi.org/10.1109/MCI.2015.2501545

Jiang, W., Zhao, L., & Lu, B.-L. (2024). Large brain model for learning generic representations with tremendous EEG data in BCI. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=QzTpTRVtrP

Jung, T.-P., Makeig, S., Westerfield, M., Townsend, J., Courchesne, E., & Sejnowski, T. J. (1998). Analyzing and visualizing single-trial event-related potentials. In *Advances in Neural Information Processing Systems* (Vol. 11, pp. 118–124). https://proceedings.neurips.cc/paper/1998/hash/0d4f4805c36dc6853edfa4c7e1638b48-Abstract.html

Kelly, S. P., & O'Connell, R. G. (2013). Internal and external influences on the rate of sensory evidence accumulation in the human brain. *Journal of Neuroscience*, *33*(50), 19434–19441. https://doi.org/10.1523/JNEUROSCI.3355-13.2013

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1412.6980

Kornhuber, H. H., & Deecke, L. (1965). Hirnpotentialaenderungen bei willkuerbewegungen und passiven bewegungen des menschen: Bereitschaftspotential und reafferente potentiale. *Pfluegers Archiv fuer die gesamte Physiologie des Menschen und der Tiere*, *284*(1), 1–17. https://doi.org/10.1007/BF00412364

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

Kostas, D., & Rudzicz, F. (2020). Thinker invariance: enabling deep neural networks for BCI across more people. *Journal of Neural Engineering*, *17*, 056008. https://doi.org/10.1088/1741-2552/abb7a7

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., & Lance, B. J. (2018). EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, *15*(5), 056013. https://doi.org/10.1088/1741-2552/aace8c

Lee, C., Zame, W. R., Yoon, J., & van der Schaar, M. (2018). DeepHit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32, pp. 2314–2321). https://doi.org/10.1609/aaai.v32i1.11842

O'Connell, R. G., Dockree, P. M., & Kelly, S. P. (2012). A supramodal accumulation-to-bound signal that determines perceptual decisions in humans. *Nature Neuroscience*, *15*(12), 1729–1735. https://doi.org/10.1038/nn.3248

Ouahidi, Y. E., Lys, J., Thölke, P., Farrugia, N., Pasdeloup, B., Gripon, V., Jerbi, K., & Lioi, G. (2025). REVE: A foundation model for EEG – adapting to any setup with large-scale pretraining on 25,000 subjects [Preprint]. arXiv. https://arxiv.org/abs/2510.21585

Ouyang, G., Herzmann, G., Zhou, C., & Sommer, W. (2011). Residue iteration decomposition (RIDE): A new method to separate ERP components on the basis of latency variability in single trials. *Psychophysiology*, *48*(12), 1631–1647. https://doi.org/10.1111/j.1469-8986.2011.01269.x

Ouyang, G., Hildebrandt, A., Sommer, W., & Zhou, C. (2017). Exploiting the intra-subject latency variability from single-trial event-related potentials in the P3 time range: A review and comparative evaluation of methods. *Neuroscience and Biobehavioral Reviews*, *75*, 1–21. https://doi.org/10.1016/j.neubiorev.2017.01.023

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922. https://doi.org/10.1162/neco.2008.12-06-420

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (*Lecture Notes in Computer Science*, Vol. 9351, pp. 234–241). Springer. https://doi.org/10.1007/978-3-319-24574-4_28

Santamaría-Vázquez, E., Martínez-Cagigal, V., Vaquerizo-Villar, F., & Hornero, R. (2020). EEG-Inception: A novel deep convolutional neural network for assistive ERP-based brain-computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *28*(12), 2773–2782. https://doi.org/10.1109/TNSRE.2020.3048106

Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., & Ball, T. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, *38*(11), 5391–5420. https://doi.org/10.1002/hbm.23730

Shibasaki, H., & Hallett, M. (2006). What is the Bereitschaftspotential? *Clinical Neurophysiology*, *117*(11), 2341–2356. https://doi.org/10.1016/j.clinph.2006.04.025

Shirazi, S. Y., Franco, A., Hoffmann, M. S., Esper, N., Truong, D., Delorme, A., Milham, M., & Makeig, S. (2024). HBN-EEG: The FAIR implementation of the healthy brain

network (HBN) electroencephalography dataset. *bioRxiv*. https://doi.org/10.1101/2024.10.03.615261

Song, Y., Zheng, Q., Liu, B., & Gao, X. (2023). EEG Conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *31*, 710–719. https://doi.org/10.1109/TNSRE.2022.3230250

Sun, B., & Saenko, K. (2016). Deep CORAL: Correlation alignment for deep domain adaptation. In *Computer Vision – ECCV 2016 Workshops* (*Lecture Notes in Computer Science*, Vol. 9915, pp. 443–450). Springer. https://doi.org/10.1007/978-3-319-49409-8_35

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1–9). https://doi.org/10.1109/CVPR.2015.7298594

Tapia-Rivas, N. I., Estevez, P. A., & Cortes-Briones, J. A. (2024). A robust deep learning detector for sleep spindles and K-complexes: Towards population norms. *Scientific Reports*, *14*, 263. https://doi.org/10.1038/s41598-023-50736-7

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)* (Vol. 30, pp. 5998–6008). https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Wang, G., Liu, W., He, Y., Xu, C., Ma, L., & Li, H. (2024a). EEGPT: Pretrained transformer for universal and reliable representation of EEG signals. In *Advances in Neural Information Processing Systems* (Vol. 37, pp. 39249–39280). Curran Associates, Inc. https://doi.org/10.52202/079017-1239

Wang, J., Zhao, S., Luo, Z., Zhou, Y., Li, S., & Pan, G. (2025). EEGMamba: An EEG foundation model with Mamba. *Neural Networks*, *192*, 107816. https://doi.org/10.1016/j.neunet.2025.107816

Wang, Y., Huang, N., Li, T., Yan, Y., & Zhang, X. (2024b). Medformer: A multi-granularity patching transformer for medical time-series classification. In *Advances in Neural Information Processing Systems* (Vol. 37, pp. 36314–36341). https://doi.org/10.52202/079017-1145

Weindel, G., Borst, J. P., & van Maanen, L. (2025). Decision-making components and times revealed by the single-trial electroencephalogram. *eLife*, *14*, RP108049. https://doi.org/10.7554/eLife.108049.3

Wu, D., Lance, B. J., Lawhern, V. J., Gordon, S., Jung, T.-P., & Lin, C.-T. (2017). EEG-based user reaction time estimation using riemannian geometry features. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *25*(11), 2157–2168. https://doi.org/10.1109/TNSRE.2017.2699784

Yao, D., Huang, S., Cadei, R., Zhang, K., & Locatello, F. (2025). The third pillar of causal analysis? a measurement perspective on causal representations. In *Advances in Neural Information Processing Systems* (Vol. 38, pp. 43423–43454). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2025/hash/3dda5b6fa9e7c98928844e776f2a657a-Abstract-Conference.html

Yu, F., & Koltun, V. (2016). Multi-scale context aggregation by dilated convolutions. In *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1511.07122

Yue, T., Gao, X., Xue, S., Tang, Y., Guo, L., Jiang, J., & Liu, J. (2025). BrainGPT: Unleashing the potential of EEG generalist foundation model by autoregressive pretraining [Preprint]. arXiv. https://arxiv.org/abs/2410.19779

Zanini, P., Congedo, M., Jutten, C., Said, S., & Berthoumieu, Y. (2018). Transfer learning: A Riemannian geometry framework with applications to brain–computer interfaces. *IEEE Transactions on Biomedical Engineering*, *65*(5), 1107–1116. https://doi.org/10.1109/TBME.2017.2742541

Zhang, R. (2019). Making convolutional networks shift-invariant again. In *Proceedings of the 36th International Conference on Machine Learning* (*Proceedings of Machine Learning Research*, Vol. 97, pp. 7324–7334). PMLR. https://proceedings.mlr.press/v97/zhang19a.html