STAGE-ALIGNED TEMPORAL REPRESENTATION LEARNING FOR MOTOR IMAGERY EEG DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Motor imagery (MI) EEG decoding relies on class-discriminative information encoded in the temporal evolution of sensorimotor-rhythm activity. Yet existing decoders largely treat temporal operators as generic architectural components, overlooking the distinct temporal learning objectives required by successive EEG representation stages. We formulate Stage-Aligned Temporal Representation Learning (SATRL), a framework that aligns distinct temporal operators with three stage-specific learning objectives. SATRL instantiates this principle through three complementary temporal operators: a multiscale temporal extractor that captures sensorimotor-rhythm responses from channel-wise EEG dynamics, a stacked small-kernel depthwise aggregator that composes local rhythm-related activations into stable short-window representations, and a context-guided dilated temporal module that integrates long-range evidence under trial-level context. SATRL elevates the recurring temporal-spatial design pattern into an explicit, principled, and experimentally testable operator-stage alignment framework. Experiments on public MI-EEG benchmarks demonstrate consistent improvements across both compact CNN and CNN-Transformer backbones. Under cross-session evaluations, SATRL improves the corresponding baselines by up to 5.26 and 10.24 percentage points, respectively, while achieving state-of-the-art classification accuracies on all datasets. Leave-one-subject-out evaluations further place SATRL in the top-performing tier, and stage-ablation studies consistently show that removing or changing any temporal stage degrades performance in all backbone types. Together, these results suggest SATRL as a principled framework for MI-EEG decoding.

1 INTRODUCTION

Electroencephalography (EEG) decoding is an important problem of representation learning from continuous electrophysiological recordings (Wolpaw et al., 2002; Chaudhary et al., 2016). In motor imagery (MI), a major source of class-discriminative information is frequency-specific, spatially organized, and temporally evolving modulations of sensorimotor rhythms, particularly mu- and beta-band event-related desynchronization patterns and their spatiotemporal organization (Pfurtscheller & Neuper, 1997; Pfurtscheller et al., 2006). Importantly, neural rhythms can manifest as brief, temporally localized bursts rather than sustained oscillations (Karvat et al., 2026). The temporal expression of these modulations is not uniform: their emergence, persistence, and lateralization vary across MI trials and task conditions (Pfurtscheller et al., 1999; Nam et al., 2011). These task-relevant temporal patterns are often weak and heterogeneous and can be obscured by physiological variability and non-neural artifacts (Makeig et al., 1996; Delorme et al., 2007; Wang et al., 2024). MI-EEG decoding therefore requires models to preserve evolving class-dependent rhythm information across successive representation stages and integrate it into discriminative trial-level representations (Wang et al., 2024; Saibene et al., 2024).

Prior MI-EEG decoding pipelines implicitly exploited temporal information across successive processing stages. Common spatial pattern (CSP) and filter-bank CSP (FBCSP) use spatial projection and frequency-specific filtering before extracting discriminative features through temporal variance estimation at a subsequent stage (Ramoser et al., 2000; Ang et al., 2008; 2012). Deep architectures make these transformations learnable: Shallow ConvNet and EEGNet employ successive temporal and spatial feature-extraction stages, whereas FBCNet and FBSF-TSCNN further process

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.

post-spatial temporal information through variance aggregation and learnable temporal decoding (Schirrmeister et al., 2017; Lawhern et al., 2018; Mane et al., 2020; 2021; Chen et al., 2020). Recent sequence-based architectures extend later-stage processing with temporal convolution, attention, and Transformer-based sequence modeling after EEG-specific front ends to capture broader temporal dependencies (Ingolfsson et al., 2020; Altaheri et al., 2023; Song et al., 2023; Liu et al., 2024; Baspinar & Tastan, 2025; Bouchane et al., 2025; Yang et al., 2025; Zhao et al., 2024; 2025; Altaheri et al., 2025). These developments have expanded the available temporal modeling mechanisms and network structures. However, temporal operators are often introduced to increase model capacity, with less attention to the changing requirements of successive representation stages. Without adaptation to these requirements, class-dependent rhythm information could be weakened or inadequately integrated across stages, limiting the discriminability of the resulting trial-level representations.

This challenge stems from the stage-dependent evolution of temporal information in EEG representations. Temporal operators are applied along the same temporal axis, but the information changes as EEG representations evolve. Effective EEG temporal modeling is therefore a stage-dependent problem of defining the learning objective for each temporal operator. At the sensor level, temporal patterns reflect channel-wise source-mixed EEG dynamics affected by volume conduction, sensor noise, and artifacts (Makeig et al., 1996; Delorme et al., 2007). After spatial reorganization, temporal patterns correspond to local activation trajectories that summarize rhythm-related responses. At later stages, temporal modeling operates on accumulated task-related evidence requiring contextual integration within a trial. Consequently, different representation stages impose distinct temporal learning requirements: early stages require capturing heterogeneous rhythm responses, intermediate stages require composing local temporal evidence, and later stages require integrating broader trial-level context.

In this study, we address this challenge by formulating Stage-Aligned Temporal Representation Learning (SATRL) for MI-EEG decoding. SATRL treats temporal modeling as an operator-stage alignment problem, where temporal operators are explicitly designed according to three distinct stage-specific temporal learning objectives: sensorimotor-rhythm response extraction, local short-window temporal aggregation, and trial-level temporal integration. This formulation elevates MI-EEG decoders into an explicit, principled, and experimentally testable framework. The main contributions of this work are:

- We identify the evolving role of temporal information across MI-EEG representation stages and formulate SATRL as an operator-stage alignment framework that explicitly matches temporal operators to three stage-specific temporal learning objectives.
- We instantiate SATRL with a sensor-level multiscale temporal extractor, an intermediate stacked small-kernel depthwise aggregator, and a late context-guided dilated temporal module, making each stage-specific objective concrete and testable across different architectures.
- We validate SATRL through compact CNN and CNN-Transformer realizations, benchmark comparisons, operator variants, and stage ablations, showing that complete stage alignment is consistently more effective than reduced-stage alternatives.

2 METHOD

2.1 OVERALL FRAMEWORK

Given an MI-EEG trial $X \in \mathbb{R}^{C\times T}$ with C channels and T temporal samples, SATRL learns a task-relevant representation $F^{(i)}$, where $i \in \{1, 2, 3\}$ indexes the three processing stages. The final representation is mapped to the predicted label $Y \in \{1, \ldots, K\}$, where K denotes the number of motor-imagery classes.

As illustrated in Figure 1, SATRL progressively transforms multichannel EEG into task-discriminative trial representations through three stage-specific temporal modules(TM). Stage-I employs a multiscale temporal extractor to capture complementary rhythm-related responses. Stage-II applies stacked small-kernel depthwise convolutions to consolidate local activations into short-window temporal representations. Stage-III uses a context-guided dilated temporal module to integrate evidence distributed across the trial. Here, a stage-specific temporal learning objective denotes

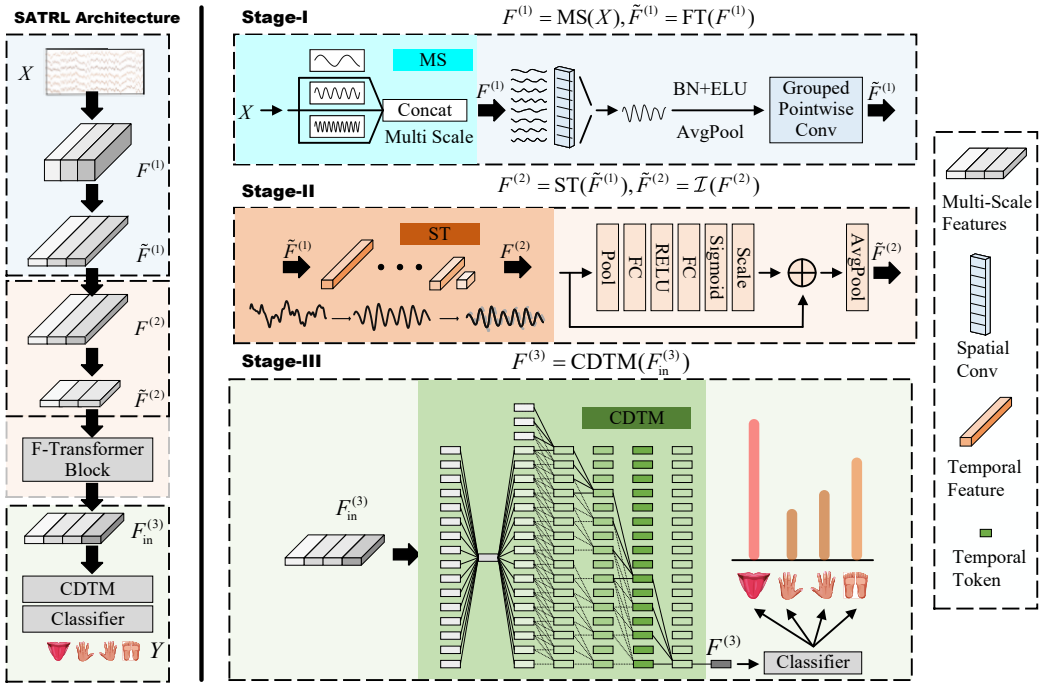


Figure 1: The overall architecture and information flow of the proposed framework.

the representational role assigned to a temporal module; Spatial and pointwise feature transformations, together with attention-based recalibration and fusion, reorganize and combine intermediate features within each specific stage. All three stages are jointly optimized through the final classification loss.

2.2 STAGE-I TM

MI-related rhythm responses exhibit a variety of temporal durations. The multiscale temporal extractor (MS) captures complementary responses using M parallel convolutional branches with kernel lengths $\{k_1, \ldots, k_M\}$. For batched computation, let $X \in \mathbb{R}^{B\times C\times T}$ denote B input trials. Each branch applies temporal convolution followed by batch normalization (BN):

$$H_i = \psi(\mathrm{Conv}_{1\times k_i}(X)), \qquad i = 1, \ldots, M, \tag{1}$$

where ψ denotes BN. The responses are concatenated along the feature-channel dimension to obtain $F^{(1)} = \mathrm{Concat}(H_1, \ldots, H_M)$. Feature transformation (FT) maps $F^{(1)}$ to the input of Stage-II TM, as detailed in Section 2.4.

2.3 STAGE-II TM

Stage-II TM consolidates local temporal dependencies in the transformed EEG features. We compare operators that share this objective, but implement local temporal aggregation in different ways, as illustrated in Figure 2. They provide richer ways to represent localized EEG segments through stacked composition, parallel temporal supports, large-kernel processing, or adaptive sampling, which demonstrates the generality of stage-II across alternative temporal aggregation mechanisms.

Stacked Small-Kernel Temporal Operator(ST). ST applies L serial depthwise temporal convolutions with kernel lengths $\{s_1, \ldots, s_L\}$. The first convolution, followed by BN and exponential linear unit (ELU) activation, maps $\widetilde{F}^{(1)}$ to $F_1^{(2)}$. Subsequent layers are defined recursively as

$$F_\ell^{(2)} = \phi\Big(\mathrm{DWConv}_{1\times s_\ell}(F_{\ell-1}^{(2)})\Big), \qquad \ell = 2, \ldots, L, \tag{2}$$

where ϕ denotes BN followed by ELU, and the final output is $F^{(2)} = F_L^{(2)}$. Unlike a single convolution, ST progressively composes temporal responses through intermediate normalized nonlinear representations. The proposed SATRL adopts ST as its default Stage-II TM, while the other operators are evaluated in the ablation study.

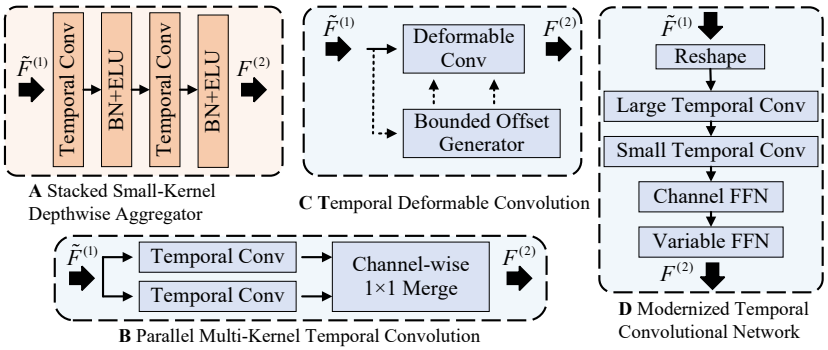


Figure 2: Schematic illustration of the operators evaluated for Stage-II TM.

Parallel-Branch Temporal Operator(PL). PL applies parallel depthwise convolutions with kernel lengths $\{p_1, \ldots, p_J\}$ to the same input, producing $B_j = \mathrm{DWConv}_{1\times p_j}(\widetilde{F}^{(1)})$, $j = 1, \ldots, J$. A grouped pointwise convolution merges the concatenated responses:

$$F^{(2)} = \mathcal{M}(\mathrm{Concat}(B_1, \ldots, B_J)), \tag{3}$$

where $\mathcal{M}$ denotes a grouped 1×1 convolution that preserves the output channel dimension.

Modern Temporal Convolutional Network Operator(mTCN). The mTCN operator combines large-kernel depthwise convolution with grouped pointwise feed-forward transformations through N residual blocks:

$$F_b^{\mathrm{mTCN}} = F_{b-1}^{\mathrm{mTCN}} + \mathcal{B}_b\big(F_{b-1}^{\mathrm{mTCN}}\big), \qquad b = 1, \ldots, N, \tag{4}$$

with $F_0^{\mathrm{mTCN}} = \widetilde{F}^{(1)}$. Each residual branch is

$$\mathcal{B}_b(F) = \mathcal{G}_b(\psi[\mathrm{DWConv}_{1\times K_L}(F) + \mathrm{DWConv}_{1\times K_S}(F)]). \tag{5}$$

Here, $F = F_{b-1}^{\mathrm{mTCN}}$ is the block input, and K_L and K_S denote the large and small kernel lengths. The transformation $\mathcal{G}_b$ comprises a channel feed-forward network (FFN) and an optional variable FFN, with Gaussian Error Linear Unit(GELU) activation and dropout. The final output is $F^{(2)} = F_N^{\mathrm{mTCN}}$.

Temporal Deformable Operator(TDC). TDC predicts bounded offsets for adaptive temporal sampling. An intermediate response is computed as $R_\Delta = \mathrm{ELU}(\mathrm{DWConv}_{1\times k_\Delta}(\widetilde{F}^{(1)}))$, and the offset field is

$$\mathbf{\Delta} = \Delta_{\max} \tanh(\mathrm{Conv}_{1\times 1}(R_\Delta)), \tag{6}$$

where k_Δ denotes the offset-generation kernel length and $\Delta_{\max}$ bounds the temporal displacement. Adaptive convolution, normalization, activation, and dropout produce

$$F^{(2)} = \mathrm{Drop}\Big(\phi\Big(\mathrm{DeformConv}_{1\times k_D}(\widetilde{F}^{(1)}, \mathbf{\Delta})\Big)\Big), \tag{7}$$

where k_D denotes the deformable kernel length and Drop denotes dropout. Regardless of the operator used in Stage-II TM, its output $F^{(2)}$ is passed to the attention block of stage-II which is implemented as the F-Transformer block.

2.4 Feature Transformation(FT) and F-Transformer Block

The FT mapping, $\widetilde{F}^{(1)} = \mathrm{FT}(F^{(1)})$, applies spatial depthwise convolution across EEG channels after stage-I TM, followed by temporal pooling, and pointwise feature projection.

After Stage-II TM, grouped squeeze-and-excitation (SE) attention (Altaheri et al., 2025)recalibrates the feature groups within a residual connection, followed by pooling and dropout:

$$\widetilde{F}^{(2)} = \mathcal{I}(F^{(2)}) = \mathrm{AvgPool}(F^{(2)} + \mathrm{SE}(F^{(2)})). \tag{8}$$

SE generates group-wise rescaling gates through temporal average pooling and fully connected transformations with ReLU and sigmoid activations. Here, $\mathrm{AvgPool}$ denotes average pooling along the temporal dimension, and dropout is applied to the resulting $\widetilde{F}^{(2)}$ before it is passed to the subsequent branch.

SATRL supports two realizations, a CNN-based realization and a CNN-Transformer realization. In the CNN-Transformer realization, the F-Transformer block transforms $\widetilde{F}^{(2)}$ into $F_{\mathrm{GQA}} = \mathrm{GQA}(\widetilde{F}^{(2)})$ through pointwise feature mixing, grouped-query attention (GQA) transformer blocks(Ainslie et al., 2023) and channel reduction. The resulting branch is concatenated with $\widetilde{F}^{(2)}$ before Stage-III TM:

$$F_{\mathrm{in}}^{(3)} = \mathrm{Concat}(\widetilde{F}^{(2)}, F_{\mathrm{GQA}}). \tag{9}$$

2.5 Stage-III TM

Because task-related evidence is distributed throughout a trial, the context-guided dilated temporal module (CDTM) integrates information across the temporal positions of $F_{\mathrm{in}}^{(3)}$. The trial-level context $\mathbf{c} = T'^{-1} \sum_{t=1}^{T'} F_{\mathrm{in}}^{(3)}(\mathrm{t})$ is computed by averaging the input features over time, where T' denotes the temporal length. The context vector is broadcast along the temporal dimension, scaled by α, and added to the input features. The TCN processes this context-enriched representation to produce the Stage-III output.

$$F^{(3)} = \mathrm{TCN}\left(F_{\mathrm{in}}^{(3)} + \alpha c\right). \tag{10}$$

The weight α controls the contribution of trial-level context. The temporal convolutional network (TCN) (Altaheri et al., 2023) consists of stacked residual temporal convolutional blocks. Each block applies two causal convolutions with the same dilation, and the dilation increases across blocks. This design expands the effective receptive field while preserving temporal order and restricting each position to preceding evidence. CDTM can therefore combine trial-level context with progressively broader temporal dependencies.

The final temporal state is passed to a grouped 1×1 pointwise classifier, which maps each feature group to class logits. The logits are averaged across groups, and the class with the highest averaged logit is selected as the prediction.

3 Experiments

3.1 Datasets and Protocols

We train and evaluate SATRL on four public EEG benchmarks for motor decoding. These comprise BCI Competition IV-2a (BCICIV2a), BCI Competition IV-2b (BCICIV2b), the High-Gamma Dataset (HGD), and the Ofner Upper-Limb Movement Dataset (Ofner) (Brunner et al., 2008; Schirrmeister et al., 2017; Ofner et al., 2017). BCICIV2a contains four motor-imagery classes, whereas BCICIV2b contains left- and right-hand imagery with fewer EEG channels. HGD provides higher-density recordings of executed movements and rest, with 44 motor-area channels retained in our implementation. Ofner includes execution and imagery of six right-upper-limb movements and rest, recorded with 61 EEG channels. Together, these benchmarks span different class structures, channel configurations, and motor tasks.

We use two evaluation protocols. In the cross-session setting, each subject's training partition is used for parameter optimization, whereas the evaluation partition serves as both validation and test data, which can preserve all available training trials when subject-specific samples are limited. For LOSO, we hold out one subject and pool the others' training and evaluation partitions for training and validation, thereby assessing cross-subject generalization. For each method, we report the accuracy at the final training epoch of the test set (Last Acc.) and the highest accuracy obtained during training (Best Acc.). All reported results are presented as the mean $\pm$ standard deviation computed with 5 different random seeds.

Table 1: Cross-session classification results on BCIC IV-2a, BCIC IV-2b, HGD, and Ofner.

Method	BCIC IV-2a		BCIC IV-2b		HGD		Ofner	
	Last Acc.	Best Acc.	Last Acc.	Best Acc.	Last Acc.	Best Acc.	Last Acc.	Best Acc.
CNN-based methods								
ShallowConvNet$_{\text{HBM'17}}$	69.74 ± 1.07	76.40 ± 0.33	84.17 ± 0.45	88.28 ± 0.19	87.99 ± 0.86	93.55 ± 0.11	25.33 ± 1.91	26.09 ± 2.18
EEG-TCNet$_{\text{SMC'20}}$	78.72 ± 1.09	81.43 ± 0.75	$\mathbf{87.08 \pm 0.33}$	$\mathbf{89.53 \pm 0.12}$	87.80 ± 0.54	90.61 ± 0.37	$\mathbf{25.43 \pm 0.63}$	25.92 ± 0.63
FBCNet$_{\text{EMBC'20}}$	71.33 ± 1.45	78.89 ± 0.12	75.79 ± 0.76	81.39 ± 0.18	89.36 ± 0.35	93.61 ± 0.07	24.78 ± 0.42	26.15 ± 0.31
EEGNet-8.2$_{\text{JNE'18}}$	75.32 ± 0.93	78.35 ± 0.43	83.65 ± 0.72	86.03 ± 0.74	88.42 ± 0.54	91.16 ± 0.45	23.24 ± 0.64	24.04 ± 0.61
Ours	$\mathbf{80.63 \pm 0.81}$	$\mathbf{82.89 \pm 0.73}$	86.50 ± 0.60	89.20 ± 0.39	$\mathbf{93.25 \pm 0.70}$	$\mathbf{95.90 \pm 0.59}$	23.71 ± 1.93	$\mathbf{29.30 \pm 0.71}$
CNN + Transformer-based methods								
ATCNet$_{\text{TII'23}}$	78.01 ± 1.12	83.14 ± 0.84	86.36 ± 0.55	90.05 ± 0.17	92.61 ± 0.44	95.97 ± 0.42	24.10 ± 0.76	33.39 ± 0.37
TCFormer$_{\text{SciRep'25}}$	85.05 ± 0.65	87.86 ± 0.16	87.22 ± 0.47	90.54 ± 0.20	94.47 ± 0.31	96.51 ± 0.18	$\mathbf{25.28 \pm 0.54}$	33.18 ± 0.23
CTNet$_{\text{SciRep'24}}$	81.53 ± 0.51	83.76 ± 0.46	87.20 ± 0.52	90.46 ± 0.17	90.14 ± 0.70	93.22 ± 0.50	24.23 ± 0.30	33.45 ± 0.56
Conformer$_{\text{TNSRE'23}}$	76.03 ± 0.80	78.55 ± 0.60	86.21 ± 0.34	88.50 ± 0.26	$\mathbf{94.50 \pm 0.27}$	96.05 ± 0.13	23.05 ± 0.69	31.36 ± 0.26
MSCFormer$_{\text{SciRep'25}}$	81.15 ± 0.55	83.11 ± 0.64	86.95 ± 0.34	89.51 ± 0.22	93.54 ± 0.28	95.78 ± 0.42	23.74 ± 0.80	31.76 ± 0.75
Ours	$\mathbf{85.96 \pm 0.66}$	$\mathbf{88.79 \pm 0.40}$	$\mathbf{87.27 \pm 0.12}$	$\mathbf{90.74 \pm 0.19}$	92.82 ± 0.22	$\mathbf{97.18 \pm 0.04}$	24.69 ± 1.56	$\mathbf{34.04 \pm 0.36}$

Table 2: Leave-one-subject-out (LOSO) classification results on BCIC IV-2a, BCIC IV-2b, HGD, and Ofner.

Method	BCIC IV-2a		BCIC IV-2b		HGD		Ofner	
	Last Acc.	Best Acc.	Last Acc.	Best Acc.	Last Acc.	Best Acc.	Last Acc.	Best Acc.
CNN-based methods								
ShallowConvNet$_{\text{HBM'17}}$	53.65 ± 0.12	56.10 ± 0.74	79.37 ± 0.55	80.24 ± 0.43	68.49 ± 0.85	71.04 ± 0.85	20.13 ± 0.45	21.20 ± 0.34
EEG-TCNet$_{\text{SMC'20}}$	56.69 ± 0.47	57.21 ± 0.47	$\mathbf{80.75 \pm 0.31}$	81.26 ± 0.45	69.00 ± 1.10	69.55 ± 1.15	21.27 ± 0.27	21.33 ± 0.23
FBCNet$_{\text{EMBC'20}}$	42.64 ± 1.09	44.97 ± 1.80	71.97 ± 0.25	73.94 ± 0.30	62.01 ± 0.95	63.49 ± 0.23	19.91 ± 0.55	20.01 ± 0.44
EEGNet-8.2$_{\text{JNE'18}}$	56.97 ± 0.53	57.18 ± 0.61	80.63 ± 0.84	$\mathbf{81.39 \pm 0.63}$	64.05 ± 0.73	65.01 ± 0.79	20.78 ± 0.51	21.03 ± 0.43
Ours	$\mathbf{58.50 \pm 0.92}$	$\mathbf{58.83 \pm 0.85}$	80.05 ± 0.55	80.86 ± 0.30	$\mathbf{71.50 \pm 1.32}$	$\mathbf{72.72 \pm 1.21}$	$\mathbf{22.34 \pm 0.25}$	$\mathbf{22.42 \pm 0.26}$
CNN + Transformer-based methods								
ATCNet$_{\text{TII'23}}$	59.16 ± 0.50	59.59 ± 0.60	80.57 ± 0.05	81.25 ± 0.06	68.66 ± 0.79	71.11 ± 0.53	22.65 ± 0.79	23.36 ± 0.80
TCFormer$_{\text{SciRep'25}}$	$\mathbf{61.31 \pm 0.70}$	$\mathbf{61.57 \pm 0.67}$	$\mathbf{81.20 \pm 0.51}$	$\mathbf{81.87 \pm 0.50}$	$\mathbf{74.23 \pm 0.56}$	$\mathbf{74.90 \pm 0.38}$	22.22 ± 0.27	22.78 ± 0.27
CTNet$_{\text{SciRep'24}}$	59.21 ± 0.36	59.41 ± 0.24	80.55 ± 0.15	81.34 ± 0.25	69.67 ± 0.76	70.65 ± 0.72	21.17 ± 0.68	23.18 ± 0.51
Conformer$_{\text{TNSRE'23}}$	54.06 ± 0.44	54.41 ± 0.54	77.83 ± 0.33	78.43 ± 0.25	72.20 ± 0.67	72.69 ± 0.67	22.16 ± 0.48	22.78 ± 0.62
MSCFormer$_{\text{SciRep'25}}$	55.68 ± 0.26	55.95 ± 0.43	79.91 ± 0.69	80.17 ± 0.63	67.19 ± 0.73	68.30 ± 0.85	17.41 ± 0.39	18.64 ± 0.29
Ours	60.42 ± 1.01	60.94 ± 0.81	80.75 ± 0.31	81.66 ± 0.68	73.42 ± 0.53	74.64 ± 0.33	$\mathbf{23.77 \pm 0.78}$	$\mathbf{24.63 \pm 0.30}$

3.2 Experimental Setup

Preprocessing. EEG signals are bandpass filtered at 0.5–40 Hz, downsampled to 100 Hz, segmented into 4 seconds trials, and standardized using statistics computed from the training data only. No information from the test session or held-out subject is used during normalization or training. Within each dataset and protocol, all methods use the same preprocessing and augmentation strategy. Backbone comparisons follow their fixed training settings and are reported for reproducibility. Specifically, the Ofner dataset uses a 3 seconds EEG window.

Compared Methods. We benchmark SATRL against representative motor-imagery electroencephalography (MI-EEG) decoders from two architectural families. The convolutional and deep-learning group includes ShallowConvNet (Schirrmeister et al., 2017), EEG-TCNet (Ingolfsson et al., 2020), FBCNet (Mane et al., 2021), and EEGNet-8.2 (Lawhern et al., 2018). The CNN–Transformer group includes ATCNet (Altaheri et al., 2023), TCFormer (Altaheri et al., 2025), CTNet (Zhao et al., 2024), Conformer (Song et al., 2023), and MSCFormer (Zhao et al., 2025). All methods are evaluated using the same preprocessing, data partitions, training protocol, and evaluation metrics. The CNN and CNN–Transformer realizations of SATRL are reported within their corresponding groups.

Implementation Details All models are implemented in PyTorch with PyTorch Lightning and trained on a single NVIDIA GeForce RTX 5090 GPU(32 GB). We use Adam with a base learning rate of 9×10^{-4}, $(\beta_1, \beta_2) = (0.5, 0.999)$, default $\epsilon = 10^{-8}$, and weight decay of 10^{-3}. The learning rate is linearly warmed up and then decayed according to a cosine schedule. The batch size is 48. Models are trained for at most 1000 epochs in cross-session evaluation, 125 epochs in LOSO evaluation. The checkpoint with the highest validation accuracy is retained.

Table 3: Stage ablation results of CNN-based SATRL under cross-session evaluation.

Stage I	Stage II	Stage III	BCIC IV-2a		BCIC IV-2b		HGD	
			Last Acc.	Best Acc.	Last Acc.	Best Acc.	Last Acc.	Best Acc.
SS	SK	-	71.44 ± 1.35	72.79 ± 1.47	85.76 ± 0.38	87.83 ± 0.48	92.43 ± 0.38	94.32 ± 0.24
MS	-	-	75.56 ± 1.06	77.27 ± 1.01	86.17 ± 0.32	88.50 ± 0.12	93.65 ± 0.65	95.49 ± 0.45
MS	SK	-	76.37 ± 0.72	78.29 ± 0.72	86.30 ± 0.43	88.77 ± 0.24	93.70 ± 0.65	95.57 ± 0.45
MS	ST	-	77.28 ± 0.34	79.66 ± 0.46	85.74 ± 0.54	87.87 ± 0.26	92.91 ± 0.38	94.65 ± 0.26
MS	PL	-	76.96 ± 0.69	79.27 ± 0.72	86.13 ± 0.56	88.31 ± 0.24	92.98 ± 0.78	94.75 ± 0.51
MS	mTCN	-	77.11 ± 0.66	79.13 ± 0.77	85.74 ± 0.25	88.69 ± 0.41	93.41 ± 0.23	96.05 ± 0.20
MS	TDC	-	77.58 ± 0.64	79.43 ± 0.51	85.21 ± 0.34	87.92 ± 0.22	93.22 ± 0.53	94.81 ± 0.57
MS	SK	CDTM	79.78 ± 0.61	81.65 ± 0.55	87.32 ± 0.44	89.35 ± 0.31	94.77 ± 0.75	96.34 ± 0.68
SS	ST	CDTM	76.48 ± 0.86	78.47 ± 0.62	86.73 ± 0.39	88.45 ± 0.64	93.89 ± 0.55	96.03 ± 0.43
MS	**ST**	**CDTM**	$\mathbf{80.63 \pm 0.81}$	$\mathbf{82.89 \pm 0.73}$	$\mathbf{86.50 \pm 0.60}$	$\mathbf{89.20 \pm 0.39}$	$\mathbf{93.25 \pm 0.70}$	$\mathbf{95.90 \pm 0.59}$

Table 4: Stage ablation results of CNN–Transformer-based SATRL under cross-session evaluation.

Stage I	Stage II	Stage III	BCIC IV-2a		BCIC IV-2b		HGD	
			Last Acc.	Best Acc.	Last Acc.	Best Acc.	Last Acc.	Best Acc.
SS	SK	-	62.34 ± 1.63	65.44 ± 1.33	82.98 ± 1.57	85.85 ± 1.31	77.05 ± 1.47	81.91 ± 0.99
MS	-	-	68.97 ± 0.41	71.98 ± 0.64	83.75 ± 0.58	86.41 ± 0.16	82.57 ± 1.43	86.53 ± 1.04
MS	SK	-	72.64 ± 1.17	75.56 ± 1.03	83.87 ± 0.19	86.53 ± 0.19	88.03 ± 1.16	91.37 ± 1.21
MS	ST	-	76.84 ± 0.44	79.17 ± 0.30	85.80 ± 0.37	88.41 ± 0.24	93.36 ± 0.81	95.59 ± 0.71
MS	PL	-	72.27 ± 1.93	75.09 ± 1.49	85.67 ± 0.15	87.74 ± 0.23	93.13 ± 0.23	95.53 ± 0.31
MS	mTCN	-	76.65 ± 0.55	79.63 ± 0.48	83.57 ± 0.46	86.83 ± 0.26	93.30 ± 0.36	96.19 ± 0.47
MS	TDC	-	77.92 ± 0.59	80.03 ± 0.48	84.16 ± 0.62	86.66 ± 0.49	92.01 ± 0.48	94.41 ± 0.17
MS	SK	CDTM	83.08 ± 0.66	85.39 ± 0.60	86.42 ± 0.21	89.96 ± 0.22	94.88 ± 0.66	97.09 ± 0.40
SS	ST	CDTM	82.78 ± 0.46	84.54 ± 0.69	86.85 ± 0.36	90.15 ± 0.38	94.51 ± 0.32	96.94 ± 0.43
MS	**ST**	**CDTM**	$\mathbf{85.96 \pm 0.12}$	$\mathbf{88.79 \pm 0.40}$	$\mathbf{87.27 \pm 0.12}$	$\mathbf{90.74 \pm 0.19}$	$\mathbf{92.82 \pm 0.22}$	$\mathbf{97.18 \pm 0.04}$

3.3 Main Results

Comparison with Compact CNN Backbones. SATRL substantially strengthened the compact CNN backbone under both evaluation protocols. Under cross-session evaluation, it achieved best accuracy of 82.89%, 89.20%, 95.90%, and 29.30% on 2a, 2b, HGD, and Ofner, respectively. These results give an average gain of 4.43 percentage points over EEGNet-8.2, with the largest dataset-level improvement being 5.26 points on Ofner. SATRL also reached the highest accuracy among the compact CNN methods on 2a, HGD, and Ofner, while remaining competitive on 2b. Under LOSO evaluation, it retained a 2.52 point average advantage over EEGNet-8.2, including a 7.71 point improvement on HGD. These results confirm that stage-aligned temporal representation learning improves robustness within a compact CNN backbone.

Comparison with CNN–Transformer Backbones. The results also show a similar overall trend in the CNN-transformer realization. Under cross-session evaluation, SATRL achieved the highest best accuracy among all CNN–Transformer methods, surpassing Conformer by 4.07% on average, with the largest improvement reaching 10.24 points on 2a. LOSO results followed the same pattern: SATRL maintained a 3.39% average advantage over Conformer, remaining in the top-performing tier across all four datasets and highlighting broad applicability of SATRL to higher capacity architectures.

Three-stage decoders (e.g., SATRL, TCFormer, and ATCNet) generally ranked above one- and two-stage alternatives, while SATRL achieved the strongest cross-session best accuracy across all datasets. By aligning stage-specific operators, SATRL transforms raw EEG into more stable and discriminative task-relevant representations. Together, these results support SATRL as a principled framework for MI-EEG decoding.

3.4 Ablation Study

We conducted stage-wise ablation experiments to evaluate the contributions of the three temporal stages in SATRL. To distinguish weakened stage designs from complete stage removal, we used controlled substitutions and omission settings. For Stage-I, SS denotes a single scale temporal extractor that retains the stage. For Stage-II, SK denotes a single kernel convolution that reduces

temporal composition. Stage-III was evaluated by removing CDTM, while a dash indicates that the corresponding stage is completely omitted. Tables 3 and 4 report the results for the CNN and CNN–Transformer realizations.

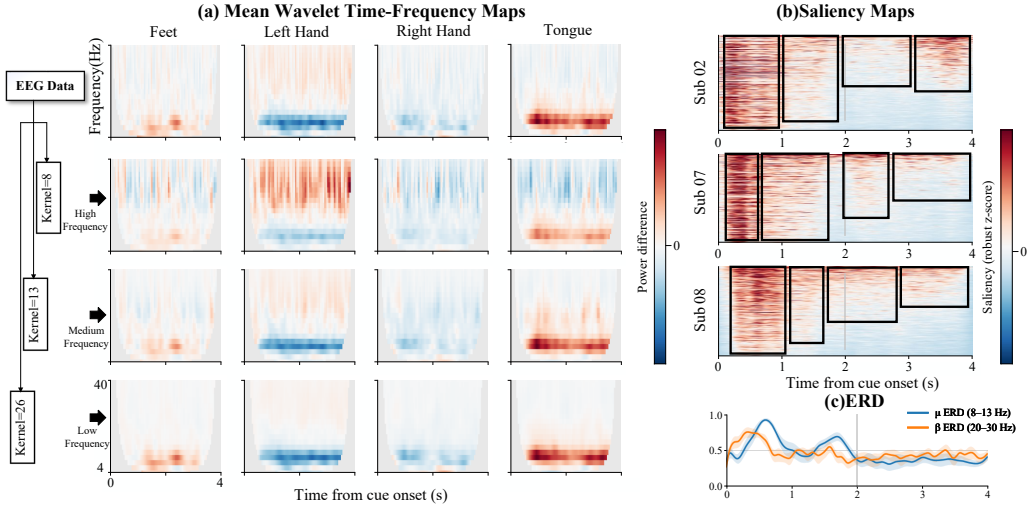


Figure 3: Time-frequency feature maps and saliency maps of SATRL.

Multiscale Pattern Extraction. In stage I, multiscale temporal filtering consistently improved best accuracy across the six dataset-backbone settings, with positive gains ranging from 0.24% to 4.42%. The largest gains occurred on 2a and HGD for the CNN-Transformer realization (4.42% and 4.25%, respectively). The SK control also yielded lower performance than the complete multiscale configuration, indicating that the improvement is not simply due to adding a temporal convolution. These results suggest that MI-EEG benefits from preserving rhythm responses at multiple temporal scales.

Short-Window Rhythm Stabilization. In stage II, removing Stage-II entirely led to a clear performance decrease, whereas retaining even the weakened SK preserved part of the temporal benefit. Across the evaluated settings, these operators generally outperformed the SK control and achieved clearer improvements over the configuration without Stage-II. It was particularly beneficial to explicit local temporal modeling in the CNN–Transformer realization, increasing average best accuracy by 6.08%. This gain indicates that, following feature transformation, locally adjacent rhythm-related activations form informative short-window evidence, for which progressive consolidation is more effective than independent processing.

Trial-Level Context Integration. With Stage-I fixed to MS, adding CDTM increased mean best acc. by 1.75 percentage points for the CNN backbone and 5.42 percentage points for the CNN–Transformer backbone, averaged over the matched SK and ST configurations and the three datasets, which improves the aggregation of distributed task-relevant EEG evidence over the full trial. Furthermore, the complete three-stage configuration achieved the highest mean performance for both backbones, supporting the effectiveness of coordinating multiscale extraction, local temporal composition, and trial-level context integration in SATRL.

3.5 Physiological Correspondence Analysis

Time-frequency feature maps. To examine whether the Stage-I branches capture distinct spectral characteristics while retaining class-related rhythms, we analyzed their time-frequency representations. Following EEGNet's hidden-feature analysis (Lawhern et al., 2018), we compared class-versus-rest Morlet power contrasts in the input EEG and Stage-I outputs before spatial convolution. Power was averaged across C3, Cz, and C4 and, for hidden features, all filters; contrasts were then averaged across nine subjects. Figure 3(a) reveals different spectral emphases: the short-

kernel branch exhibits prominent higher-frequency contrasts, while longer-kernel branches emphasize lower-frequency structure. Class-related differences remain visible, including relatively lower left-hand and higher tongue power around 8–12 Hz. These observations support the rationale for multiscale extraction: different temporal supports represent broad spectral information while preserving task-related rhythms, rather than restricting feature extraction to a single temporal scale.

Saliency maps. To characterize when the model relies on EEG evidence during a trial and assess the temporal relevance of local window modeling, we computed trial-wise temporal saliency using absolute Gradient×Input averaged across EEG channels (Shrikumar et al., 2017). Figure 3(b) shows smoothed, subject-wise scaled profiles for Subjects 02, 07, and 08, with trials sorted by mean relevance over 1.5–4.0 s. Relevance is most prominent within the first second after cue onset, broadly coinciding with the early peaks in normalized μ- and β-band ERD strength (Figure 3(c)). Localized relevance also appears between 1 and 4 s, although it is generally less prominent and varies across trials. These patterns support the rationale for local temporal windows designed to capture sustained activity and rhythmic changes as stable local features. The early preference, together with later relevant fragments, motivates extracting local patterns throughout the trial rather than restricting analysis to its initial response. Computational details for both analyses are provided in Appendix B.

3.6 Sensitivity Analysis

We conducted a sensitivity analysis to determine the configurations of MS, ST, and weighting of CDTM. For MS, extracting slow, intermediate, and fast rhythm-related temporal patterns preserves complementary EEG information. Omitting any of these scales leaves part of the temporal pattern space insufficiently modeled, whereas dividing the signal into additional scales provides no clear benefit. For ST, serially connecting convolutional kernels with suitable sizes effectively forms temporal-window representations, while excessive stacking introduces unnecessary nonlinear transformations. For the contextual weighting, performance tends to decline as the weight increases, indicating that overly strong contextual modulation can weaken local task-relevant evidence. These results suggest that SATRL benefits from balanced temporal-scale coverage, controlled local composition, and restrained contextual modulation.

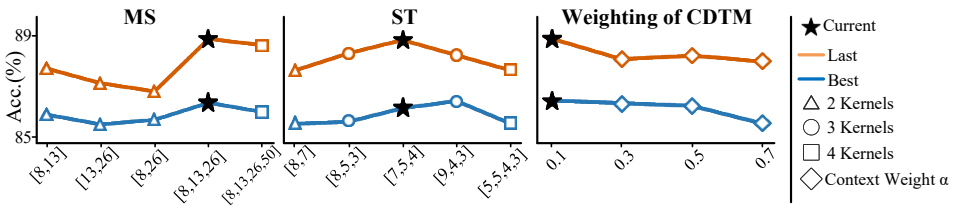


Figure 4: The results of the sensitivity analysis.

4 Conclusion

We introduced Stage-Aligned Temporal Representation Learning (SATRL), a framework for MI-EEG decoding that aligns temporal operators with the evolving requirements of successive representation stages. Its three-stage design connects multiscale rhythm extraction, local temporal aggregation, and context-guided integration to transform EEG recordings into task-relevant trial representations. Experiments on four datasets demonstrate competitive cross-session and leave-one-subject-out performance across compact CNN and CNN–Transformer architectures. Among the tested configurations, the complete three-stage model achieves the highest average best accuracy in the stage ablations for both architecture families. Time–frequency and saliency analyses further reveal complementary spectral responses and temporally localized model relevance. Together, these findings support matching temporal operators to their representation stages as a practical design principle for MI-EEG decoding.

AI USE STATEMENT

In this work, we used generative AI tools for language editing, grammar correction, and assistance with LaTeX formatting. We have not used generative AI tools for generating scientific claims, experimental results, datasets, figures, or quantitative analyses, and authorship decisions, peer review, or fabrication of citations are not applicable to this work. Additionally, we used generative AI tools for polishing manuscript expressions. We have reviewed all AI-assisted work. All AI-assisted text and formatting suggestions were revised and verified by the authors. The authors independently checked the correctness of technical content, equations, citations, code, experimental settings, and reported results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4895–4901, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.298.

Hamdi Altaheri, Ghulam Muhammad, and Mansour Alsulaiman. Physics-informed attention temporal convolutional network for EEG-based motor imagery classification. *IEEE Transactions on Industrial Informatics*, 19(2):2249–2258, 2023. doi: 10.1109/TII.2022.3197419.

Hamdi Altaheri, Fakhri Karray, and Amir-Hossein Karimi. Temporal convolutional transformer for EEG based motor imagery decoding. *Scientific Reports*, 15(1):32959, 2025. doi: 10.1038/s41598-025-16219-7.

Kai Keng Ang, Zheng Yang Chin, Haihong Zhang, and Cuntai Guan. Filter bank common spatial pattern (FBCSP) in brain-computer interface. In *2008 IEEE International Joint Conference on Neural Networks*, pp. 2390–2397, 2008. doi: 10.1109/IJCNN.2008.4634130.

Kai Keng Ang, Zheng Yang Chin, Chuanchu Wang, Cuntai Guan, and Haihong Zhang. Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b. *Frontiers in Neuroscience*, 6:39, 2012. doi: 10.3389/fnins.2012.00039.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

Ulvi Baspinar and Yahya Tastan. Enhancing motor imagery EEG classification in brain–computer interfaces via transformernet. *IEEE Access*, 13:143291–143301, 2025. doi: 10.1109/ACCESS.2025.3594083.

Mouna Bouchane, Wei Guo, and Shuojin Yang. Hybrid CNN-GRU models for improved EEG motor imagery classification. *Sensors*, 25(5):1399, 2025. doi: 10.3390/s25051399.

Clemens Brunner, Robert Leeb, Gernot R. Muller-Putz, Alois Schlogl, and Gert Pfurtscheller. BCI competition 2008–graz data set A: a train/competition set. In *Proceedings of the 4th International IEEE/EMBS Conference on Neural Engineering*, 2008. Dataset description for BCI Competition IV data set 2a.

Ujwal Chaudhary, Niels Birbaumer, and Ander Ramos-Murguialday. Brain-computer interfaces for communication and rehabilitation. *Nature Reviews Neurology*, 12(9):513–525, 2016. doi: 10.1038/nrneurol.2016.113.

Junjian Chen, Zhuliang Yu, Zhenghui Gu, and Yuanqing Li. Deep temporal-spatial feature learning for motor imagery-based brain–computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(11):2356–2366, 2020. doi: 10.1109/TNSRE.2020.3023417.

Arnaud Delorme, Terrence Sejnowski, and Scott Makeig. Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis. *NeuroImage*, 34(4):1443–1449, 2007. doi: 10.1016/j.neuroimage.2006.11.004.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *International Conference on Learning Representations*, 2024.

Albert Gu, Karan Goel, and Christopher Re. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022.

Sepp Hochreiter and Jurgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

Thorir Mar Ingolfsson, Michael Hersche, Xiaying Wang, Nobuaki Kobayashi, Lukas Cavigelli, and Luca Benini. EEG-TCNet: an accurate temporal convolutional network for embedded motor-imagery brain-machine interfaces. In *2020 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 2958–2965, 2020. doi: 10.1109/SMC42975.2020.9283028.

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

Golan Karvat, Maité Crespo-García, Gal Vishne, Michael C. Anderson, and Ayelet N. Landau. Universal rhythmic architecture uncovers two modes of neural dynamics. *Nature Communications*, 17:7024, 2026. doi: 10.1038/s41467-026-73553-8. URL https://doi.org/10.1038/s41467-026-73553-8.

Vernon J. Lawhern, Amelia J. Solon, Nicholas R. Waytowich, Stephen M. Gordon, Chou P. Hung, and Brent J. Lance. EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018. doi: 10.1088/1741-2552/aace8c.

Ke Liu, Tao Yang, Zhuliang Yu, Weibo Yi, Hong Yu, Guoyin Wang, and Wei Wu. MSVTNet: multi-scale vision transformer neural network for EEG-based motor imagery decoding. *IEEE Journal of Biomedical and Health Informatics*, 28(12):7126–7137, 2024. doi: 10.1109/JBHI.2024.3450753.

Scott Makeig, Anthony J. Bell, Tzyy-Ping Jung, and Terrence J. Sejnowski. Independent component analysis of electroencephalographic data. In *Advances in Neural Information Processing Systems*, volume 8, pp. 145–151. MIT Press, 1996.

Ravikiran Mane, Neethu Robinson, A. P. Vinod, Seong-Whan Lee, and Cuntai Guan. A multi-view CNN with novel variance layer for motor imagery brain computer interface. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 2950–2953, 2020. doi: 10.1109/EMBC44109.2020.9175874.

Ravikiran Mane, Effie Chew, Karen Chua, Kai Keng Ang, Neethu Robinson, A. P. Vinod, Seong-Whan Lee, and Cuntai Guan. FBCNet: A multi-view convolutional neural network for brain-computer interface. *arXiv preprint arXiv:2104.01233*, 2021. URL https://arxiv.org/abs/2104.01233.

Chang S. Nam, Yongwoong Jeon, Young-Joo Kim, Insuk Lee, and Kyungkyu Park. Movement imagery-related lateralization of event-related (de)synchronization (ERD/ERS): motor-imagery duration effects. *Clinical Neurophysiology*, 122(3):567–577, 2011. doi: 10.1016/j.clinph.2010.08.002.

Patrick Ofner, Andreas Schwarz, Joana Pereira, and Gernot R. Muller-Putz. Upper limb movements can be decoded from the time-domain of low-frequency EEG. *PLOS ONE*, 12(8):e0182578, 2017. doi: 10.1371/journal.pone.0182578.

Hua Peng and Ping Yan. A novel EEG motor imagery classification approach based on graph signal processing and dimensionality reduction. *Sadhana*, 51(2), 2026. doi: 10.1007/s12046-026-03077-w.

Gert Pfurtscheller and Christa Neuper. Motor imagery activates primary sensorimotor area in humans. *Neuroscience Letters*, 239(2–3):65–68, 1997. doi: 10.1016/S0304-3940(97)00889-6.

Gert Pfurtscheller, Christa Neuper, Herbert Ramoser, and Johannes Muller-Gerking. Visually guided motor imagery activates sensorimotor areas in humans. *Neuroscience Letters*, 269(3):153–156, 1999. doi: 10.1016/S0304-3940(99)00452-8.

Gert Pfurtscheller, Clemens Brunner, Alois Schlogl, and Fernando H. Lopes da Silva. Mu rhythm (de)synchronization and EEG single-trial classification of different motor imagery tasks. *NeuroImage*, 31(1):153–159, 2006. doi: 10.1016/j.neuroimage.2005.12.003.

Herbert Ramoser, Johannes Muller-Gerking, and Gert Pfurtscheller. Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE Transactions on Rehabilitation Engineering*, 8(4):441–446, 2000. doi: 10.1109/86.895946.

Aurora Saibene, Hafez Ghaemi, and Eda Dagdevir. Deep learning in motor imagery EEG signal decoding: A systematic review. *Neurocomputing*, 610:128577, 2024. doi: 10.1016/j.neucom.2024.128577.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, 38(11):5391–5420, 2017. doi: 10.1002/hbm.23730.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3145–3153. PMLR, 2017.

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, 2023. doi: 10.1109/TNSRE.2022.3230250.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017.

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 *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=NPNUHgHF2w.

Xianheng Wang, Veronica Liesaputra, Zhaobin Liu, Yi Wang, and Zhiyi Huang. An in-depth survey on deep learning-based motor imagery electroencephalogram (EEG) classification. *Artificial Intelligence in Medicine*, 147:102738, 2024. doi: 10.1016/j.artmed.2023.102738.

Jonathan R. Wolpaw, Niels Birbaumer, Dennis J. McFarland, Gert Pfurtscheller, and Theresa M. Vaughan. Brain-computer interfaces for communication and control. *Clinical Neurophysiology*, 113(6):767–791, 2002. doi: 10.1016/S1388-2457(02)00057-3.

Yingjie Yang, Xiu Zhang, Xin Zhang, and Changyi Yu. MCMTNet: advanced network architectures for EEG-based motor imagery classification. *Neurocomputing*, 620:129255, 2025. doi: 10.1016/j.neucom.2024.129255.

Wei Zhao, Xiaolu Jiang, Baocan Zhang, Shixiao Xiao, and Sujun Weng. CTNet: a convolutional transformer network for EEG-based motor imagery classification. *Scientific Reports*, 14 (1):20237, 2024. doi: 10.1038/s41598-024-71118-7.

Wei Zhao, Baocan Zhang, Haifeng Zhou, Dezhi Wei, Chenxi Huang, and Quan Lan. Multi-scale convolutional transformer network for motor imagery brain-computer interface. *Scientific Reports*, 15(1), 2025. doi: 10.1038/s41598-025-96611-5.

A Related Work

EEG Decoding. End-to-end EEG decoding organizes feature extraction and temporal integration across successive stages to learn task-relevant representations. Convolutional models establish temporal and spatial feature extraction stages with parameter-efficient designs (Schirrmeister et al., 2017; Lawhern et al., 2018), while recurrent, temporal-convolutional, and attention-based architectures introduce subsequent stages for dependency modeling (Ingolfsson et al., 2020; Bouchane et al., 2025; Song et al., 2023; Altaheri et al., 2023; 2025). Multi-scale hybrids enrich feature extraction stages with complementary temporal patterns (Liu et al., 2024; Zhao et al., 2025), whereas graph-based methods explicitly model inter-channel relationships during spatial representation learning (Peng & Yan, 2026). More recently, EEG foundation models have enabled transferable representation learning through large-scale self-supervised pretraining (Jiang et al., 2024; Wang et al., 2025). These advances enrich the modules and architectural structures available for EEG decoding and strengthen spatial-temporal feature learning. Nevertheless, it still remains challenging that maintaining high decoding accuracy and reliable generalization across sessions and subjects. The suitability of temporal operators for different representation stages remains insufficiently understood. SATRL is motivated by the need to preserve complementary multi-scale temporal information, consolidate local temporal patterns, and integrate distributed evidence for EEG decoding.

Temporal Modeling in Time Series. Time Series Models offer complementary mechanisms for extracting local patterns and modeling dependencies over broader temporal contexts. Recurrent networks summarize temporal history through hidden-state updates (Hochreiter & Schmidhuber, 1997), whereas temporal convolutional networks extract local patterns and expand temporal coverage through stacking and dilation (Bai et al., 2018). Self-attention directly connects distant positions through input-dependent interactions (Vaswani et al., 2017). Structured and selective state-space models support efficient long-sequence processing through learned state dynamics, with selective mechanisms controlling which information is retained or propagated (Gu et al., 2022; Gu & Dao, 2024). These approaches provide different solutions for local feature extraction, temporal memory, and long-range interaction, rather than simply increasing network depth or receptive field size. These approaches inform temporal module design for EEG decoding, highlighting the need to preserve localized rhythm-related activity while integrating information distributed across a trial.

Hierarchical Representation Learning. Stage-wise EEG decoders progressively transform local EEG features into trial-level representations through successive feature transformations. Early stages combine temporal filtering and spatial integration to extract rhythm-related features from multichannel recordings (Schirrmeister et al., 2017; Lawhern et al., 2018). Later stages build on these intermediate features by applying sequence modeling or attention mechanisms to organize broader temporal dependencies (Ingolfsson et al., 2020; Altaheri et al., 2023; Bouchane et al., 2025; Altaheri et al., 2025). Statistical aggregation provides another route for summarizing temporal feature sequences into compact classification descriptors, as exemplified by the temporal variance pooling of FBCNet (Mane et al., 2021). These designs establish a hierarchy in which temporal information is progressively preserved, composed, and summarized as EEG features develop into trial-level evidence. Effective temporal modeling thus depends not only on the capability of individual operators, but also on their placement and coordination across representation stages. SATRL focuses on aligning temporal operators with this hierarchical organization.

B Interpretability Analysis

B.1 Time-frequency feature analysis

Principle. Following EEGNet's hidden-feature analysis (Lawhern et al., 2018), we characterize internal signals through their time-resolved spectral power. Class contrasts reveal which rhythmic differences remain represented after temporal filtering. This is a representation-level analysis, not an attribution of individual frequencies to the classification decision.

Computation. We extracted the outputs of temporal convolution and batch normalization before spatial convolution. For each trial i and branch k, feature power was

$$P_{i,k}(f,t) = \frac{1}{J|\mathcal{C}|} \sum_{j=1}^{J} \sum_{c \in \mathcal{C}} \left| \mathcal{W}_f\{H_{i,k,j,c}\}(t) \right|^2, \tag{11}$$

where $J = 32$, $\mathcal{C} = \{\mathrm{C3}, \mathrm{Cz}, \mathrm{C4}\}$, and $\mathcal{W}_f$ denotes the Morlet transform. We used five-cycle, zero-mean wavelets at $4, 6, \ldots, 40$ Hz, retaining time points every 0.05 s. Power was computed before averaging across filters and electrode positions.

For subject s and ground-truth class q, we calculated

$$\Delta P_{s,k,q} = \overline{P}_{s,k,q} - \frac{1}{3} \sum_{r \neq q} \overline{P}_{s,k,r}, \tag{12}$$

where each $\overline{P}$ is a within-class trial mean. The displayed maps average these contrasts equally across nine subjects, each with 72 test trials per class. Input EEG was processed analogously, without filter averaging. We excluded boundary regions using the wavelet's five-standard-deviation half-support plus the temporal kernel's approximate half-support. Color scales were set independently by row. These descriptive contrasts are neither statistical significance maps nor baseline-referenced ERD/ERS.

B.2 Temporal saliency analysis

Principle. Gradient-based attribution measures local output sensitivity. Under a first-order approximation, $\Delta z_{\hat{y}} \approx \nabla_x z_{\hat{y}}^\top \Delta x$. Gradient×Input weights this sensitivity by the observed input (Shrikumar et al., 2017). Its magnitude indicates local input-weighted sensitivity, not the causal effect of removing an entire interval.

Computation. For each trial, we targeted the predicted-class logit $z_{\hat{y}_i}$ and computed

$$R_i(t) = \frac{1}{22} \sum_{c=1}^{22} \left| x_{i,c,t} \frac{\partial z_{\hat{y}_i}(x_i)}{\partial x_{i,c,t}} \right|. \tag{13}$$

Profiles were Gaussian-smoothed with $\sigma = 40$ ms. For visualization, we subtracted the global median within each subject, divided by the interquartile range, and clipped the resulting values to $[-3, 3]$. All 288 test trials were retained and sorted by descending mean relevance over 1.5–4.0 s.

Red therefore indicates greater absolute attribution, not necessarily a positive contribution. The concentration of later high-relevance trials near the top follows from the sorting criterion. These maps neither isolate Stage-II nor establish temporal correspondence with μ/β ERD; the latter requires a separate alignment analysis.

Temporal Module Configurations. Stage-I uses $M = 3$ parallel temporal convolutional branches with kernel lengths $\{8, 13, 26\}$. Stage-II employs $L = 3$ serial depthwise temporal convolutions with kernel lengths $\{7, 5, 4\}$, unit stride and dilation, and length-preserving padding. The resulting receptive field is $R_L = 1 + \sum_{\ell=1}^{L} (s_\ell - 1) = 14$ temporal steps in the transformed feature sequence. In Stage-III, the trial-level context weight is set to $\alpha = 0.1$. For the TDC variant evaluated in the Stage-II ablation, the offset-prediction layer is initialized to produce zero temporal displacements, so adaptive sampling initially follows the regular convolution grid.