TRACE: TEMPORAL ROUTING WITH AUTOREGRESSIVE CROSS-CHANNEL EXPERTS FOR EEG REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning transferable representations for electroencephalography (EEG) remains challenging because EEG signals are inherently multi-channel and non-stationary. Channels observed at the same time provide coupled measurements of neural activity, while the relevant temporal dynamics vary across contexts. This structure is poorly matched by architectures that apply uniform computation across time or route each channel patch independently. To this end, we propose TRACE, an autoregressive EEG pre-training framework that predicts future EEG patches from observed history while performing temporally adaptive and cross-channel coherent computation. At each temporal step, TRACE derives an expert routing decision from the cross-channel history and applies it jointly to all channels at that step. This preserves instantaneous cross-channel coherence while allowing different temporal regimes to activate different computation. Since routing is defined over the available channel set and temporal history, TRACE is compatible with heterogeneous pre-training across corpora with different channel counts, montages, sequence lengths, and recording domains. Across eight downstream EEG benchmarks, TRACE is evaluated in both settings: when downstream domains are seen only as unlabeled pre-training data and when downstream datasets are completely unseen during pre-training. It achieves the best mean performance on six of the eight benchmarks and remains competitive on the two motor imagery tasks, with ablations supporting the importance of cross-channel temporal routing.

1 INTRODUCTION

Electroencephalography (EEG) is widely used in brain-computer interfaces (BCIs) and clinical monitoring because it is non-invasive and provides high temporal resolution (Lawhern et al., 2018; Song et al., 2022; Ma et al., 2025). However, learning transferable EEG representations remains challenging: EEG signals are noisy, non-stationary, and highly variable across montages, recording protocols, temporal scales, and subject populations (Jiang et al., 2024b; Wang et al., 2025; Ma et al., 2025). Consequently, models trained on one dataset or electrode layout often transfer poorly to another, motivating foundation models that learn general-purpose representations from large unlabeled EEG corpora.

Recent EEG foundation models are first pre-trained on large-scale unlabeled EEG and then fine-tuned on downstream tasks. Representative methods, including BIOT (Yang et al., 2023), LaBraM (Jiang et al., 2024b), CBraMod (Wang et al., 2025), and CodeBrain (Ma et al., 2025), have achieved noticeable improvements over task-specific architectures. Most follow a BERT-style paradigm (Devlin et al., 2018), masking portions of the input and reconstructing missing signals, tokens, or latent codes. This objective learns useful representations from unlabeled EEG, but reconstruction can draw on observations on both sides of a masked interval. It therefore does not explicitly require the representation at a given time to retain only the observed history and predict what follows. For EEG recorded as a continuous stream, forecasting future signal patches provides a complementary way to learn representations of temporal evolution.

Autoregressive (AR) pre-training formalizes this idea by predicting future EEG patches from past and current observations. Recent EEG models have explored related generative or predictive objectives,

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.

including NeuroLM (Jiang et al., 2024a) and EEGPT (Yue et al., 2024). Yet the prediction objective alone does not determine how a model should process multi-channel EEG: channels recorded at the same time provide complementary views of the evolving signal, while the useful temporal context can change across a recording. An effective architecture must therefore use the available cross-channel information at each step and adapt its computation over time without accessing future patches.

Mixture-of-experts (MoE) offers a way to make this computation conditional, activating a subset of specialized pathways according to the input (Shazeer et al., 2017; Zoph et al., 2022). Transformer MoE systems such as GShard and Switch Transformer typically select experts separately for each token (Lepikhin et al., 2021; Fedus et al., 2022). Applied directly to multi-channel EEG, this allows patches recorded at the same instant on different electrodes to take different computational paths, without explicitly coordinating their expert selection. Using one routing decision for an entire recording would instead forgo adaptation as its temporal context changes. Our goal is therefore to share expert selection across channels at each step while updating that selection from the observed history.

To couple future-patch prediction with temporally adaptive, channel-coordinated computation, we introduce TRACE, a **Temporal Routing with Autoregressive Cross-channel Experts** framework for EEG pre-training. Its autoregressive objective trains the model to forecast future patches using only observed EEG history. Within each Temporal Routing MoE (TR-MoE) block, spatial-temporal attention processes that history, and TemporalFormer summarizes the resulting context across available channels. The Cross-Channel Temporal Routing FFN (CTR-FFN) then selects the same specialized expert subset for all channels at a given step and updates that selection as new patches are observed. This coordinates computation across simultaneous channels without merging their distinct representations, while allowing the active experts to change over time.

TRACE also supports heterogeneous EEG pre-training across different montages, recording protocols, and sequence lengths without projecting all recordings onto a common montage. Instead of assuming a fixed electrode layout, TRACE encodes each available channel as a temporal patch sequence and computes routing from the cross-channel context available up to the current step. Our pre-training corpus combines clinical EEG, high-density healthy-population EEG, and task-related EEG, covering 16-128 channels and 4-30 second windows. This diversity exposes the model to a broad range of acquisition settings and temporal dynamics while keeping the encoder and CTR-FFN compatible with variable channel counts.

We further evaluate TRACE on eight downstream datasets spanning six EEG task categories under two transfer regimes: seen-domain transfer, where the downstream domain is observed only through unlabeled pre-training data, and unseen-dataset generalization, where the downstream dataset is completely excluded from pre-training. TRACE achieves the best mean performance on six of the eight benchmarks and remains competitive on the two motor imagery tasks, while ablations confirm the benefits of autoregressive pre-training, CTR-FFN, and multi-source pre-training. To summarize, our main contributions are four-fold:

- We propose TRACE, an autoregressive EEG pre-training framework that forecasts future EEG patches from observed history with temporally adaptive and cross-channel coherent computation.
- We introduce the TR-MoE block, combining spatial-temporal attention with a CTR-FFN that routes same-step channels via a shared temporal-state decision.
- We pre-train TRACE on a heterogeneous multi-source EEG corpus spanning different channel counts, acquisition settings, and task paradigms, without enforcing a common montage.
- We show that TRACE achieves the best mean performance on six of eight downstream benchmarks, and validate our designs through comprehensive ablations.

2 METHOD

2.1 PROBLEM SETUP AND AUTOREGRESSIVE PATCH REPRESENTATION

Given an EEG segment $\mathbf{S} \in \mathbb{R}^{C \times T}$ with C channels and T timestamps, we split each channel into $n = \lfloor T/t \rfloor$ non-overlapping patches of length t:

$$\mathbf{X} = \{\mathbf{x}_{i,j} \mid i = 1, \ldots, C,\ j = 1, \ldots, n\}, \qquad \mathbf{x}_{i,j} \in \mathbb{R}^t \tag{1}$$

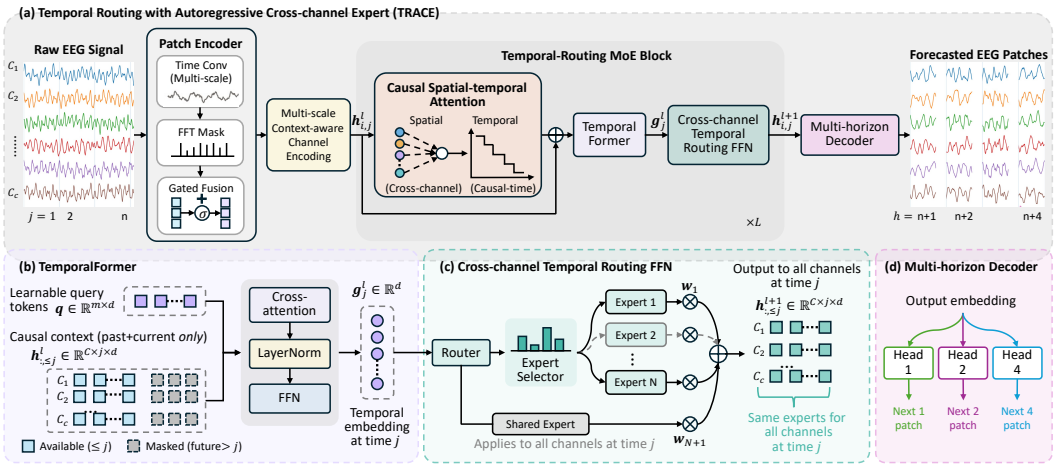


Figure 1: **Overview of TRACE.** TRACE partitions raw EEG into channel-wise temporal patches, embeds each patch with time-frequency features and multi-scale context-aware channel encoding, and processes the resulting sequence using stacked Temporal Routing MoE (TR-MoE) blocks. Each TR-MoE block combines causal spatial-temporal attention with a Cross-Channel Temporal Routing FFN (CTR-FFN). For each temporal step, TemporalFormer summarizes the cross-channel history, and a router selects top-K temporal experts shared by all channels at that step, together with an always-on shared expert. A multi-horizon decoder predicts future EEG patches under autoregressive and expert-balancing objectives.

where i indexes channels and j indexes temporal patches. TRACE learns an autoregressive patch representation: the representation at temporal step j is computed only from patches up to j, *i.e.*, it cannot access $\{\mathbf{x}_{i,j'} \mid j' > j\}$. The learned representation is then used to predict future patches autoregressively.

2.2 Overview of Trace

As shown in Figure 1, TRACE first maps raw channel-wise patches into time-frequency embeddings with a patch encoder and channel encoding. A causally masked transformer backbone then processes the patch sequence under the autoregressive constraint. The core component is a Temporal Routing MoE (TR-MoE) block, where a Cross-Channel Temporal Routing FFN (CTR-FFN) uses TemporalFormer to select temporal experts shared across channels at each step. Finally, TRACE predicts multiple future patch ranges from each past-conditioned representation.

2.3 Patch Encoder and Channel Encoding

The patch encoder maps each channel-wise temporal patch into a d-dimensional time-frequency embedding. It uses multi-scale convolution, FFT-based spectral weighting, and gated fusion. These auxiliary choices follow prior multi-scale EEG encoders and frequency-domain self-supervision (Santamaria-Vazquez et al., 2020; Wu et al., 2022; Wang et al., 2023). A *temporal patch* is $\mathbf{x}_{i,j}$ of length t from Section 2.1.

In-Patch Multi-scale Convolution. The first module applies several 1D convolutional branches with different temporal kernel sizes to each patch, allowing the encoder to capture local waveform patterns at multiple scales within the same patch. Let $\mathcal{Q}_{\text{temp}}$ denote the set of temporal kernel sizes. We summarize this module as:

$$\phi_q(\mathbf{x}_{i,j}) = \text{GELU}(\text{GN}(\text{Conv1d}_3\left(\text{GELU}(\text{GN}(\text{Conv1d}_q(\mathbf{x}_{i,j}))\right))\right)), \qquad q \in \mathcal{Q}_{\text{temp}}, \tag{2}$$

$$\mathbf{e}_{i,j}^{\text{temp}} = \mathbf{W}_{\text{temp}}\left[\phi_q(\mathbf{x}_{i,j})\right]_{q\in\mathcal{Q}_{\text{temp}}} \in \mathbb{R}^d, \tag{3}$$

where $\phi_q(\cdot)$ is the branch output for kernel size q, $[\cdot]_{q\in\mathcal{Q}_{\text{temp}}}$ denotes concatenation over kernel sizes, and $\mathrm{GN}(\cdot)$ denotes group normalization.

FFT-Based Spectral Masking. The second module transforms the same patch into the frequency domain and applies a learnable spectral mask to emphasize informative frequency components before projection by:

$$\mathbf{e}_{i,j}^{\text{freq}} = \mathrm{FC}(|\mathbf{M}_\theta \odot \mathcal{F}(\mathbf{x}_{i,j})|) \in \mathbb{R}^d, \tag{4}$$

where $\mathcal{F}(\cdot)$ is the discrete Fourier transform, $\mathbf{M}_\theta$ is a learnable spectral mask, and $\odot$ denotes element-wise multiplication.

Gated Fusion. The third module fuses the time-domain embedding $\mathbf{e}_{i,j}^{\text{temp}}$ and the frequency-domain embedding $\mathbf{e}_{i,j}^{\text{freq}}$ with a learned gate via:

$$\mathbf{z}_{i,j} = \sigma\Big(\mathbf{W}_{\text{g}}[\mathbf{e}_{i,j}^{\text{temp}}; \mathbf{e}_{i,j}^{\text{freq}}]\Big), \tag{5}$$

$$\mathbf{e}_{i,j} = \mathbf{e}_{i,j}^{\text{temp}} + \mathbf{z}_{i,j} \odot \mathbf{e}_{i,j}^{\text{freq}}. \tag{6}$$

Thus, $\mathbf{e}_{i,j}$ is the time-frequency embedding of patch $\mathbf{x}_{i,j}$, and the resulting tensor is $\mathbf{E} \in \mathbb{R}^{C\times n\times d}$.

Multi-scale Context-aware Channel Encoding (MS-CCE). To encode channel context without using future patches, we add multi-scale context-aware channel encoding (MS-CCE). It applies input-conditioned convolutions along the fixed, dataset-consistent native channel order to capture channel dependencies at multiple scales. Inspired by conditional positional encoding (Chu et al., 2023; Wang et al., 2025), MS-CCE uses zero-padded depth-wise 1D convolutions:

$$\mathbf{E}^{\text{p}} = \sum_{r=1}^{R} \mathrm{DWConv1d}_{q_r}(\mathbf{E}), \tag{7}$$

$$\mathbf{E}^{\text{o}} = \mathbf{E} + \mathbf{E}^{\text{p}}. \tag{8}$$

The convolution is applied along the channel axis at each patch step, so the channel encoding at step j only uses same-step channel embeddings. We set $\mathbf{h}_{i,j}^0 = \mathbf{e}_{i,j}^{\text{o}}$ as the input to the backbone.

2.4 Temporal Routing MoE

The Temporal Routing MoE (TR-MoE) block first applies causal spatial-temporal attention (CSTA), then replaces the standard feed-forward sublayer with a Cross-Channel Temporal Routing FFN (CTR-FFN):

$$\mathbf{u}_{i,j}^l = \mathrm{CSTA}\big(\mathrm{LN}(\mathbf{h}_{i,j}^l)\big) + \mathbf{h}_{i,j}^l, \tag{9}$$

$$\mathbf{h}_{i,j}^{l+1} = \mathrm{CTRFFN}\big(\mathrm{LN}(\mathbf{u}_{i,j}^l),\, \mathbf{g}_j^l\big) + \mathbf{u}_{i,j}^l, \tag{10}$$

where $\mathrm{CSTA}(\cdot)$ denotes causal spatial-temporal multi-head attention, which combines cross-channel attention within the same temporal step with causal temporal attention along the patch axis; implementation details are provided in Appendix G. Cross-channel attention supports heterogeneous EEG inputs because each temporal step can be treated as a set of channel tokens: regardless of the dataset-specific channel count or montage, CSTA can model instantaneous inter-channel interactions over the available channels. $\mathrm{CTRFFN}(\cdot)$ denotes the CTR-FFN, and $\mathbf{g}_j^l \in \mathbb{R}^d$ is a temporal routing context produced by TemporalFormer.

Cross-channel Temporal Routing FFN (CTR-FFN). It contains N specialized experts and one shared expert. For each temporal step j, a router maps $\mathbf{g}_j^l$ to top-K specialized experts. The selected experts are shared across all channels at that temporal step, so channels from different montages are routed according to their shared temporal context rather than independent channel-specific decisions.

This channel-shared routing preserves within-step coherence and further supports heterogeneous EEG pre-training with variable channel counts:

$$\mathrm{CTRFFN}(\mathbf{x}, \mathbf{g}_j^l) = \sum_{k=1}^{N} r_k(\mathbf{g}_j^l)\,\mathrm{Expert}_k(\mathbf{x}) + \mathrm{Expert}_{N+1}(\mathbf{x}), \tag{11}$$

$$r_k(\mathbf{g}_j^l) = \begin{cases} \dfrac{\exp(\mathbf{w}_k^\top \mathbf{g}_j^l)}{\sum_{k' \in \mathcal{T}_j^l} \exp(\mathbf{w}_{k'}^\top \mathbf{g}_j^l)}, & k \in \mathcal{T}_j^l, \\ 0, & \text{otherwise}, \end{cases} \tag{12}$$

where $\mathcal{T}_j^l$ is the top-K expert set selected at layer l and step j. The shared expert Expert_{N+1} is always active and does not participate in the top-K competition, providing common computation across datasets and channel layouts.

TemporalFormer. TemporalFormer constructs the routing context $\mathbf{g}_j^l$ from the cross-channel history. It uses m learnable query tokens $\mathbf{Q} \in \mathbb{R}^{m \times d}$ that attend to all channel patches up to step j:

$$\tilde{\mathbf{G}}_j^l = \mathrm{MSA}\big(\mathbf{Q},\, \mathbf{H}_{:,\le j}^l,\, \mathbf{H}_{:,\le j}^l\big), \qquad \mathbf{H}_{:,\le j}^l = \{\mathbf{h}_{i,j'}^l \mid i = 1, \ldots, C,\ j' \le j\}. \tag{13}$$

The query outputs are pooled and passed through a feed-forward layer:

$$\mathbf{g}_j^l = \mathrm{FFN}\Big(\mathrm{LN}\Big(\mathrm{Pool}(\tilde{\mathbf{G}}_j^l)\Big)\Big). \tag{14}$$

Since keys and values are restricted to $\mathbf{H}_{:,\le j}^l$, the routing decision at step j uses no future patches. The fixed number of query tokens also lets the router summarize variable-length and variable-channel EEG contexts into a fixed-dimensional routing vector.

2.5 Autoregressive Multi-horizon Pre-training

TRACE is trained with a multi-horizon autoregressive forecasting loss and an auxiliary expert-balancing loss.

Multi-horizon Forecasting. At each channel i and temporal step j, the final hidden state $\mathbf{h}_{i,j}^L$ summarizes the context up to patch j. Rather than predicting only the immediate next patch, TRACE predicts multiple future ranges from this same representation. Specifically, for each horizon $\rho \in \mathcal{H}$, a prediction head outputs the next ρ patches:

$$\hat{\mathbf{x}}_{i,j+1:j+\rho}^{(\rho)} = \mathrm{Head}_\rho(\mathbf{h}_{i,j}^L). \tag{15}$$

For example, when $\mathcal{H} = \{1, 2, 4\}$, the model learns to predict the next 1 patch, the next 2 patches, and the next 4 patches in parallel. This multi-horizon design encourages the shared backbone to capture both short-range waveform continuity and longer-range EEG dynamics. Because raw EEG is non-stationary and often contains transient artifacts, we use the Huber loss (Huber, 1992). The autoregressive loss is thus defined as:

$$\mathcal{L}_{\mathrm{AR}} = \frac{1}{|\mathcal{H}|} \sum_{\rho \in \mathcal{H}} \frac{1}{|\mathcal{V}_\rho|} \sum_{(i,j) \in \mathcal{V}_\rho} \mathcal{L}_\delta\Big(\mathbf{x}_{i,j+1:j+\rho},\, \hat{\mathbf{x}}_{i,j+1:j+\rho}^{(\rho)}\Big), \tag{16}$$

where $\mathcal{L}_\delta$ denotes the Huber loss with threshold δ, and $\mathcal{V}_\rho = \{(i,j) \mid j + \rho \le n\}$ is the set of valid prediction positions for horizon ρ.

Expert Balancing. To prevent routing collapse, we use a load-balancing objective adapted from MoE training (Dai et al., 2024). Let f_k be the fraction of temporal steps routed to expert k within a batch, and let p_k be the average routing probability assigned to expert k. The auxiliary loss is:

$$\mathcal{L}_{\mathrm{aux}} = N \sum_{k=1}^{N} f_k p_k. \tag{17}$$

Total Objective. The final pre-training objective is thus formulated as:

$$\mathcal{L} = \mathcal{L}_{\mathrm{AR}} + \lambda_{\mathrm{aux}}\mathcal{L}_{\mathrm{aux}}, \tag{18}$$

where λ_{aux} controls the strength of the expert-balancing regularization. We set λ_{aux}=1e$-$2 in all pre-training experiments (Table 7). Appendix H.9 analyzes its effect on expert utilization, and Appendix H.10 discusses rare-event and localization limitations.

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

Pre-training Data and Implementation Details. We pre-train TRACE on a heterogeneous multi-source EEG corpus consisting of the Temple University Hospital EEG Corpus (TUEG) (Obeid & Picone, 2016), HBN (Alexander et al., 2017), and several task-related datasets, with channel counts ranging from 16 to 128. Following prior EEG foundation model preprocessing protocols (Jiang et al., 2024b; Wang et al., 2025; Ma et al., 2025), EEG signals are band-pass filtered between 0.5-75 Hz, notch-filtered at 60 Hz, resampled to 200 Hz, and divided into non-overlapping windows. Across the six sources, preprocessing yields over 1.5 million available EEG segments; pre-training uses only the training-split subjects from sources that overlap with downstream evaluation. For downstream datasets that overlap with task-related pre-training sources, only the training-split subjects are used as unlabeled pre-training data, while validation and test subjects are strictly held out. This creates two evaluation regimes: *seen-domain transfer*, where the downstream domain is observed only through unlabeled pre-training data, and *unseen-dataset generalization*, where the downstream dataset is completely excluded from pre-training. TRACE is pre-trained with a multi-horizon autoregressive forecasting objective using $\mathcal{H} = \{1, 2, 4\}$ on 4 NVIDIA H100 GPUs. Additional training details are provided in Appendix D.

Downstream Tasks and Datasets. We assess the generalization ability of TRACE across a diverse set of EEG tasks, including *Sleep Staging* (*e.g.*, ISRUC (Khalighi et al., 2016)), *Emotion Recognition* (*e.g.*, SEED-V (Liu et al., 2021), FACED (Chen et al., 2023)), *Motor Imagery* (*e.g.*, PhysioNet-MI (Schalk et al., 2004), SHU-MI (Ma et al., 2022)), *Seizure Detection* (*e.g.*, CHB-MIT (Shoeb, 2009)), *Imagined Speech* (*e.g.*, BCIC2020-3 (Jeong et al., 2022)), and *Event Classification* (*e.g.*, TUEV (Obeid & Picone, 2016)). Among the six main downstream benchmarks, CHB-MIT and PhysioNet-MI evaluate seen-domain transfer, whereas FACED, SEED-V, ISRUC, and BCIC2020-3 evaluate unseen-dataset generalization; among the two additional benchmarks, SHU-MI also evaluates seen-domain transfer, and TUEV, a subset of the TUH EEG corpus, falls under seen-domain transfer with TUEG pre-training, consistent with prior EEG foundation models (Jiang et al., 2024b; Wang et al., 2025; Ma et al., 2025). Detailed dataset information is provided in Appendix B.

Baselines and Evaluation Metrics. We compare TRACE against several state-of-the-art methods, grouped into two categories: *1) Task-specific models*, including EEGNet (Lawhern et al., 2018) and EEGConformer (Song et al., 2022), and *2) EEG foundation models*, such as BIOT (Yang et al., 2023), LaBraM (Jiang et al., 2024b), CBraMod (Wang et al., 2025), and CodeBrain (Ma et al., 2025). For evaluation, we report Balanced Accuracy (B-Acc), area under the precision-recall curve (AUC-PR), and area under the receiver operating characteristic curve (AUROC) on binary tasks, and B-Acc, Cohen's Kappa (Kappa), and weighted F1 (F1-W) on multi-class tasks. TRACE results are averaged over five runs with different random seeds; baseline results are taken from the corresponding publications. We attach a three-layer MLP classification head and fine-tune the encoder and head end-to-end. Dataset-specific settings are listed in Appendix E and Table 9.

3.2 COMPARISON WITH STATE-OF-THE-ART METHODS

Table 1 presents comparisons of TRACE against state-of-the-art task-specific and EEG foundation models across six main downstream EEG datasets; additional results on SHU-MI and TUEV are reported in Appendix F. Color highlights indicate numerical ranks over five-seed means. These results demonstrate strong transfer performance, with clear gains on several benchmarks and competitive results on the remaining tasks, leading to two main conclusions:

Table 1: Performance comparison of various models across six main EEG datasets. The highest and second-highest reported means are highlighted in **red** and blue, respectively. Colors indicate numerical ranks only, not statistically significant differences.

Method	FACED (9-Class)			SEED-V (5-Class)		
	Balanced Acc	Kappa	Weighted F1	Balanced Acc	Kappa	Weighted F1
EEGNet	0.4090 ± 0.0122	0.3342 ± 0.0251	0.4124 ± 0.0141	0.2961 ± 0.0102	0.1006 ± 0.0143	0.2749 ± 0.0098
EEGConformer	0.4559 ± 0.0125	0.3858 ± 0.0186	0.4514 ± 0.0107	0.3537 ± 0.0112	0.1772 ± 0.0174	0.3487 ± 0.0136
CNN-Transformer	0.4697 ± 0.0132	0.4017 ± 0.0168	0.4720 ± 0.0125	0.3678 ± 0.0078	0.2072 ± 0.0183	0.3642 ± 0.0088
FFCL	0.4673 ± 0.0158	0.3987 ± 0.0383	0.4699 ± 0.0145	0.3641 ± 0.0092	0.2078 ± 0.0201	0.3645 ± 0.0132
ST-Transformer	0.4810 ± 0.0079	0.4137 ± 0.0133	0.4795 ± 0.0096	0.3052 ± 0.0072	0.1083 ± 0.0121	0.2833 ± 0.0105
BIOT	0.5118 ± 0.0118	0.4476 ± 0.0254	0.5136 ± 0.0112	0.3837 ± 0.0187	0.2261 ± 0.0262	0.3856 ± 0.0203
LaBraM-Base	0.5273 ± 0.0107	0.4698 ± 0.0188	0.5288 ± 0.0102	0.3976 ± 0.0138	0.2386 ± 0.0209	0.3974 ± 0.0111
CBraMod	0.5509 ± 0.0089	0.5041 ± 0.0122	0.5618 ± 0.0093	0.4091 ± 0.0097	0.2569 ± 0.0143	0.4101 ± 0.0108
CodeBrain	0.5941 ± 0.0098	0.5406 ± 0.0084	0.5953 ± 0.0113	0.4137 ± 0.0023	0.2735 ± 0.0032	0.4235 ± 0.0022
TRACE (Ours)	**0.6352** ± 0.0037	**0.5883** ± 0.0042	**0.6390** ± 0.0043	**0.4186** ± 0.0032	**0.2759** ± 0.0030	**0.4253** ± 0.0024

Method	ISRUC (5-Class)			BCIC2020-3 (5-Class)		
	Balanced Acc	Kappa	Weighted F1	Balanced Acc	Kappa	Weighted F1
EEGNet	0.7154 ± 0.0121	0.7040 ± 0.0173	0.7513 ± 0.0124	0.4413 ± 0.0096	0.3016 ± 0.0123	0.4413 ± 0.0102
EEGConformer	0.7400 ± 0.0133	0.7143 ± 0.0162	0.7634 ± 0.0151	0.4506 ± 0.0133	0.3133 ± 0.0183	0.4488 ± 0.0154
CNN-Transformer	0.7363 ± 0.0087	0.7129 ± 0.0121	0.7719 ± 0.0105	0.4533 ± 0.0092	0.3166 ± 0.0118	0.4506 ± 0.0127
FFCL	0.7277 ± 0.0182	0.7016 ± 0.0291	0.7614 ± 0.0197	0.4678 ± 0.0197	0.3301 ± 0.0359	0.4689 ± 0.0205
ST-Transformer	0.7381 ± 0.0205	0.7013 ± 0.0352	0.7681 ± 0.0175	0.4126 ± 0.0122	0.2941 ± 0.0159	0.4247 ± 0.0138
BIOT	0.7527 ± 0.0121	0.7192 ± 0.0231	0.7790 ± 0.0146	0.4920 ± 0.0086	0.3650 ± 0.0176	0.4917 ± 0.0079
LaBraM	0.7633 ± 0.0102	0.7231 ± 0.0182	0.7810 ± 0.0133	0.5060 ± 0.0155	0.3800 ± 0.0242	0.5054 ± 0.0205
CBraMod	0.7865 ± 0.0110	0.7442 ± 0.0152	0.8011 ± 0.0099	0.5373 ± 0.0108	0.4216 ± 0.0163	0.5383 ± 0.0096
CodeBrain	0.7856 ± 0.0031	0.7671 ± 0.0091	0.8202 ± 0.0071	0.6101 ± 0.0052	0.5127 ± 0.0065	0.6101 ± 0.0053
TRACE (Ours)	**0.7997** ± 0.0013	**0.7711** ± 0.0048	**0.8209** ± 0.0040	**0.6315** ± 0.0055	**0.5393** ± 0.0069	**0.6316** ± 0.0057

Method	CHB-MIT (2-Class)			PhysioNet-MI (4-Class)		
	Balanced Acc	AUC-PR	AUROC	Balanced Acc	Kappa	Weighted F1
EEGNet	0.5658 ± 0.0106	0.1914 ± 0.0182	0.8048 ± 0.0136	0.5814 ± 0.0125	0.4468 ± 0.0199	0.5796 ± 0.0115
EEGConformer	0.5976 ± 0.0141	0.2209 ± 0.0215	0.8226 ± 0.0170	0.6049 ± 0.0104	0.4736 ± 0.0171	0.6062 ± 0.0095
CNN-Transformer	0.6389 ± 0.0067	0.2479 ± 0.0227	0.8662 ± 0.0082	0.6053 ± 0.0118	0.4725 ± 0.0223	0.6041 ± 0.0105
FFCL	0.6262 ± 0.0104	0.2049 ± 0.0346	0.8271 ± 0.0051	0.5726 ± 0.0092	0.4323 ± 0.0182	0.5701 ± 0.0079
ST-Transformer	0.5915 ± 0.0195	0.1422 ± 0.0094	0.8237 ± 0.0491	0.6035 ± 0.0081	0.4712 ± 0.0199	0.6053 ± 0.0075
BIOT	0.7068 ± 0.0457	0.3277 ± 0.0460	0.8761 ± 0.0284	0.6153 ± 0.0154	0.4875 ± 0.0272	0.6158 ± 0.0197
LaBraM	0.7075 ± 0.0358	0.3287 ± 0.0402	0.8679 ± 0.0199	0.6173 ± 0.0122	0.4912 ± 0.0192	0.6177 ± 0.0141
CBraMod	0.7398 ± 0.0284	0.3689 ± 0.0382	0.8892 ± 0.0154	**0.6417** ± 0.0091	**0.5222** ± 0.0169	**0.6427** ± 0.0100
CodeBrain	0.7273 ± 0.0240	0.4377 ± 0.0288	0.8961 ± 0.0174	–	–	–
TRACE (Ours)	**0.7453** ± 0.0178	**0.5921** ± 0.0795	**0.9220** ± 0.0232	0.6342 ± 0.0090	0.5122 ± 0.0120	0.6365 ± 0.0096

1) TRACE achieves higher mean performance on several evaluated datasets. The model demonstrates notable improvements on challenging decoding benchmarks. To be specific, on the imagined speech task, BCIC2020-3, TRACE surpasses CodeBrain and CBraMod in Balanced Accuracy by $+2.14$ and $+9.42$ points, respectively. Similar gains are observed on the emotion recognition task (FACED), where TRACE improves Balanced Accuracy over CodeBrain and CBraMod by $+4.11$ and $+8.43$ points. TRACE also attains the highest reported means on SEED-V, ISRUC, and CHB-MIT in Table 1. These numerical results support the effectiveness of TRACE on several representative EEG transfer settings.

2) TRACE generalizes across diverse EEG tasks and recording settings. TRACE performs well in both evaluation regimes: it achieves strong transfer when the downstream domain is seen only as unlabeled pre-training data, such as CHB-MIT and PhysioNet-MI, and it also improves on several datasets that are completely unseen during pre-training, including FACED, SEED-V, ISRUC, and BCIC2020-3. On PhysioNet-MI, TRACE ranks second behind CBraMod. Together, these results demonstrate transfer across different channel configurations, recording domains, and task protocols.

3.3 ABLATION STUDY

Autoregressive vs. Masked Pre-training. To compare pre-training objectives under a controlled setup, we use the same parameterized backbone, corpus, optimization budget, and downstream protocol for TRACE-AR and a masked-reconstruction variant; the attention mask and input corruption

Table 2: Effect of TemporalFormer query number. Results are reported as mean ± std over 5 random seeds, with the best result in each row highlighted in **bold**.

Dataset	1 Query			2 Queries			4 Queries		
	Balanced Acc (%)	Kappa (%)	Weighted F1 (%)	Balanced Acc (%)	Kappa (%)	Weighted F1 (%)	Balanced Acc (%)	Kappa (%)	Weighted F1 (%)
FACED	61.92 ± 0.47	57.00 ± 0.47	62.31 ± 0.47	62.31 ± 0.45	57.42 ± 0.50	62.65 ± 0.45	**63.52** ± 0.37	**58.83** ± 0.42	**63.90** ± 0.43
PhysioNet-MI	61.98 ± 0.53	49.30 ± 0.71	62.29 ± 0.57	63.02 ± 0.52	50.69 ± 0.69	63.27 ± 0.57	**63.42** ± 0.90	**51.22** ± 1.20	**63.65** ± 0.96
SEED-V	**40.50** ± 0.37	**25.78** ± 0.50	**40.98** ± 0.47	40.18 ± 0.42	25.41 ± 0.44	40.69 ± 0.32	40.32 ± 0.47	25.48 ± 0.57	40.76 ± 0.44
Average	54.80 ± 0.46	44.03 ± 0.56	55.19 ± 0.50	55.17 ± 0.46	44.51 ± 0.54	55.54 ± 0.45	**55.75** ± 0.58	**45.18** ± 0.73	**56.10** ± 0.61

follow each objective. TRACE-AR outperforms TRACE-Masked on four of five datasets (all except PhysioNet-MI, −1.05 points), with the largest gains on CHB-MIT (+16.73) and BCIC2020-3 (+8.56), raising macro-average balanced accuracy from 58.56% to 64.59% (+6.03 points). Full results are in Appendix H.4.

Effect of CTR-FFN. We first validate the core architectural design of TRACE. This comparison and its compute profile use a reduced configuration (N=8 specialized experts, top-K=2), whereas the final model in Table 1 uses N=64, K=8; other ablations vary only the settings specified in their respective comparisons. For a fair comparison, we construct two controlled variants: 1) a Non-MoE baseline where the CTR-FFN is replaced with a standard dense FFN, and 2) a Vanilla MoE version that keeps the expert structure but allows each channel patch to select experts via a token-wise router.

Figure 2 shows two key observations. *First*, standard token-wise MoE does not consistently improve performance: Vanilla MoE is nearly tied with Non-MoE on FACED and even decreases Balanced Accuracy from 0.398 to 0.384 on SEED-V, suggesting that simply replacing the FFN with token-wise MoE is insufficient for EEG modeling. *Second*, CTR-FFN achieves the highest mean performance across the three datasets. Taking FACED as an example, CTR-FFN improves over Vanilla MoE by +1.13 points in Balanced Accuracy, +1.28 in Kappa, and +1.24 in Weighted F1. This supports our hypothesis that shared same-step routing is useful for cross-channel coordination and temporal adaptation. FLOPs-matched expert-scaling and fine-grained routing analyses are in Appendices H.5 and H.10.

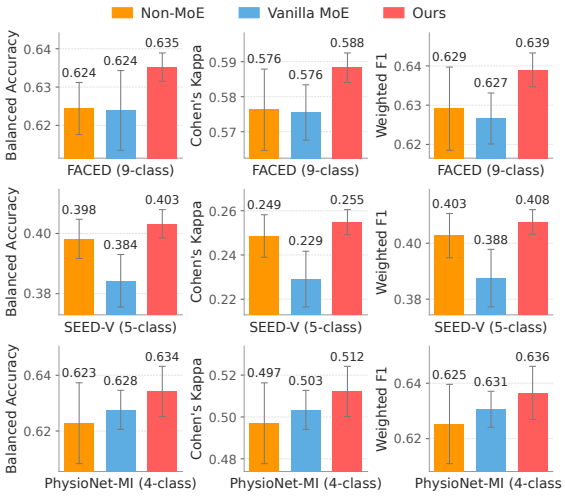


Figure 2: Routing ablations on three EEG datasets. Bars show means and error bars show ± one standard deviation over five seeds. Panel-specific vertical axes are truncated to show differences. Numerical values are in Appendix H.3.

Expert Pathways, Routing Behavior, and Cost. Removing the shared expert lowers balanced accuracy by 1.08 points on FACED and 1.18 points on BCIC2020-3 (Appendix H.6). Routing diagnostics on 256 held-out windows per dataset show within-window changes in expert weights while dataset, subject, and montage remain fixed; these changes do not by themselves identify physiological states (Appendix H.10). The controlled N=8, K=2 TRACE variant uses 24.72 GFLOPs per sample versus 18.08 for Non-MoE, so its gains accompany additional routing computation (Appendix H.11). These costs describe the ablation configuration, not the final N=64, K=8 model.

TemporalFormer Design. We further study the number of learnable query tokens in TemporalFormer, which produces the temporal-state representation used by the router for shared same-step routing. We vary the query number from 1 to 2 and 4, with all other settings fixed. As shown in Table 2, using more query tokens improves average downstream performance. The 4-query setting achieves the best average Balanced Accuracy, Kappa, and Weighted F1, outperforming the 1-query setting by +0.95, +1.15, and +0.91 points, respectively. It performs best on FACED and PhysioNet-

Table 3: MS-CCE ablation (%). $\{q_r\}_{r=1}^{R}$: kernel sizes of the R depth-wise convolutions along the channel axis; single-scale uses R=1; *w/o* MS-CCE removes the channel encoding.

MS-CCE Design	FACED		BCIC2020-3	
	Balanced Acc	Kappa	Balanced Acc	Kappa
Multi-scale $\{q_r\}_{r=1}^{R} = \{5, 11, 19\}$	**63.52** $\pm$ 0.37	**58.83**	**63.15** $\pm$ 0.55	**53.93**
Single-scale (R=1, q_1=5)	61.08 $\pm$ 0.78	56.04	60.16 $\pm$ 1.22	50.20
Single-scale (R=1, q_1=19)	62.34 $\pm$ 0.66	57.53	61.57 $\pm$ 1.18	51.97
w/o MS-CCE	59.80 $\pm$ 0.52	54.55	57.36 $\pm$ 1.13	46.70

Table 4: Pre-training data composition on CHB-MIT (%, mean $\pm$ std over five seeds). Phys-MI/Phys-MA: PhysioNet motor imagery/mental arithmetic.

Pre-training Sources						CHB-MIT		
TUEG	HBN	Phys-MI	Phys-MA	SHU-MI	CHB-MIT	Balanced Acc	AUC-PR	AUROC
✓	✗	✗	✗	✗	✗	58.91 $\pm$ 1.89	39.25 $\pm$ 10.81	89.46 $\pm$ 4.14
✗	✓	✗	✗	✗	✗	70.35 $\pm$ 0.87	56.54 $\pm$ 8.69	90.72 $\pm$ 4.24
✗	✗	✗	✗	✗	✓	71.08 $\pm$ 2.48	58.93 $\pm$ 7.49	91.95 $\pm$ 0.93
✓	✗	✓	✓	✓	✓	65.88 $\pm$ 4.10	44.00 $\pm$ 6.43	89.85 $\pm$ 1.89
✓	✓	✓	✓	✓	✓	**74.53** $\pm$ 1.78	**59.21** $\pm$ 7.95	**92.20** $\pm$ 2.32

MI, while SEED-V slightly favors 1 query, indicating task-dependent routing capacity. We therefore use 4 queries as the default TemporalFormer setting in TRACE.

Multi-scale Context-aware Channel Encoding. We next examine the MS-CCE introduced in Section 2.3. Removing it decreases Balanced Accuracy by 3.72 points on FACED and 5.79 points on BCIC2020-3. Wider single-scale kernels outperform narrower ones, while the multi-scale design performs best. These results support context-aware channel encoding over the fixed native order.

Effect of Multi-source Pre-training. Table 4 separates in-domain exposure from heterogeneous pre-training. CHB-MIT-only pre-training reaches 71.08% Balanced Accuracy and outperforms the TUEG-only and HBN-only settings. Using all six sources further raises Balanced Accuracy to 74.53%, a gain of 3.45 points over CHB-MIT-only pre-training, and gives the best mean AUC-PR and AUROC. Heterogeneous pre-training therefore adds value beyond target-domain exposure alone. Additional SEED-V results in Appendix H.2 show a consistent benefit from the full multi-source pre-training corpus.

4 Conclusion and Limitations

We presented TRACE, an autoregressive EEG foundation model for heterogeneous multi-source EEG. It combines multi-horizon future-patch forecasting with temporal routing that preserves same-step cross-channel coherence. Across eight benchmarks, TRACE achieves leading or competitive transfer performance. Matched-objective and pathway ablations show complementary benefits from autoregressive forecasting, shared computation, and routed experts.

Limitations. TRACE shares expert selection across all channels at each step, which may be suboptimal for focal events confined to a few electrodes, and the balancing loss may discourage experts from specializing on rare patterns. Its past-only objective forgoes bidirectional context that offline tasks could exploit, and MS-CCE assumes a consistent channel order within each dataset. Moreover, routing decisions are learned without supervision, so individual experts are not guaranteed to correspond to interpretable neural states. Finally, the routed expert pool increases storage and pre-training cost despite sparse activation.

REPRODUCIBILITY STATEMENT

We describe the model architecture and objectives in Section 2.1–2.5. The pre-training corpus, preprocessing, and optimization settings are detailed in Appendix D and Table 7; downstream datasets, splits, and fine-tuning hyperparameters are given in Appendix B, Appendix E, and Table 9. All pre-training and downstream datasets are publicly available, and all TRACE results are averaged over five random seeds. We will release code and pre-trained weights upon publication.

AI USE STATEMENT

In preparing this manuscript, the authors used generative AI tools to assist with language editing, manuscript organization, literature discovery, and the summarization and presentation of existing experimental results. The tools were not used to generate experimental data, determine the reported numerical results, or replace the authors' scientific judgment. All AI-assisted text, citations, claims, and numerical results were reviewed and verified by the authors, who take full responsibility for the final content of the manuscript.

REFERENCES

Lindsay M. Alexander, Jasmine Escalera, Lei Ai, Charissa Andreotti, Karina Febre, Alexander Mangone, Natan Vega-Potler, Nicolas Langer, Alexis Alexander, Meagan Kovacs, et al. An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Scientific Data*, 4:170181, 2017.

Jingjing Chen, Xiaobin Wang, Chen Huang, Xin Hu, Xinke Shen, and Dan Zhang. A large finer-grained affective computing EEG dataset. *Scientific Data*, 10(1):740, 2023.

Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, and Chunhua Shen. Conditional positional encodings for vision transformers. In *International Conference on Learning Representations (ICLR)*, 2023.

Damai Dai, Chengqi Deng, Chenggang Zhao, R. X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Yu Wu, et al. DeepSeekMoE: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*, 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformer: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Joshua I. Glaser, Matthew Whiteway, John P. Cunningham, Liam Paninski, and Scott W. Linderman. Recurrent switching dynamical systems models for multiple interacting neural populations. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Peter J. Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pp. 492–518. Springer, 1992.

Ji-Hoon Jeong, Jeong-Hyun Cho, Young-Eun Lee, Seo-Hyun Lee, Gi-Hwan Shin, Young-Seok Kweon, José del R. Millán, Klaus-Robert Müller, and Seong-Whan Lee. 2020 international brain–computer interface competition: A review. *Frontiers in Human Neuroscience*, 16:898300, 2022.

Wei-Bang Jiang, Yansen Wang, Bao-Liang Lu, and Dongsheng Li. NeuroLM: A universal multi-task foundation model for bridging the gap between language and EEG signals. *arXiv preprint arXiv:2409.00101*, 2024a. doi: 10.48550/arXiv.2409.00101.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *International Conference on Learning Representations (ICLR)*, 2024b.

Sirvan Khalighi, Teresa Sousa, José Moutinho Santos, and Urbano Nunes. ISRUC-Sleep: A comprehensive public dataset for sleep researchers. *Computer Methods and Programs in Biomedicine*, 124:180–192, 2016.

Demetres Kostas, Stephane Aroca-Ouellette, and Frank Rudzicz. BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data. *Frontiers in Human Neuroscience*, 15:653659, 2021.

Vernon J. Lawhern, Amelia J. Solon, Nicholas R. Waytowich, Stephen M. Gordon, Chou P. Hung, and Brent J. Lance. EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. GShard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations (ICLR)*, 2021.

Hongli Li, Man Ding, Ronghua Zhang, and Chunbo Xiu. Motor imagery EEG classification algorithm based on CNN-LSTM feature fusion network. *Biomedical Signal Processing and Control*, 72:103342, 2022.

Scott Linderman, Matthew Johnson, Andrew Miller, Ryan Adams, David Blei, and Liam Paninski. Bayesian learning and inference in recurrent switching linear dynamical systems. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 914–922, 2017.

Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2):715–729, 2021.

Jingying Ma, Feng Wu, Qika Lin, Yucheng Xing, Chenyu Liu, Ziyu Jia, and Mengling Feng. CodeBrain: Bridging decoupled tokenizer and multi-scale architecture for EEG foundation model. *arXiv preprint arXiv:2506.09110*, 2025.

Jun Ma, Banghua Yang, Wenzheng Qiu, Yunzhe Li, Shouwei Gao, and Xinxing Xia. A large EEG dataset for studying cross-session variability in motor imagery brain-computer interface. *Scientific Data*, 9(1):531, 2022.

Iyad Obeid and Joseph Picone. The Temple University Hospital EEG data corpus. *Frontiers in Neuroscience*, 10:196, 2016.

Wei Yan Peh, Yuanyuan Yao, and Justin Dauwels. Transformer convolutional neural networks for automated artifact detection in scalp EEG. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 3599–3602. IEEE, 2022.

Eduardo Santamaria-Vazquez, Victor Martinez-Cagigal, Fernando Vaquerizo-Villar, and Roberto Hornero. 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, 2020. doi: 10.1109/TNSRE.2020.3048106.

Gerwin Schalk, Dennis J. McFarland, Thilo Hinterberger, Niels Birbaumer, and Jonathan R. Wolpaw. BCI2000: A general-purpose brain-computer interface (BCI) system. *IEEE Transactions on Biomedical Engineering*, 51 (6):1034–1043, 2004.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017.

Ali Hossam Shoeb. *Application of machine learning to epileptic seizure onset detection and treatment*. PhD thesis, Massachusetts Institute of Technology, 2009.

Yonghao Song, Xueyu Jia, Lie Yang, and Longhan Xie. Transformer-based spatial-temporal feature learning for EEG decoding. *arXiv preprint arXiv:2106.11170*, 2021.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. EEG Conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2022.

Christopher Wang, Vighnesh Subramaniam, Adam Uri Yaari, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. Brainbert: Self-supervised representation learning for intracranial recordings. In *International Conference on Learning Representations*, 2023.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBraMod: A criss-cross brain foundation model for EEG decoding. In *International Conference on Learning Representations (ICLR)*, 2025.

Di Wu, Siyuan Li, Jie Yang, and Mohamad Sawan. Neuro-BERT: Rethinking masked autoencoding for self-supervised neurological pretraining. *arXiv preprint arXiv:2204.12440*, 2022.

Chaoqi Yang, M. Brandon Westover, and Jimeng Sun. BIOT: Biosignal transformer for cross-data learning in the wild. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Zhizhang Yuan, Daoze Zhang, Junru Chen, Geifei Gu, and Yang Yang. Brant-2: Foundation model for brain signals. *arXiv preprint arXiv:2402.10251*, 2024.

Tongtian Yue, Shuning Xue, Xuange Gao, Yepeng Tang, Longteng Guo, Jie Jiang, and Jing Liu. EEGPT: Unleashing the potential of EEG generalist foundation model by autoregressive pre-training. *arXiv preprint arXiv:2410.19779*, 2024. doi: 10.48550/arXiv.2410.19779.

Daoze Zhang, Zhizhang Yuan, Yang Yang, Junru Chen, Jingjing Wang, and Yafeng Li. Brant: Foundation model for intracranial neural signal. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. ST-MoE: Designing stable and transferable sparse expert models. *arXiv preprint arXiv:2202.08906*, 2022.

Igor Zyma, Sergii Tukaev, Ivan Seleznov, Ken Kiyono, Anton Popov, Mariia Chernykh, and Oleksii Shpenkov. Electroencephalograms during mental arithmetic task performance. *Data*, 4(1):14, 2019.

A Related Work

EEG Foundation Models. Inspired by the success of foundation models in vision and language, recent EEG research has moved from task-specific models toward general-purpose EEG foundation models. Existing efforts mainly follow contrastive or reconstruction-based pre-training. BENDR (Kostas et al., 2021) demonstrated contrastive pre-training on large EEG corpora, while Brant-style models extended this direction across physiological signals (Zhang et al., 2023; Yuan et al., 2024). BIOT (Yang et al., 2023), LaBraM (Jiang et al., 2024b), CBraMod (Wang et al., 2025), and CodeBrain (Ma et al., 2025) use unified tokenization or reconstruction objectives. NeuroLM (Jiang et al., 2024a) and EEGPT (Yue et al., 2024) explore autoregressive EEG pre-training. TRACE follows this direction with past-only next-patch forecasting and Cross-Channel Temporal Routing FFNs (CTR-FFNs).

Mixture-of-Experts. The mixture-of-experts (MoE) architecture increases model capacity without proportionally increasing computation by sparsely activating a subset of experts for each input (Shazeer et al., 2017). Efficient routing strategies, such as Top-1 routing in Switch Transformer (Fedus et al., 2022) and Top-2 gating in GShard (Lepikhin et al., 2021), have made MoE widely used in modern Transformer models. However, standard MoE architectures are mostly designed for language-like token sequences and route each token independently. This token-wise routing is not well aligned with EEG signals, where channels at the same temporal step jointly describe an instantaneous brain state. Routing each patch independently may fragment this multi-channel coherence, while a uniform feed-forward block cannot adapt to the temporal non-stationarity of EEG dynamics. TRACE introduces a CTR-FFN for autoregressive EEG pre-training: TemporalFormer summarizes the cross-channel history, and a router assigns all channels at the same temporal step to the same specialized expert subset. This enables expert specialization over evolving temporal contexts while preserving multi-channel EEG coherence.

Switching neural dynamics. Recurrent switching linear dynamical systems (rSLDS) use a discrete latent state to select among dynamical regimes, with state transitions conditioned on evolving latent activity (Linderman et al., 2017). Extensions model changing interactions among neural populations (Glaser et al., 2020).

TR-MoE brings this switching-regime perspective to a nonlinear Transformer, where the router selects computational pathways from past context.

B Dataset Description

We use a diverse collection of public EEG datasets for TRACE pre-training and downstream evaluation. These datasets span large-scale clinical EEG, high-density healthy-population EEG, affective decoding, motor imagery, mental arithmetic, sleep staging, seizure detection, clinical event classification, and imagined speech. The collection covers heterogeneous recording domains, channel configurations, sampling rates, and temporal resolutions for learning and evaluating transferable EEG representations.

TUEG. The Temple University Hospital EEG Corpus (TUEG) is a large-scale public clinical EEG corpus collected in hospital settings (Obeid & Picone, 2016). It contains long-duration scalp EEG recordings from diverse clinical populations and covers routine, ambulatory, and continuous EEG studies. Following prior EEG foundation-model preprocessing protocols (Jiang et al., 2024b; Wang et al., 2025), we use the 19-channel international $10-20$ montage, apply band-pass filtering, notch filtering, resample the signals to 200 Hz, and segment recordings into 30-second non-overlapping windows. TUEG provides large-scale clinical EEG background activity under real-world acquisition conditions.

HBN. The Healthy Brain Network (HBN) dataset provides high-density EEG recordings from children and adolescents across multiple experimental paradigms (Alexander et al., 2017). Different from TUEG, which uses a 19-channel clinical montage, HBN uses a 128-channel HydroCel Geodesic Sensor Net and covers a healthy-population developmental cohort. This dataset introduces richer spatial coverage, different channel configurations, and non-clinical recording conditions. In our

preprocessing pipeline, HBN recordings are resampled to 200 Hz and segmented into 30-second windows while preserving the native 128-channel layout.

Emotion Recognition. We conduct emotion recognition experiments on FACED and SEED-V. FACED is a fine-grained affective EEG dataset containing 32-channel EEG recordings collected during affective video watching (Chen et al., 2023). It is formulated as a 9-class emotion classification task covering multiple positive, negative, and neutral affective states. We follow the standard cross-subject evaluation protocol, resample the EEG signals to 200 Hz, and segment trials into 10-second windows.

SEED-V is a 5-class EEG emotion recognition dataset with happy, sad, neutral, disgust, and fear categories (Liu et al., 2021). It contains 62-channel multi-session EEG recordings from 16 subjects, originally sampled at 1,000 Hz. Each subject participates in three sessions with 15 trials per session. We split the 15 trials in each session into training, validation, and test partitions of 5 trials each, resample the signals to 200 Hz, and segment them into 1-second windows.

Motor Imagery Classification. We evaluate motor imagery classification on PhysioNet-MI and SHU-MI. PhysioNet-MI contains 64-channel EEG recordings from 109 subjects performing motor execution and motor imagery tasks, originally sampled at 160 Hz (Schalk et al., 2004). We formulate it as a 4-class classification task, resample the recordings to 200 Hz, and segment the signals into 4-second windows. Following a subject-wise split, subjects $1-70$ are used for training, $71-89$ for validation, and $90-109$ for testing.

SHU-MI is a motor imagery dataset collected from 25 subjects using a 32-channel EEG cap, originally sampled at 250 Hz (Ma et al., 2022). Each subject performs five sessions of left-hand and right-hand motor imagery. We formulate it as a binary classification task, resample the recordings to 200 Hz, and segment them into 4-second non-overlapping windows. Following a subject-wise split, subjects $1-15$ are used for training, $16-20$ for validation, and $21-25$ for testing.

Mental Arithmetic. PhysioNet-MA contains EEG recordings from 36 subjects under two conditions: resting state and mental arithmetic (Zyma et al., 2019). The original recordings consist of 72 EDF files, corresponding to two recordings per subject, sampled at 500 Hz with 23 electrodes. We retain 20 standard $10-20$ channels, resample the signals to 200 Hz, and segment them into 5-second non-overlapping windows. Following a subject-wise split, subjects $1-28$ are used for training, $29-32$ for validation, and $33-36$ for testing.

Sleep Staging. ISRUC is used for sleep stage classification (Khalighi et al., 2016). It contains overnight polysomnographic recordings annotated according to standard sleep-stage categories: Wake, N1, N2, N3, and REM. Each EEG segment corresponds to a 30-second epoch. We adopt subject-wise splits used in prior EEG foundation-model benchmarks: for ISRUC-S1, which contains 100 subjects, we use 80 subjects for training, 10 for validation, and 10 for testing; for ISRUC-S3, which contains 10 subjects, we follow an $8:1:1$ subject-wise split.

Seizure Detection. CHB-MIT is a widely used pediatric scalp EEG dataset for seizure detection (Shoeb, 2009). It contains long-term EEG recordings from subjects with epilepsy collected during clinical monitoring. We formulate the task as binary classification between seizure and non-seizure EEG segments. Following common preprocessing protocols, we use 16 channels in the standard bipolar montage, resample recordings to 200 Hz, and segment them into 10-second non-overlapping windows. We adopt a subject-wise split with chb01–chb20 for training, chb21-chb22 for validation, and chb23-chb24 for testing, with chb12, chb13, and chb17 excluded due to insufficient or noisy recordings.

EEG Event Classification. TUEV is used for clinical EEG event type classification (Obeid & Picone, 2016). The annotations cover six event categories: spike and sharp wave (SPSW), generalized periodic epileptiform discharges (GPED), periodic lateralized epileptiform discharges (PLED), eye movement (EYEM), artifact (ARTF), and background activity (BCKG). Following standard preprocessing, we select 16 bipolar montage channels, resample the recordings to 200 Hz, and segment them into 5-second windows. We follow the official train/test split and further divide the training subjects into an $8:2$ train/validation partition.

Imagined Speech Classification. BCI Competition 2020 Track 3 (BCIC2020-3) is used for imagined speech classification (Jeong et al., 2022). The dataset contains 64-channel EEG recordings collected while subjects silently imagine speaking five short words or phrases: "hello", "help me", "stop", "thank you", and "yes". It is formulated as a 5-class classification task. We resample the recordings to 200 Hz and use the train, validation, and test splits provided by the competition organizers.

C Baselines and Metrics Description

C.1 Baselines

Here, we introduce the details of the baselines used for performance evaluation.

EEGNet (Lawhern et al., 2018) is a lightweight convolutional neural network designed for EEG decoding. It uses temporal convolution and depthwise spatial convolution to learn compact EEG representations.

EEGConformer (Song et al., 2022) is an EEG model that combines CNN-based local feature extraction with Transformer-based self-attention to model global temporal dependencies.

CNN-Transformer (Peh et al., 2022) is a hybrid architecture that uses CNN layers to extract local EEG features and a Transformer encoder to capture long-range temporal dependencies.

FFCL (Li et al., 2022) is an EEG classification model that combines convolutional and recurrent modules, where CNNs capture local spatial-temporal patterns and recurrent layers model sequential dynamics.

ST-Transformer (Song et al., 2021) is a Transformer-based EEG model that learns spatial and temporal dependencies through attention mechanisms.

BIOT (Yang et al., 2023) is a biosignal foundation model that tokenizes heterogeneous physiological signals into unified token sequences and learns transferable representations through cross-dataset pre-training.

LaBraM (Jiang et al., 2024b) is a large EEG foundation model that learns neural tokens with a vector-quantized tokenizer and performs masked EEG token prediction for pre-training.

CBraMod (Wang et al., 2025) is an EEG foundation model that uses criss-cross spatial-temporal attention and reconstruction-based pre-training to learn transferable EEG representations.

CodeBrain (Ma et al., 2025) is an EEG foundation model that learns discrete EEG representations through temporal-frequency tokenization, aiming to capture structured EEG patterns for downstream transfer.

C.2 Metrics

In this section, we introduce the evaluation metrics adopted in this paper. Consistent with prior EEG foundation model evaluations (Jiang et al., 2024b; Wang et al., 2025; Ma et al., 2025), we report different metrics according to the task type.

Balanced Accuracy measures the average recall across all classes and is less affected by class imbalance than standard accuracy. It is reported for both binary and multi-class classification tasks.

Kappa quantifies the agreement between predicted labels and ground-truth labels after correcting for chance agreement. This metric is particularly useful for multi-class EEG classification benchmarks with imbalanced class distributions.

Weighted F1 computes the class-wise F1 score and averages it according to class support. It provides a precision-recall summary for multi-class tasks where different classes may have different numbers of samples.

AUC-PR denotes the area under the precision-recall curve. It is especially informative for imbalanced binary detection tasks, where the positive class is rare.

AUROC denotes the area under the receiver operating characteristic curve and evaluates threshold-independent discrimination between positive and negative classes.

D More Details for Experimental Settings on Pre-Training

We next provide additional details on the pre-training corpus, model configuration, and optimization setup used for TRACE pre-training.

Pre-training corpus. We construct the pre-training corpus from six EEG sources: two large-scale sources, TUEG and HBN, together with four task-related datasets. TUEG provides large-scale 19-channel clinical EEG, while HBN provides 128-channel high-density EEG from a healthy-population cohort. The task-related sources further cover seizure monitoring, motor imagery, and mental arithmetic recordings. For datasets that also appear in downstream evaluation, only the training-split subjects are included in the pre-training corpus; validation and test subjects are strictly held out from pre-training. The sources and preprocessing details are summarized in Table 5. The sample counts in the table indicate the total number of available preprocessed segments for each source; for downstream-overlapping datasets, pre-training uses only the training split.

Table 5: Pre-training corpus and preprocessing details.

Dataset	EEG Domain	# Channels	# Samples	Window
TUEG	Clinical EEG	19	1.1M	30s
HBN	Healthy-population EEG	128	132K	30s
CHB-MIT	Seizure detection	16	326,993	10s
PhysioNet-MI	Motor imagery	64	9,837	4s
SHU-MI	Motor imagery	32	11,988	4s
PhysioNet-MA	Mental arithmetic	20	1,343	5s

Model configuration. The architecture hyperparameters of TRACE are summarized in Table 6. The main configuration uses $N{=}64$ specialized experts plus one shared expert and activates $K{=}8$ experts at each temporal step.

Optimization setup. The pre-training optimization hyperparameters are summarized in Table 7.

Table 7: Hyperparameters of pre-training.

Hyperparameters	Values
Schedule	
Epochs	30
Batch size	256
Mixed precision	bf16
Optimizer	
Linear params (dim ≥ 2)	Muon, lr=1e$-$3
Embeddings / LN / biases	AdamW, lr=2e$-$5
Weight decay	1e$-$2
LR schedule	cosine decay
Warmup steps	2,000
Gradient clipping (max norm)	1.0
Forecasting objective	
Horizons $\mathcal{H}$	$\{1, 2, 4\}$
Huber threshold δ	1.0
Autoregressive loss $\mathcal{L}_{\mathrm{AR}}$	unit weight
Auxiliary losses	
Expert-balancing weight λ_{aux}	1e$-$2

Table 6: Hyperparameters of TRACE architecture.

Hyperparameters	Values
Patch embedding	
Temporal kernel set $\lvert\mathcal{Q}_{\text{temp}}\rvert$	3
Temporal kernels $\mathcal{Q}_{\text{temp}}$	$\{25, 49, 99\}$
CNN stride	25
Filters per branch	8
Spectral branch	FFT log-amplitude
Spectral frequency bins	101
Fusion	sigmoid-gated
Patch embedding dimension d	200
Multi-Scale Context-aware Channel Encoding (MS-CCE)	
Channel-kernel branches R	3 depth-wise Conv2d
Channel kernels $\{q_r\}_{r=1}^{R}$	$\{5, 11, 19\}$
Time kernel	1
Transformer backbone	
Layers L	12
Hidden size d	200
FFN intermediate size	800
Attention heads	8
KV heads	8
Positional encoding	RoPE (θ=10,000)
Activation	SiLU
Normalization	RMSNorm (ϵ=1e$-$6)
Cross-Channel Temporal Routing FFN (CTR-FFN)	
Specialized experts N	64
Activated specialized experts K	8
Shared expert Expert_{N+1}	1
Routing context	TemporalFormer
Expert selection	Top-K router
TemporalFormer query tokens m	4
TemporalFormer attention heads	4

E FINE-TUNING SETTINGS ON DOWNSTREAM TASKS

Downstream Tasks. To evaluate the transferability of TRACE, we fine-tune the pretrained model on eight downstream EEG datasets covering emotion recognition, motor imagery classification, sleep staging, seizure detection, imagined speech classification, and EEG event classification. For all downstream datasets, EEG signals are resampled to 200 Hz, and each EEG patch is set to 1s, corresponding to 200 data points, following the same tokenization setting used in pre-training. Details of the downstream tasks and datasets are summarized in Table 8.

Table 8: Overview of downstream EEG tasks and datasets.

Downstream Tasks	Dataset	# Channels	# Samples	Window	Label
Emotion Recognition	FACED	32	10,332	10s	9-class
	SEED-V	62	117,744	1s	5-class
Motor Imagery Classification	PhysioNet-MI	64	9,837	4s	4-class
	SHU-MI	32	11,988	4s	2-class
Sleep Staging	ISRUC	6	89,240	30s	5-class
Seizure Detection	CHB-MIT	16	326,993	10s	2-class
Imagined Speech Classification	BCIC2020-3	64	6,000	3s	5-class
Event Type Classification	TUEV	16	112,491	5s	6-class

Fine-tuning setup. For each downstream task, we attach a lightweight three-layer MLP classification head on top of the pretrained TRACE encoder and fine-tune the full model end-to-end. The fine-tuning hyperparameters are selected for each dataset based on the task type and dataset scale. Table 9 lists the learning rate, weight decay, dropout rate, and batch size used for each downstream benchmark.

Table 9: Fine-tuning hyperparameters for downstream tasks.

Dataset	Learning rate	Weight decay	Dropout	Batch size
FACED	2e−3	5e−2	0.05	16
SEED-V	2e−3	5e−2	0.10	256
PhysioNet-MI	5e−3	5e−2	0.10	64
SHU-MI	1e−3	5e−3	0.10	32
ISRUC	5e−3	5e−2	0.10	16
CHB-MIT	2e−3	5e−3	0.10	64
BCIC2020-3	5e−3	5e−2	0.10	32
TUEV	2e−3	5e−2	0.10	32

F MORE RESULTS ON OTHER DOWNSTREAM BCI TASKS

We provide additional downstream results on SHU-MI and TUEV to further evaluate the transfer ability of TRACE across motor imagery and EEG event classification tasks.

SHU-MI. As shown in Table 10, CodeBrain achieves the best overall performance on SHU-MI, reaching 0.6431 Balanced Accuracy, 0.7166 AUC-PR, and 0.7124 AUROC. TRACE remains competitive among EEG foundation models, obtaining 0.6329 Balanced Accuracy, 0.6978 AUC-PR, and 0.6988 AUROC. Notably, its AUROC matches CBraMod (0.6988), while its Balanced Accuracy remains higher than BIOT and LaBraM-Base. Compared with task-specific baselines such as EEGConformer and ST-Transformer, TRACE also shows a clear improvement in Balanced Accuracy and AUC-PR. These results suggest that although TRACE does not achieve the best result on SHU-MI, it maintains strong transfer performance on motor imagery classification.

Table 10: Performance comparison on the SHU-MI (2-Class) dataset. The best and second-best results are highlighted in **red** and blue, respectively.

Method	Balanced Acc	AUC-PR	AUROC
EEGNet	0.5889 ± 0.0177	0.6311 ± 0.0142	0.6283 ± 0.0152
EEGConformer	0.5900 ± 0.0107	0.6370 ± 0.0093	0.6351 ± 0.0101
CNN-Transformer	0.5975 ± 0.0169	0.6412 ± 0.0076	0.6323 ± 0.0082
FFCL	0.5692 ± 0.0252	0.5943 ± 0.0172	0.6014 ± 0.0168
ST-Transformer	0.5992 ± 0.0206	0.6394 ± 0.0122	0.6431 ± 0.0111
BIOT	0.6179 ± 0.0183	0.6770 ± 0.0119	0.6609 ± 0.0127
LaBraM-Base	0.6166 ± 0.0192	0.6761 ± 0.0083	0.6604 ± 0.0091
CBraMod	0.6370 ± 0.0151	0.7139 ± 0.0088	0.6988 ± 0.0068
CodeBrain	**0.6431** ± 0.0066	**0.7166** ± 0.0106	**0.7124** ± 0.0050
TRACE (Ours)	0.6329 ± 0.0061	0.6978 ± 0.0059	0.6988 ± 0.0048

TUEV. As shown in Table 11, TUEV is a challenging six-class clinical EEG event classification benchmark. TRACE obtains 0.6782 Balanced Accuracy, 0.7070 Kappa, and 0.8501 Weighted F1, achieving the best result on all three metrics. It substantially outperforms task-specific baselines such as EEGConformer and ST-Transformer, and also improves over BIOT across all three metrics. Compared with BIOT, TRACE increases Balanced Accuracy by 15.01, Kappa by 17.97, and Weighted F1 by 10.09 percentage points. These results indicate that TRACE transfers effectively to clinical event classification.

Table 11: Performance comparison on the TUEV (6-Class) dataset. The best and second-best results are highlighted in **red** and blue, respectively.

Method	Balanced Acc	Kappa	Weighted F1
EEGNet	0.3876 ± 0.0143	0.3577 ± 0.0155	0.6539 ± 0.0120
EEGConformer	0.4074 ± 0.0164	0.3967 ± 0.0195	0.6983 ± 0.0152
CNN-Transformer	0.4087 ± 0.0161	0.3815 ± 0.0134	0.6854 ± 0.0293
FFCL	0.3979 ± 0.0104	0.3732 ± 0.0188	0.6783 ± 0.0120
ST-Transformer	0.3984 ± 0.0228	0.3765 ± 0.0306	0.6823 ± 0.0190
BIOT	0.5281 ± 0.0225	0.5273 ± 0.0249	0.7492 ± 0.0082
LaBraM-Base	0.6409 ± 0.0065	0.6637 ± 0.0093	0.8312 ± 0.0052
CBraMod	0.6671 ± 0.0107	0.6772 ± 0.0096	0.8342 ± 0.0064
CodeBrain	0.6428 ± 0.0062	0.6912 ± 0.0101	0.8362 ± 0.0048
TRACE (Ours)	**0.6782** ± 0.0036	**0.7070** ± 0.0244	**0.8501** ± 0.0114

G IMPLEMENTATION DETAILS

G.1 CAUSAL SPATIAL-TEMPORAL ATTENTION

This section provides the detailed implementation of the causal spatial–temporal attention (CSTA) module used in each TR-MoE block. Let the input to the l-th block be $\mathbf{H}^l \in \mathbb{R}^{C\times n\times d}$, where C is the number of EEG channels, n is the number of temporal patches, and d is the hidden dimension. CSTA factorizes full attention over the Cn patch tokens into two structured attention branches: a spatial branch over channels at the same temporal step and a causal temporal branch over past patches within each channel. This design captures both instantaneous cross-channel coupling and within-channel temporal dynamics while preserving the autoregressive constraint.

Spatial attention branch. For each temporal step j, we collect the channel tokens $\mathbf{H}^l_{:,j} \in \mathbb{R}^{C\times d}$ and apply multi-head self-attention along the channel axis:

$$\mathbf{S}^l_{:,j} = \mathrm{MHA}_{\mathrm{sp}}\left(\mathbf{H}^l_{:,j}, \mathbf{H}^l_{:,j}, \mathbf{H}^l_{:,j}\right). \tag{19}$$

The attention weights are computed only among channels that belong to the same temporal patch index j. Therefore, the spatial branch models global electrode interactions at each temporal step without using information from any future temporal patch. In practice, this branch is implemented by reshaping the tensor from (B, C, n, d) to (Bn, C, d) and applying standard multi-head attention with sequence length C.

Causal temporal attention branch. For each channel i, we collect its temporal sequence $\mathbf{H}^l_{i,:} \in \mathbb{R}^{n\times d}$ and apply multi-head self-attention along the patch axis with a lower-triangular causal mask:

$$\mathbf{T}^l_{i,:} = \mathrm{MHA}_{\mathrm{time}}\left(\mathbf{H}^l_{i,:}, \mathbf{H}^l_{i,:}, \mathbf{H}^l_{i,:}; \mathbf{M}_{\mathrm{causal}}\right), \tag{20}$$

where $\mathbf{M}_{\mathrm{causal}}(j, j') = 0$ if $j' \le j$ and $-\infty$ otherwise. Thus, the output token at temporal step j can attend only to patches $\{1, \ldots, j\}$ from the same channel. This branch is implemented by reshaping (B, C, n, d) to (BC, n, d) and applying standard masked multi-head attention with sequence length n.

Branch fusion. The two attention outputs are reshaped back to (B, C, n, d) and fused with an output projection:

$$\mathrm{CSTA}(\mathbf{H}^l)_{i,j} = \mathbf{W}_{\mathrm{o}}\left[\mathbf{S}^l_{i,j}; \mathbf{T}^l_{i,j}\right], \tag{21}$$

where $[\cdot;\cdot]$ denotes feature concatenation and $\mathbf{W}_{\mathrm{o}} \in \mathbb{R}^{2d\times d}$ projects the fused representation back to the model dimension. The projected output is then used in the pre-norm residual update in the TR-MoE block, as defined in Section 2.4.

CSTA is causal because the spatial branch only mixes channels at the same temporal step and the temporal branch uses a causal mask along the patch axis. Compared with full attention over all Cn tokens, which has complexity $O(C^2n^2)$, the factorized implementation has complexity $O(nC^2 + Cn^2)$. This is important for EEG pre-training because both the channel count and sequence length can vary substantially across datasets.

H More Experimental Results

H.1 Effect of Multi-Horizon Forecasting

We examine the autoregressive forecasting objective by comparing $\mathcal{H} = \{1\}$, $\{1, 2\}$, and $\{1, 2, 4\}$ across FACED, SEED-V, and PhysioNet-MI. On FACED, we additionally test the single-horizon settings $\{2\}$, $\{4\}$, and $\{8\}$.

Table 12: Forecasting-horizon ablation on FACED. Results are mean ± standard deviation over five seeds; the best mean in each metric is highlighted in **bold**.

Horizon Set	Balanced Acc	Kappa	Weighted F1
$\mathcal{H} = \{1\}$	62.14 ± 0.35	57.25 ± 0.33	62.47 ± 0.36
$\mathcal{H} = \{2\}$	62.20 ± 0.64	57.35 ± 0.69	62.62 ± 0.67
$\mathcal{H} = \{4\}$	62.58 ± 0.63	57.77 ± 0.65	63.00 ± 0.51
$\mathcal{H} = \{8\}$	61.89 ± 0.78	56.98 ± 0.85	62.35 ± 0.63
$\mathcal{H} = \{1, 2\}$	61.31 ± 1.39	56.25 ± 1.63	61.57 ± 1.44
$\mathcal{H} = \{1, 2, 4\}$	**62.83** ± 0.59	**58.07** ± 0.69	**63.17** ± 0.58

Table 13: Multi-horizon forecasting on SEED-V and PhysioNet-MI, with the three-dataset average including FACED from Table 12. Results are mean ± standard deviation over five seeds; the best result in each block is highlighted in **bold**.

Dataset	Horizon Set	Balanced Acc (%)	Kappa (%)	Weighted F1 (%)
SEED-V	$\mathcal{H} = \{1\}$	**40.95** ± 0.46	**26.51** ± 0.76	**41.64** ± 0.65
	$\mathcal{H} = \{1, 2\}$	40.73 ± 0.51	26.16 ± 0.71	41.37 ± 0.55
	$\mathcal{H} = \{1, 2, 4\}$	40.63 ± 0.33	25.94 ± 0.42	41.19 ± 0.35
PhysioNet-MI	$\mathcal{H} = \{1\}$	63.07 ± 1.45	50.75 ± 1.93	63.28 ± 1.41
	$\mathcal{H} = \{1, 2\}$	63.49 ± 0.67	51.31 ± 0.89	63.71 ± 0.76
	$\mathcal{H} = \{1, 2, 4\}$	**63.62** ± 0.30	**51.49** ± 0.40	**63.90** ± 0.25
Average	$\mathcal{H} = \{1\}$	55.39 ± 0.75	44.84 ± 1.01	55.80 ± 0.81
	$\mathcal{H} = \{1, 2\}$	55.18 ± 0.86	44.57 ± 1.08	55.55 ± 0.92
	$\mathcal{H} = \{1, 2, 4\}$	**55.69** ± 0.41	**45.17** ± 0.50	**56.09** ± 0.39

As shown in Table 13, using $\mathcal{H} = \{1, 2, 4\}$ gives the best average performance across the three downstream datasets. Compared with the single-step objective $\mathcal{H} = \{1\}$, the multi-horizon objective improves the average Balanced Accuracy from 55.39% to 55.69%, Kappa from 44.84% to 45.17%, and Weighted F1 from 55.80% to 56.09%. The gain is most evident on FACED and PhysioNet-MI. Table 12 shows that $\mathcal{H} = \{1, 2, 4\}$ has the highest mean on FACED among all six tested settings, improving Balanced Accuracy by 0.69 points and Weighted F1 by 0.70 points over $\{1\}$. The longer single horizon $\{8\}$ does not improve the FACED result, and adding horizons is not monotonically beneficial: $\{1, 2\}$ is below $\{1\}$. On PhysioNet-MI, $\{1, 2, 4\}$ also achieves the best result across all metrics, with a smaller standard deviation than the single-step setting. SEED-V shows a different pattern, where the single-step horizon performs best. This suggests that the optimal temporal prediction scale can vary across downstream tasks, while combining multiple horizons provides the strongest average transfer among the three common settings.

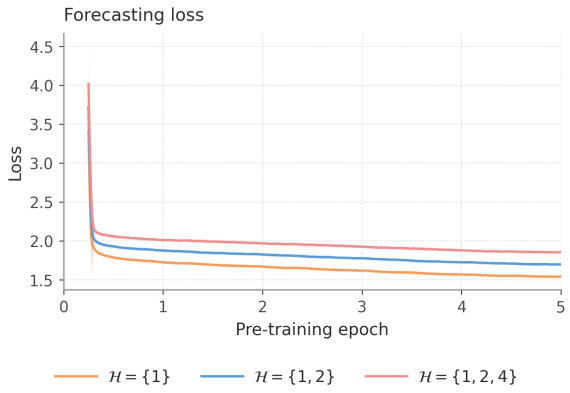


Figure 3: Pre-training loss dynamics under different horizon sets.

Figure 3 further compares the pre-training dynamics under different horizon sets. Adding longer horizons increases the forecasting loss, which is expected because farther future patches are harder to predict from observed history. Importantly, this higher forecasting loss does not indicate weaker downstream transfer: $\mathcal{H} = \{1, 2, 4\}$ shows the largest forecasting loss but achieves the best average performance in Table 13. This suggests that multi-horizon forecasting creates a more challenging but more informative temporal prediction task.

H.2 Effect of Multi-source Pre-training

We evaluate the impact of pre-training data composition by comparing four source combinations, ranging from single-source pre-training to the full multi-source corpus. This experiment aims to

determine whether source diversity, channel configuration, and task coverage improve downstream transfer.

As shown in Table 14, the full multi-source setting achieves the best performance on SEED-V across B-Acc, Kappa, and F1-W. This highlights two key findings. First, pre-training source domain matters. HBN-only slightly outperforms TUEG-only despite using fewer samples, suggesting that high-density 128-channel healthy-population EEG provides complementary information to the 19-channel clinical EEG in TUEG. Second, task-oriented sources are most effective when combined with large-scale EEG corpora. The 4ds-only setting performs the worst, while the full setting achieves the strongest result after adding the task-oriented datasets to TUEG and HBN. These results suggest that TRACE benefits from pre-training on EEG data that spans different channel counts, acquisition conditions, and task domains, rather than relying on a single source or merely increasing sample count.

Table 14: Analysis of pre-training data composition for SEED-V.

Pre-training Sources						Downstream on SEED-V		
TUEG	**HBN**	**Phys-MI**	**Phys-MA**	**SHU-MI**	**CHB-MIT**	**Balanced Acc (%)**	**Kappa (%)**	**Weighted F1 (%)**
✓	✗	✗	✗	✗	✗	41.10 ± 0.08	26.57 ± 0.14	41.69 ± 0.15
✗	✓	✗	✗	✗	✗	41.24 ± 0.55	26.84 ± 0.73	42.06 ± 0.60
✗	✗	✓	✓	✓	✓	40.81 ± 0.61	26.33 ± 0.52	41.42 ± 0.51
✓	✓	✓	✓	✓	✓	**41.86** ± 0.32	**27.59** ± 0.30	**42.53** ± 0.24

As shown in Table 14, the choice of pre-training data has a clear impact on downstream transfer. The mixed-source corpus provides the strongest overall transfer, suggesting that diverse EEG sources help TRACE learn more general temporal representations. We also observe task-dependent transfer patterns, where sources closer to the target task tend to provide stronger gains on the corresponding downstream benchmark.

H.3 EFFECT OF CTR-FFN

We investigate the essential design of TRACE, *i.e.*, the CTR-FFN. We evaluate the necessity of this core architectural design by comparing the full TRACE model against three variants: 1) a Non-MoE baseline where the CTR-FFN is replaced with a standard FFN, 2) a Vanilla MoE version where each patch selects its own experts independently via a token-wise router, and 3) a Mean Temporal variant where the routing context at each temporal step is obtained by mean-pooling channel embeddings instead of using TemporalFormer. All experiments are conducted on three datasets: FACED, PhysioNet-MI, and SEED-V. As shown in Table 15, our proposed CTR-FFN consistently outperforms all baseline variants across the evaluated datasets.

Table 15: Effect of CTR-FFN.

Dataset	Routing Design	Balanced Acc (%)	Kappa (%)	Weighted F1 (%)
FACED 9-Class	Non-MoE	62.44 ± 0.68	57.63 ± 1.16	62.91 ± 1.06
	Vanilla MoE	62.39 ± 1.04	57.55 ± 0.79	62.66 ± 0.65
	Mean Temporal	63.45 ± 0.53	58.76 ± 0.61	63.75 ± 0.59
	TRACE (Ours)	**63.52** ± 0.37	**58.83** ± 0.42	**63.90** ± 0.43
PhysioNet-MI 4-Class	Non-MoE	62.28 ± 1.45	49.70 ± 1.93	62.53 ± 1.43
	Vanilla MoE	62.76 ± 0.70	50.34 ± 0.93	63.06 ± 0.65
	Mean Temporal	63.00 ± 0.84	50.66 ± 1.13	63.30 ± 0.87
	TRACE (Ours)	**63.42** ± 0.90	**51.22** ± 1.20	**63.65** ± 0.96
SEED-V 5-Class	Non-MoE	39.82 ± 0.65	24.86 ± 0.96	40.27 ± 0.79
	Vanilla MoE	38.43 ± 0.87	22.91 ± 1.26	38.75 ± 1.03
	Mean Temporal	39.42 ± 0.50	24.18 ± 0.59	39.82 ± 0.46
	TRACE (Ours)	**40.32** ± 0.47	**25.48** ± 0.57	**40.76** ± 0.44
Average	Non-MoE	54.85 ± 0.93	44.06 ± 1.35	55.24 ± 1.09
	Vanilla MoE	54.53 ± 0.87	43.60 ± 0.99	54.82 ± 0.78
	Mean Temporal	55.29 ± 0.62	44.53 ± 0.78	55.62 ± 0.64
	TRACE (Ours)	**55.75** ± 0.58	**45.18** ± 0.73	**56.10** ± 0.61

On average, TRACE achieves the best balanced accuracy, surpassing the Non-MoE baseline by +0.90 percentage points, the Vanilla MoE by +1.22 percentage points, and the Mean Temporal variant by +0.46 percentage points. This leads to two observations. First, replacing the FFN with a generic token-wise MoE does not necessarily improve performance, suggesting that expert routing strategy is important for EEG modeling. Second, TemporalFormer-based temporal routing further improves over both Vanilla MoE and mean-based temporal routing, indicating that routing all channels at the same temporal step with a learned cross-channel temporal context better matches the multi-channel temporal structure of EEG signals. Overall, these results show that the effectiveness of TRACE comes from both sparse expert specialization and temporal routing.

H.4 MATCHED AUTOREGRESSIVE FORECASTING AND MASKED RECONSTRUCTION

To isolate the pre-training objective, we compare TRACE-AR with a matched masked-reconstruction variant.

We fix the parameterized backbone, patch encoder, CTR-FFN, expert configuration, pre-training corpus, optimization budget, and downstream fine-tuning protocol; the attention mask and input corruption differ according to the objective.

TRACE-AR uses multi-horizon future-patch forecasting. TRACE-Masked instead uses bidirectional attention to reconstruct randomly masked temporal patches. For TRACE-Masked, each 1-second temporal patch is independently masked with probability 0.5, and the mask is resampled at every update. The attention constraint and input corruption follow the respective objective.

Table 16: Matched balanced-accuracy comparison between autoregressive forecasting and masked reconstruction. The architecture, corpus, optimization budget, and downstream protocol are fixed.

Dataset	TRACE-AR (%)	TRACE-Masked (%)	AR − Masked (pp)
SEED-V	**41.86**	40.63	+1.23
PhysioNet-MI	63.42	**64.47**	−1.05
BCIC2020-3	**63.15**	54.59	+8.56
CHB-MIT	**74.53**	57.80	+16.73
ISRUC	**79.97**	75.30	+4.67
Macro average	**64.59**	58.56	+6.03

TRACE-AR performs better on four of five datasets. Macro-average balanced accuracy rises from 58.56% to 64.59%, a gain of 6.03 percentage points.

The largest gains occur on BCIC2020-3 ($+8.56$), CHB-MIT ($+16.73$), and ISRUC ($+4.67$); SEED-V also improves by 1.23 points.

On PhysioNet-MI, masked reconstruction is 1.05 percentage points higher than AR forecasting.

H.5 MODEL SCALING UNDER MATCHED ACTIVATED COMPUTATION

We vary the number of specialized experts N and the number of activated experts K while keeping the activated expert width, and hence the dominant expert FLOPs, fixed. Figure 4 reports downstream fine-tuning results over five seeds.

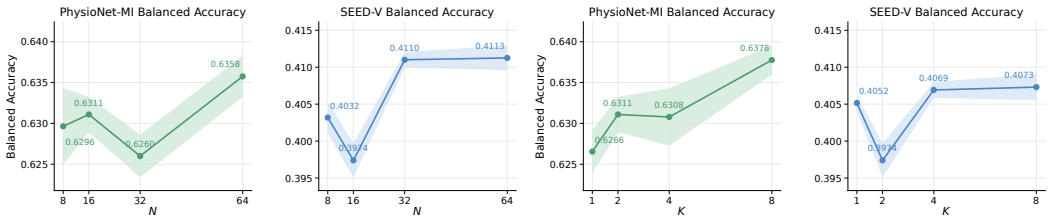
In Figure 4(a), we fix $K{=}2$ and the per-expert width at 800, then vary $N \in \{8, 16, 32, 64\}$. The largest pool gives the highest mean on both datasets, although performance is not monotonic in N.

On SEED-V, the mean largely plateaus from $N{=}32$ to $N{=}64$ (0.4110 to 0.4113). On PhysioNet-MI, the best mean is also at $N{=}64$ (0.6358), but the intermediate settings fluctuate.

In Figure 4(b), we fix $N{=}16$ and vary $K \in \{1, 2, 4, 8\}$. The corresponding per-expert widths are $\{1600, 800, 400, 200\}$, keeping the total activated expert width at $Kd_{\text{ff}}{=}1600$.

Activating eight narrower experts gives the highest mean on both datasets. The differences on SEED-V are small apart from the lower $K{=}2$ result, so the sweep does not support a monotonic scaling claim.

Together, the results show that how a fixed activated-compute budget is partitioned across available and selected experts can affect downstream transfer.



(a) Number of total experts N with top-$K = 2$ fixed **(b) Number of activated experts K with $N = 16$ fixed**

Figure 4: **Expert scaling at approximately matched FLOPs.** Fine-tuned balanced accuracy on PhysioNet-MI and SEED-V when varying (a) N with $K{=}2$, and (b) K with $N{=}16$. The activated expert width is fixed at $Kd_{\text{ff}}{=}1600$. Lines show means over five fine-tuning seeds; shading shows $\pm$SEM.

H.6 CONTRIBUTIONS OF SHARED AND ROUTED EXPERT PATHWAYS

Appendix H.3 provides a dense no-routing control: the Non-MoE variant replaces CTR-FFN with a dense FFN and has no routed experts. This changes the architecture and whole-model FLOPs, as reported in Appendix H.11; it does not isolate the shared expert by itself. The full model improves average balanced accuracy by 0.90 points over this control.

To isolate the shared pathway within the MoE architecture, we remove the shared expert while retaining the router and specialized experts. All other settings are kept fixed.

Table 17: Effect of the shared expert.

Dataset	Variant	Balanced Acc (%)	Kappa (%)	Weighted F1 (%)
FACED	*w/o* Shared Expert	62.44 ± 0.14	57.60 ± 0.13	62.77 ± 0.22
	w/ Shared Expert	**63.52** ± 0.37	**58.83** ± 0.42	**63.90** ± 0.43
BCIC2020-3	*w/o* Shared Expert	61.97 ± 0.78	52.46 ± 0.98	61.97 ± 0.82
	w/ Shared Expert	**63.15** ± 0.55	**53.93** ± 0.69	**63.16** ± 0.57

As shown in Table 17, including a shared expert provides consistent improvements on both FACED and BCIC2020-3 across all metrics. On average, this leads to an improvement of approximately $+1.13$ percentage points in Balanced Accuracy, $+1.35$ percentage points in Kappa, and $+1.16$ percentage points in Weighted F1. The most notable gain is observed on BCIC2020-3, where Kappa increases by nearly 1.5 percentage points. We hypothesize that the shared expert learns to model fundamental EEG patterns that are broadly present across temporal states and channels, such as common signal properties, baseline neural activity, or general rhythmic components. The shared expert can provide a common representation pathway for general EEG patterns, while the specialized experts capture more fine-grained temporal variations. This suggests that the shared expert provides a common representation pathway for general EEG patterns, complementing the specialized experts that capture more fine-grained temporal variations.

H.7 Effect of Patch Length

Patch length t determines how raw EEG signals are grouped into temporal patches. In our setting, the signals are resampled to 200 Hz, so patch lengths of 50, 100, 200, and 400 samples correspond to temporal windows of 0.25s, 0.5s, 1s, and 2s, respectively. Each patch is encoded into one token by the time-frequency patch encoder. This token is then used as the prediction unit in multi-horizon autoregressive forecasting and as the temporal unit for expert routing. Therefore, patch length t directly controls the temporal granularity of both forecasting and expert selection in TRACE.

Table 18: Ablation on patch length t.

Dataset	Patch length t	Balanced Acc (%)	Kappa (%)	Weighted F1 (%)
FACED	50	58.36 ± 1.70	52.80 ± 1.88	58.22 ± 1.64
	100	58.74 ± 1.35	53.23 ± 1.49	58.56 ± 1.29
	200	**63.52** ± 0.37	**58.83** ± 0.42	**63.90** ± 0.43
	400	58.82 ± 0.73	53.50 ± 0.83	59.13 ± 0.64
ISRUC	50	78.77 ± 0.15	75.37 ± 0.72	80.65 ± 0.55
	100	78.26 ± 1.12	73.44 ± 2.86	78.87 ± 2.63
	200	**79.41** ± 0.58	**75.41** ± 0.29	**80.77** ± 0.18
	400	78.77 ± 0.56	75.06 ± 1.12	80.57 ± 0.95

We compare four patch lengths t while keeping the model architecture, pre-training corpus, and downstream fine-tuning protocol fixed. As shown in Table 18, the 1s patch achieves the best performance on both FACED and ISRUC across all three metrics. The effect is especially clear on FACED, where the 1s patch improves Balanced Accuracy by nearly 4.7 percentage points over the neighboring settings. ISRUC shows a smaller but consistent gain, suggesting that the 1s patch also transfers well to longer sleep-stage epochs.

Compatibility with downstream windows. A 1s patch naturally matches common EEG window lengths used in downstream benchmarks. For example, SEED-V uses 1s clips, PhysioNet-MI uses 4s windows, and ISRUC uses 30s epochs. With a 1s patch, these inputs can be represented as an integer number of tokens without unnecessary padding or truncation. This choice is also consistent with recent EEG foundation models, where 1s EEG segments are commonly used as the basic modeling unit (Jiang et al., 2024b; Wang et al., 2025; Ma et al., 2025). Smaller patches increase the number of

temporal tokens and the routing cost of the CTR-FFN, while larger patches can be less suitable for short-window datasets.

Preserving waveform structure. A 1s patch also better preserves physiologically meaningful EEG patterns. Many EEG events occur at sub-second to second-level scales, such as K-complexes around 1s, sleep spindles lasting 0.5–2s, and event-related potentials such as P300 around 0.3–0.6s. Very small patches may split these patterns across multiple tokens, while very large patches may merge fast transients into a coarse representation. The 1s setting therefore provides a practical balance between local waveform preservation, temporal abstraction, and computational efficiency.

H.8 Embedding Design Ablation

We further ablate the embedding design of TRACE. All variants keep the backbone, pre-training objective, and downstream fine-tuning protocol fixed, and only modify the patch encoder. The results are summarized in Tables 19–21.

Branch design. We first examine whether both the time-domain CNN branch and the spectral FFT branch are necessary. Removing the time branch causes a substantial drop, decreasing Balanced Accuracy by 23.25 percentage points on FACED and 6.03 points on BCIC2020-3. In contrast, removing the FFT branch only mildly affects performance, with drops of 0.40 and 2.11 points on the two datasets. This suggests that the time-domain CNN branch carries the dominant discriminative signal, while the FFT branch provides complementary spectral cues for tasks with stronger frequency-band structure.

Table 19: Analysis of time-frequency branch design.

Branch Design	FACED		BCIC2020-3	
	Balanced Acc	**Kappa**	**Balanced Acc**	**Kappa**
Full dual-branch	**0.6352** $\pm$ 0.0037	**0.5883**	**0.6315** $\pm$ 0.0055	**0.5393**
Time-only (*w/o* FFT)	0.6312 $\pm$ 0.0083	0.5835	0.6104 $\pm$ 0.0151	0.5130
FFT-only (*w/o* Time)	0.4027 $\pm$ 0.0129	0.3250	0.5712 $\pm$ 0.0214	0.4640

Fusion strategy. We examine whether the learnable sigmoid gate is necessary for fusing the time-domain and FFT-based branches. Compared with the default sigmoid-gated fusion, fixed summation decreases Balanced Accuracy by 1.36 percentage points on FACED and 0.99 points on BCIC2020-3. It also leads to a larger standard deviation on BCIC2020-3. This suggests that the two branches should not be combined with a fixed equal-weight rule; the learnable gate provides an adaptive fusion mechanism that dynamically balances temporal and spectral information.

Table 20: Analysis of branch fusion strategy.

Fusion Design	FACED		BCIC2020-3	
	Balanced Acc	**Kappa**	**Balanced Acc**	**Kappa**
Sigmoid-gated fusion	**0.6352** $\pm$ 0.0037	**0.5883**	**0.6315** $\pm$ 0.0055	**0.5393**
Fixed sum fusion	0.6216 $\pm$ 0.0118	0.5731	0.6216 $\pm$ 0.0216	0.5270

CNN kernel scale. We further examine the effect of temporal kernel scale in the time-domain CNN branch. As shown in Table 21, both single-kernel variants underperform the multi-scale default. On FACED, using only the narrow kernel $q = 25$ and the wide kernel $q = 99$ decreases Balanced Accuracy by 1.56 and 2.09 percentage points, respectively. On BCIC2020-3, the gap is smaller: $q = 25$ trails the default by 0.88 points, while $q = 99$ is only 0.27 points lower.

These results suggest that informative EEG waveform patterns appear at multiple temporal scales. A single narrow or wide kernel can capture part of the signal, but the multi-scale design $\mathcal{Q}_{\text{temp}} =$

$\{25, 49, 99\}$ provides a more robust time-domain embedding across tasks with different temporal dynamics.

Table 21: Ablation on temporal CNN kernel scale.

Temporal Kernel Set	FACED		BCIC2020-3	
	Balanced Acc	Kappa	Balanced Acc	Kappa
Multi-scale $\mathcal{Q}_{\text{temp}} = \{25, 49, 99\}$	**0.6352** $\pm$ 0.0037	**0.5883**	**0.6315** $\pm$ 0.0055	**0.5393**
Single-scale $\mathcal{Q}_{\text{temp}} = \{25\}$	0.6196 $\pm$ 0.0089	0.5697	0.6227 $\pm$ 0.0206	0.5283
Single-scale $\mathcal{Q}_{\text{temp}} = \{99\}$	0.6143 $\pm$ 0.0140	0.5641	0.6288 $\pm$ 0.0137	0.5360

H.9 Effect of Auxiliary Expert-Balancing Loss

We examine whether the auxiliary expert-balancing loss $\mathcal{L}_{\text{aux}}$ is necessary for maintaining diverse expert routing in CTR-FFN. This analysis uses the ablation setting with N=16, K=4, and Temporal-Former routing. We compare the default model in this setting with an ablated variant trained without the auxiliary expert-balancing loss, and analyze the routed experts on four downstream datasets: FACED, SEED-V, BCIC2020-3, and PhysioNet-MI. To quantify cross-dataset expert overlap, we compute the Jaccard similarity between the top-4 routed expert sets of different datasets. For two expert sets A and B, it is defined as $J(A, B) = |A \cap B|/|A \cup B|$. A higher Jaccard similarity indicates that different datasets rely on more similar experts, suggesting more concentrated and less task-specific routing.

Table 22: Top-4 routed experts with and without auxiliary expert-balancing loss.

Dataset	Default	*w/o* Aux. Loss
FACED	$\{6, 3, 5, 15\}$	$\{12, \mathbf{9}, 6, 2\}$
SEED-V	$\{13, 16, 1, 8\}$	$\{4, \mathbf{9}, \underline{8}, 5\}$
BCIC2020-3	$\{2, 13, 11, 8\}$	$\{\mathbf{9}, \underline{8}, 12, 10\}$
PhysioNet-MI	$\{13, 6, 5, 3\}$	$\{\mathbf{9}, \underline{8}, 7, 13\}$

Note: Expert **9** appears in the top-4 experts for all four datasets under the ablated setting; expert $\underline{8}$ appears in three out of four datasets.

Table 23: Quantitative evidence of routing concentration.

Metric	Default	w/o Aux. Loss	Change
Mean pairwise Jaccard of top-4 experts	0.203	0.314	+55%
Universal experts in top-4 across 4 datasets	0	1	+1
Experts appearing in ≥ 3 datasets' top-4	1	2	+1
Task-specific experts appearing in only 1 dataset	5	5	0

1) Removing the auxiliary loss increases cross-dataset expert overlap. As shown in Table 23, the mean pairwise Jaccard similarity rises from 0.203 to 0.314 after removing the auxiliary expert-balancing loss, corresponding to a relative increase of 55%. This indicates that the dominant experts selected by different downstream datasets become more similar without the balancing constraint.

2) Removing the auxiliary loss creates a universal dominant expert. As shown in Table 22, in the default model, no expert appears in the top-4 routed experts for all four datasets. In contrast, without the auxiliary expert-balancing loss, expert 9 appears in the top-4 routed experts for FACED, SEED-V, BCIC2020-3, and PhysioNet-MI, while expert 8 appears in three out of four datasets. This suggests that the router becomes more dependent on a small subset of experts when the auxiliary

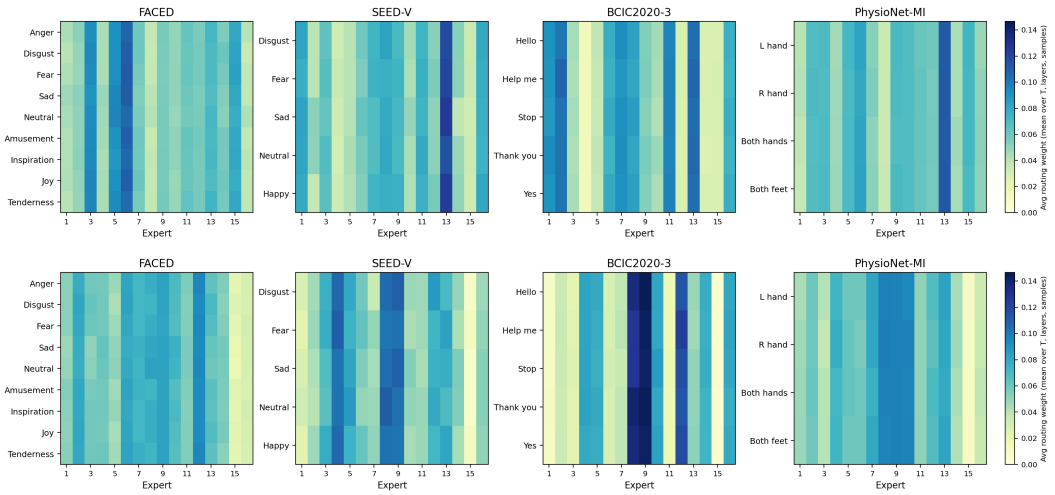


Figure 5: **Effect of auxiliary expert-balancing loss on per-class routing.** Average routing weights over time, layers, and samples for four downstream datasets under the auxiliary-loss ablation setting with $N{=}16$, $K{=}4$, and TemporalFormer routing. *Top:* default model with auxiliary expert-balancing loss. *Bottom:* ablated model without auxiliary expert-balancing loss. Removing the auxiliary loss leads to more concentrated expert usage: expert 9 appears in the top-4 experts for all datasets, and the mean pairwise top-4 Jaccard similarity increases from 0.203 to 0.314 ($+55\%$).

loss is removed. Overall, these results indicate that the auxiliary expert-balancing loss helps prevent routing collapse and encourages more diverse expert utilization across downstream tasks.

Figure 5 provides a qualitative visualization of the same trend, where the ablated model shows more concentrated routing weights across datasets.

Appendix H.10 next examines how routing varies across layers and temporal steps within fixed recording windows.

H.10 Layer-wise and Within-recording Routing Diagnostics

We analyze every layer and temporal step using 256 held-out windows from each of FACED and PhysioNet-MI.

Within each window, dataset, subject, and montage remain fixed while temporal step varies. This controlled comparison tests whether routing remains temporally non-static after these identity variables are held constant.

Table 24: Most frequently selected expert at each layer. Parentheses give its selection rate among the top-K=4 experts; the uniform reference for N=16 is 0.25.

Layer	FACED	PhysioNet-MI	Same Expert
L0	E11 (1.00)	E11 (1.00)	✓
L1	E8 (0.68)	E13 (0.57)	✗
L2	E8 (0.81)	E5 (0.96)	✗
L3	E5 (0.99)	E5 (0.99)	✓
L4	E4 (0.99)	E4 (0.98)	✓
L5	E12 (0.99)	E12 (1.00)	✓
L6	E13 (0.91)	E11 (0.83)	✗
L7	E14 (0.99)	E14 (1.00)	✓
L8	E2 (0.99)	E2 (1.00)	✓
L9	E5 (0.88)	E8 (0.92)	✗
L10	E12 (0.73)	E12 (0.87)	✓
L11	E10 (0.96)	E10 (0.97)	✓

The dominant expert agrees across datasets in eight of twelve layers. The remaining layers show dataset-dependent preferences, indicating shared depth-wise organization alongside dataset-dependent adaptation.

Table 25: Layer-3 routing within fixed dataset, subject, and montage windows. Values are routing weights; a dash denotes an expert not reported for that sequence.

Dataset	Timestep	E12	E5	E6	Highest-weight Expert
FACED	0	**0.50**	0.15	0.13	E12
	1	**0.27**	0.22	0.21	E12
	2	0.14	**0.27**	0.26	E5
	3–5	0.04–0.08	**0.29–0.30**	0.28–0.30	E5
	6–9	$\leq$ 0.03	0.27–0.30	**0.30–0.31**	E6
PhysioNet-MI	0	**0.50**	0.18	–	E12
	1	**0.31**	0.26	–	E12
	2	0.16	**0.31**	–	E5
	3	0.09	**0.34**	–	E5

At layer 3, the highest-weight expert changes from E12 to E5 in both datasets and later to E6 in FACED. Across layers, the observed changes are concentrated mainly in L1–L5, while deeper routing is more stable.

These within-window transitions demonstrate that routing varies over time even when dataset, subject, and montage are fixed.

Together, Appendix H.9 shows that expert balancing reduces global routing concentration, while this analysis shows time-varying selection within fixed recording windows.

H.11 Computational Profile and Efficiency Trade-off

We profile the routing variants from Appendix H.3 with the same 570-token input. The MoE variants use N=8 specialized experts plus one shared expert, top-K=2 activation, and an intermediate expert width of 800.

Table 26: Computational profile for the $N=8$, $K=2$ routing ablations with a 570-token input. Activated parameters count the active paths for one forward pass.

Configuration	Total Params. (M)	Activated Params. (M)	GFLOPs/Sample
Non-MoE	15.86	15.86	18.08
Vanilla MoE	33.16	15.88	18.10
Mean Temporal	33.16	15.88	18.10
TRACE / CTR-FFN	38.96	21.68	24.72

Sparse activation lets Vanilla MoE increase stored capacity with almost unchanged forward cost. TRACE adds the TemporalFormer router, raising active parameters by 5.80M and computation by 6.62 GFLOPs over Vanilla MoE.

Relative to Non-MoE, the controlled TRACE variant uses about 36.7% more forward computation. This cost should be interpreted together with the downstream improvements in Appendix H.3.