RCT-MNM: Region-Channel-Temporal Masked Neural Modeling for Emotion and Psychiatric Disorder Analysis

Anonymous Authors[1]

Abstract

We propose Region-Channel-Temporal Masked Neural Modeling (RCT-MNM), a self-supervised learning framework for EEG and sEEG-based emotion recognition and psychiatric disorder analysis. RCT-MNM integrates region-channel-temporal self-attention to model spatial-temporal dependencies and conditioned self-supervised learning for robust representation alignment. A post-pretraining refinement step further enhances task performance. Experimental results on multiple EEG and sEEG datasets demonstrate state-of-the-art performance in emotion recognition and task engagement classification. Additionally, RCT-MNM provides region- and channel-level interpretability, revealing meaningful neural dynamics for cognitive health applications. Our framework advances neural representation learning, enhancing the generalizability and interpretability of EEG/sEEG-based models for brain signal decoding and therapeutic insights.

1. Introduction

Psychiatric and mood disorders such as depression and anxiety, along with cognitive impairments, impact millions of individuals worldwide, significantly affecting mental well-being and cognitive function (Organization, 2001). These conditions are often difficult to diagnose and treat due to their complex neurophysiological underpinnings and reliance on subjective assessments (Kret & Ploeger, 2015; Goschke, 2014). As a result, there is a growing interest in leveraging neurophysiological signals, to develop objective and data-driven diagnostic tools for emotion recognition and cognitive state analysis (Houssein et al., 2022). Understanding the neural basis of these disorders can lead to improved treatment strategies, brain-computer interfaces (BCIs), and personalized mental health interventions (Ng & Weisz, 2016).

Emotion recognition and cognitive analysis from electroencephalography (EEG) and stereo-EEG (sEEG) are essential for understanding human affective states and psychiatric disorders (Drane et al., 2021). However, these neural signals exhibit complex spatio-temporal dependencies, posing significant challenges for machine learning models in achieving robust and accurate classification (Gao et al., 2019). Traditional approaches rely on handcrafted features or shallow models, which fail to capture the intricate interactions across spatial and temporal scales (Rouast et al., 2019; Liu et al., 2024). Recent advancements in deep learning, particularly transformer-based models, have shown promise in neural representation learning in global level (Vaswani et al., 2017; Dosovitskiy et al., 2021; He et al., 2022; Assran et al., 2023).

Despite advancements in EEG/sEEG neural modeling, several challenges remain. (1) Hierarchical Spatial Representation: Existing models overlook the structured organization from electrodes to brain regions, limiting spatial feature extraction (Lachaux et al., 2003; Gevins et al., 1995). (2) Long-Term Temporal Dependencies: Conventional models struggle to capture sustained neural patterns essential for emotion and cognitive analysis. (3) Structured Attention Mechanisms: Global self-attention for spatial representation learning is computationally expensive, requiring more efficient structured attention. (4) Generalization Across Datasets: High inter-subject variability hinders model robustness (Saha & Baumert, 2020). Self-supervised learning (SSL) can improve feature extraction without labeled data (Rafiei et al., 2024; He et al., 2022; Selvaraju et al., 2017; Jiang et al., 2024; Wang et al., 2024). Addressing these challenges requires a model that integrates spatio-temporal dependencies, regional interactions, and efficient SSL strategies for improved EEG-based emotion and cognitive analysis.

We introduce **Region-Channel-Temporal Masked Neural Modeling (RCT-MNM)**, a self-supervised learning framework that leverages hierarchical attention mechanisms to capture multi-scale dependencies in EEG and sEEG signals. Our contributions are as follows:

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

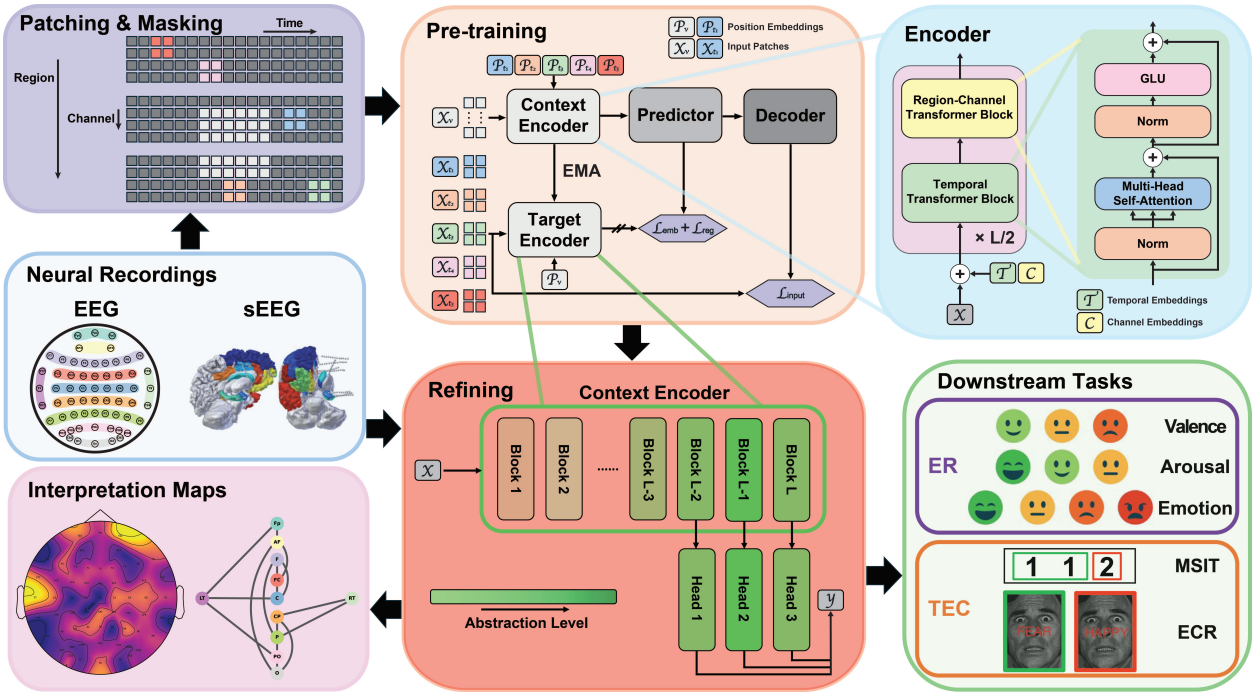

Figure 1. **RCT-MNM.** The model consists of three stages: Patching & Masking, segmenting and masking EEG/sEEG signals; Pre-training, learning region-channel-temporal representations; and Refining, fine-tuning the Encoder for improved downstream performance. It supports emotion recognition (ER) and task engagement classification (TEC) with interpretable neural representations.

1. We propose a novel Temporal-Region-Channel encoder that captures hierarchical neural dependencies in EEG and sEEG data through a structured transformer-based approach.

2. We introduce a Region-Channel Self-Attention Mechanism (RCSA) mechanism that models functional interactions between brain regions and electrode channels, enabling structured spatial dependency modeling.

3. We develop a region-conditioned self-supervised learning paradigm that employs Region-Channel-Temporal masking for representation alignment and input reconcstruction to enhance model generalizability across diverse EEG/sEEG tasks.

4. We introduce a post-pre-training Adaptive Attention-Based Layer Refinement step that fine-tunes intermediate representations using to improve task-specific classification performance.

5. Our model achieves state-of-the-art classification performance for emotion recognition and task engagement classification in subject-dependent and subject-independent settings.

6. We enhance interpretability by generating region- and channel-level activation maps and self-attention graphs, providing insights into functional brain dynamics for affective and psychiatric disorder analysis.

2. Related Work

Graph-Based Models for EEG Analysis. Graph-based models have been widely explored for EEG-based emotion recognition and cognitive state classification. Dynamic Graph Convolutional Neural Networks (DGCNN) dynamically construct EEG graph representations and apply graph convolution to model spatial dependencies among electrode channels (Song et al., 2018). Regularized Graph Neural Networks (RGNN) integrate a neuroscience-inspired adjacency matrix and employ node-wise domainadversarial training (NodeDAT) for cross-subject adaptation, and emotion-aware distribution learning (EmotionDL) for noise robustness in EEG-based emotion recognition. (Zhong et al., 2020). Local-Global Graph Networks (LGG-Net) extend this approach by leveraging local and global graph representations, enhancing the model's ability to capture fine-grained spatial structures (Ding et al., 2023). HDGCN employs dual-branch hierarchical dynamic graph convolution to capture global and local EEG spatial patterns, integrating a layerwise adjacency matrix and an auxiliary information module for enhanced feature fusion and interpretability (Ye et al., 2022). However, these models exhibit limited

spatio-temporal integration and generalizability in spatial representations.

Temporal Modeling for Neural Signals. Temporal dependencies in neural signals have been modeled using convolutional and transformer-based architectures. TSception employs multi-scale 1D convolutional kernels to capture both short- and long-term temporal dependencies while leveraging spatial asymmetry in EEG signals (Ding et al., 2022). EEG Conformer introduces a hybrid architecture combining convolution and self-attention to capture local and global temporal dependencies in EEG data (Song et al., 2022). Despite their effectiveness, convolution-based methods struggle with long-range dependencies, and temporal self-attention itself does not explicitly encode inductive biases about EEG/sEEG signal structures, which can make training inefficient and lead to overfitting on small datasets.

Transformer-Based Self-Supervised Learning for EEG. As self-supervised learning has gained traction in computer vision, there is increasing interest in adapting these techniques to the neural domain, where the inherent spatio-temporal structure of EEG signals presents unique challenges for representation learning. Masked Autoencoders (MAEEG) perform masked input reconstruction to learn robust EEG representations (Chien et al., 2022). MMM unifies EEG channel topologies and employs multi-dimensional position encoding, hierarchical channel representations, and multi-stage pre-training to enhance cross-dataset generalization for EEG-based emotion recognition (Yi et al., 2024). LaBraM introduces a biologically inspired approach to self-supervised EEG representation learning, leveraging latent brain dynamics modeling to capture underlying neural processes across diverse cognitive states (Jiang et al., 2024). EEGPT extends large-scale pre-training techniques to EEG analysis, leveraging a self-supervised dual pretraining strategy that combines spatio-temporal representation alignment and masked reconstruction (Wang et al., 2024). SEEGnificant framework tokenizes sEEG signals with convolutions, applies temporal and spatial self-attention, integrates electrode locations, and employs subject-specific heads for multi-subject behavioral decoding and cross-subject transfer learning (Mentzelopoulos et al., 2024). However, they either lack hierarchical spatial modeling or fail to effectively integrate long-range temporal dependencies in EEG and sEEG signals.

3. Methodology

3.1. Problem Setting and Formulation

We represent the neural data (EEG/sEEG) as $\mathbf{X}_{\text{input}} \in \mathbb{R}^{C\times T}$, where C is the number of channels, and T is the number of time points. The problem of task engagement classification (TEC) and low-high valence/arousal classification is formulated as binary classification tasks predicting rest-stage versus task-stage and low versus high valence/arousal with labels $y \in \{0, 1\}$. The problem of valence recognition is formulated as a multi-class classification task predicting different types of emotions with labels $y \in \{0, ..., E-1\}$, where E is the number of types of emotions. The goal for these tasks is to predict the label y based on a given neural input $\mathbf{X}_{\text{input}}$.

We first segment the data into non-overlapping patches. The input patch representation is defined as: $\mathbf{X}_{\text{input}} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}, \quad \mathbf{x}_n \in \mathbb{R}^{1\times P}$, where P is the temporal patch size and $N = \frac{CT}{P}$ is the number of patches. Each patch undergoes a linear projection, such that $\mathbf{X} = \mathbf{X}_{\text{input}} W_P$, where $\mathbf{X} \in \mathbb{R}^{N\times d}$, and d is the embedding size.

3.2. Temporal-Region-Channel Encoder

The Temporal-Region-Channel Encoder is designed to capture the intricate spatio-temporal dependencies in EEG and sEEG data through a hierarchical transformer-based architecture. It consists of two primary components: the Region-Channel Transformer Block and the Temporal Transformer Block, which work together to extract generalized neural representations.

The Region-Channel Transformer Block models spatial dependencies across EEG channels by applying a multi-head self-attention mechanism across the region-channel space. To construct region-level representations, we add extra region tokens to the channels tokens and use Region-Channel Self-Attention mechanism (**RCSA**) to learn region representations from channel-level information based on pre-defined functional region-channel map and region-channel attention mask. The region-channel maps and corresponding attention masks are defined based on the localtion of electrodes on function areas for EEG (Alarcao & Fonseca, 2017), and using electrode labeling algorithm (ELA) for sEEG (Peled et al., 2017). RCSA is a pivotal element in capturing both localized and global interactions within neural data, which especially captures the functional connectivity and coordination among neural patches from distinct brain regions. These regional-channel dynamics are essential for understanding higher-order brain functions, such as task execution or cognitive processing, which involve communication across spatially distributed networks.

Given an input representation $\mathbf{X}$ consisting of channel tokens $\mathbf{X}_c \in \mathbb{R}^{C\times d}$, we append randomly initialized region tokens $\mathbf{X}_r \in \mathbb{R}^{R\times d}$, forming the combined input: $\mathbf{X}_{cr} = [\mathbf{X}_c; \mathbf{X}_r] \in \mathbb{R}^{(C+R)\times d}$, where R is the number of regions. We define a binary region-channel attention mask $M_{\text{rc}} \in \{0, 1\}^{(C+R)\times(C+R)}$ based on the functional region-channel map, ensuring that each region token only attends to its corresponding channels and other regions, and each channel only attends its corresponding region and other channels

within the same region. The region-channel self-attention mechanism (**RCSA**) operates as follows:

$$\begin{aligned} &\mathbf{Q} = \mathbf{X}_{cr} W_Q, \quad \mathbf{K} = \mathbf{X}_{cr} W_K, \quad \mathbf{V} = \mathbf{X}_{cr} W_V \\ &A_{\text{rc}} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}} \odot M_{\text{rc}}\right) \\ &\mathbf{X}'_{cr} = A_{\text{rc}} \mathbf{V} \end{aligned} \tag{1}$$

where W_Q, W_K, W_V are linear projection matrices for Query, Key, and Value, and A_{rc} represents the region-channel attention matrix, capturing dependencies across neural channels and regions, and $\odot$ denotes element-wise multiplication.

These region-channel tokens further undergo an additional feedforward network to enhance spatial modeling (Shazeer, 2020):

$$\begin{aligned} \mathbf{X}''_{cr} &= \text{FFN}_{\text{SwiGLU}}(\mathbf{X}'_{cr}) \\ &= (\text{Swish}_1(\mathbf{X}'_{cr} W) \odot \mathbf{X}'_{cr} V) W_2 \end{aligned} \tag{2}$$

where $\text{Swish}_\beta(x) = x\sigma(\beta x)$. By the adding residual connections, the Region-Channel Transformer Block can be represented as:

$$\begin{aligned} \mathbf{X}'_{cr} &= \text{RCSA}(\text{Norm}(\mathbf{X}_{cr})) + \mathbf{X}_{cr} \\ \mathbf{X}''_{cr} &= \text{FFN}_{\text{SwiGLU}}(\text{Norm}(\mathbf{X}'_{cr})) + \mathbf{X}'_{cr} \end{aligned} \tag{3}$$

To model long-range dependencies in neural activity, we apply a Temporal Transformer Block, which operates across the time dimension while preserving spatial information. The attention mechanism is a standard temporal self-attention (TSA), and the same feedforward network is used for feature mixing. Therefore, the Temporal Transformer Block can be formulated as:

$$\begin{aligned} \mathbf{X}'_t &= \text{TSA}(\text{Norm}(\mathbf{X}_t)) + \mathbf{X}_t \\ \mathbf{X}''_t &= \text{FFN}_{\text{SwiGLU}}(\text{Norm}(\mathbf{X}'_t)) + \mathbf{X}'_t \end{aligned} \tag{4}$$

where $\mathbf{X}_t \in \mathbb{R}^{S \times d}$, and $S = \frac{T}{P}$ is the number of time segments.

We construct the Encoder in a two-stage hierarchical manner. First, we apply Temporal Transformer Block to model long-range dependencies across time, ensuring that dynamic temporal patterns are well-represented. Then, we employ the Region-Channel Transformer Block to integrate spatial dependencies within and across functional brain regions. This design choice is motivated by the fundamental structure of neural signals, where temporal dynamics exhibit continuity and correlation over time, while spatial interactions are constrained by anatomical and functional organization. By first encoding temporal dependencies, we allow the model to learn stable temporal representations before aggregating region-level spatial information. This sequential processing also improves computational efficiency by reducing spatial redundancy in raw channel-wise inputs. The Predictor is also in the same structure as the Encoder, and the Decoder is constructed in a Region-Channel-then-Temporal Transformer Block manner for reversed reconstruction.

To enhance the representation of temporal and spatial dependencies, we incorporate Rotary Position Embeddings (RoPE) (Su et al., 2024) for the temporal dimension and learnable spatial embeddings (Dosovitskiy et al., 2021) for both channel and region representations. This design choice is driven by the strong inductive bias introduced by RCSA, which enforces structured connectivity constraints between channels and brain regions.

3.3. Region-Conditioned Masked Neural Modeling

Our proposed framework utilizes Masked Neural Modeling (MNM) as an input-representation predictive architecture for self-supervised learning of EEG and sEEG patches. The model consists of three core components: a context encoder $f(\cdot;\theta)$, a target encoder $f(\cdot;\tilde{\theta})$, and a predictor $g(\cdot;\psi)$, which collectively learn to align target representations. Additionally, a decoder $h(\cdot;\phi)$ is applied for input reconstruction, ensuring robust feature generalizations across spatio-temporal dimensions. We further introduce in detials about our proposed masking strategy **RCT-Mask**, **Regional Conditioning** for guiding the learning process, and **MNM-Loss** function to improve generalization of neural representations.

3.3.1. Region-Channel-Temporal Masking

Our pre-training framework is designed to predict neural inputs $\mathbf{X}$ and representations $\mathbf{Z}$ across multiple target blocks based on the representation of a context block. In training, the model learns to predict representations of other sections of the neural data within the latent space based on a single observation. However, random sampling of targets, as seen in methods like MAE, may lead the model to rely on simplistic shortcuts, such as interpolating time series or defaulting to more prevalent patterns in the data, potentially reducing its generalizability (Assran et al., 2023). It is important to consider that neural patches exhibit spatial variability based on functional brain organization, as well as temporal variability related to brain states and task conditions. Additionally, complex nonlinear relationships among brain networks add layers of interaction across different brain patches, further challenging the model's capacity to learn robust, generalized representations.

The Region-Channel-Temporal Masking (**RCT-Mask**) approach is a foundational component of our pre-training framework. It is designed to encourage the model to learn robust and generalized representations from neural data by masking signals selectively across temporal, and localized/distributed spatial domains, either independently or in

combination. This mechanism prevents the model from relying on simplistic patterns or interpolation, driving it to uncover deeper, nonlinear relationships.

For a given EEG/sEEG input $\mathbf{X} \in \mathbb{R}^{N \times d}$, where $N = C \times S$, and C is the number of channels, and S is the number of time sequences, the framework identifies a context block $\mathbf{x}_c$. This context block is randomly selected from the input neural segments within the range $\{\rho_c, \eta_c\}$, where ρ_c and η_c are the range ratio for the context block along the spatial and temporal dimensions respectively. The corresponding binary mask for the context block can be defined as $\mathbf{M}_c \in \{0,1\}^{C \times S}$. The context block serves as the observation basis for predicting masked neural representations of other segments, termed target blocks $\mathbf{x}_t$, with $\mathbf{x}_c, \mathbf{x}_t \in \mathbf{X}$ and $\mathbf{x}_c \cap \mathbf{x}_t = \varnothing$. The objective of the model is to predict $\mathbf{x}_t$ and $\mathbf{z}_t$ from $\mathbf{x}_c$, where $\mathbf{z}$ represents the latent representations derived from the encoders, and $\mathbf{z}_t$ is the latent representations of the target blocks.

We categorize the remaining patches excluding the context block into five distinct and non-overlapping domains: cross-time, cross-channel, cross-time-channel, cross-region, cross-region-channel. We randomly sample multiple blocks from each of the five domains as targets blocks, with binary masks $\mathbf{M} \in \{0,1\}^{C \times S}$ corresponding to each domain as $\mathbf{M}^t$, $\mathbf{M}^c$, $\mathbf{M}^{ct}$, $\mathbf{M}^r$, and $\mathbf{M}^{rt}$, within the range $\{\rho_t, \eta_t\}$, where ρ_t and η_t are the range ratio for the target blocks along the spatial and temporal dimensions respectively. Therefore, the RCT-Mask $\mathbf{M}_{\text{RCT}}$, context block, and target blocks can be written as:

$$\begin{aligned} \mathbf{M}_{\text{RCT}} &= \mathbf{M}^t \cup \mathbf{M}^c \cup \mathbf{M}^{ct} \cup \mathbf{M}^r \cup \mathbf{M}^{rt} \\ \mathbf{x}_c &= \mathbf{X} \odot \mathbf{M}_c \\ \mathbf{x}_t &= \mathbf{X} \odot \mathbf{M}_{\text{RCT}} \end{aligned} \tag{5}$$

3.3.2. Regional Conditioning

Given a tokenized input representation $\mathbf{x} \in \mathbb{R}^{N \times d}$ and corresponding positional embeddings $\mathbf{p} \in \mathbb{R}^{N \times d}$, where N is the number of tokens, we define a self-supervised learning framework based on context-target prediction and reconstruction. Let c denote the set of indices corresponding to the context block, such that the context tokens and positional embeddings are represented as $\mathbf{x}_c$ and $\mathbf{p}_c$. Similarly, let t represents target blocks with target tokens and positional embeddings $\mathbf{x}_t$ and $\mathbf{p}_t$. The encoder function $f(\cdot;\theta)$ maps the context input to latent representations: $\mathbf{z}_c = f(\mathbf{x}_c;\theta)$, where θ denotes the encoder parameters. The learned context representations $\mathbf{z}_c$ are then used to predict the representations of the masked target tokens $\mathbf{z}_t = f(\mathbf{x};\tilde{\theta})$, where $\tilde{\theta}$ represents an exponential moving average (EMA) of the encoder weights θ to stabilize training. A predictor function $g(\cdot;\psi)$ further refines the predictions by incorporating both the context representations $\mathbf{z}_c$ and the positional embeddings of the target tokens $\mathbf{p}_t$, where ψ are the learnable parameters of the predictor function. The decoder function $h(\cdot;\phi)$ finally reconstruct the input target tokens using predicted target representations $\hat{\mathbf{z}}_t$ and target positional embeddings, such that:

$$\hat{\mathbf{z}}_t = g(\mathbf{z}_c, \mathbf{p}_t; \psi), \quad \hat{\mathbf{x}}_t = h(\hat{\mathbf{z}}_t, \mathbf{p}_t; \phi) \tag{6}$$

where $\hat{\mathbf{z}}_t$ is the predicted target representation, $\hat{\mathbf{x}}_t$ is the reconstructed target input, and ϕ are the learnable parameters of the decoder function. The encoder function f, predictor function g, and decoder function h are implemented as transformer blocks.

Our proposed methodology also incorporates explicit regional conditioning into the encoder during pre-training and inference to enhance the self-supervised learning process. Specifically, the regional conditioning is achieved by appending region-specific positional embeddings as additional tokens in the input sequence processed by the transformer blocks. This formulation can be expressed as follows:

$$\mathbf{z}_c^t = f(\mathbf{x}_c, \mathbf{p}_t^{r_t}; \theta), \quad \mathbf{z}_t^c = f(\mathbf{x}, \mathbf{p}_c^{r_c}; \tilde{\theta}) \tag{7}$$

where $\mathbf{z}_c^t$ represents the context embedding conditioned on other regional targets, while $\mathbf{z}_t^c$ denotes the target representation conditioned on the regional contexts, and $\mathbf{p}_t^{r_t}$ and $\mathbf{p}_c^{r_c}$ are the region positional embedding for regions r_t and r_c, which are the unique regions existed in target and context blocks. Therefore, the each region positional embedding is only used once as an additional token, and all unique tokens are appended to the input tokens.

EEG and sEEG datasets often exhibit variations in the number of channels, sampling rates, and recording durations. Explicit temporal and channel-wise conditioning may require recalibration for different settings and reduce generalizability. By conditioning only on region-level embeddings, our model maintains flexibility across diverse datasets without requiring channel- and temporal-level adjustments. Furthermore, since each region embedding is only used once as an additional token in the input sequence, rather than being repeatedly appended per time segment or channel, the model avoids excessive computational overhead. This allows for faster training and inference while retaining critical distributed spatial dependencies.

3.3.3. Masked Neural Modeling Loss

The loss function for our framework (**MNM-Loss**) is a combination of the input reconstruction loss, the representation alignment loss, and the VICReg-inspired regularization terms, structured as follows:

$$\mathcal{L}_{\text{MNM}} = \lambda_1 \mathcal{L}_{\text{input}} + \lambda_2 \mathcal{L}_{\text{rep}} + \lambda_3 \mathcal{L}_{\text{reg}} \tag{8}$$

Input Reconstruction & Representation Alignment Loss. The MNM objective is optimized by the average l_2 distance

between the predicted patch-level representations/inputs of the target blocks $\{\hat{\mathbf{z}}_t, \hat{\mathbf{x}}_t\}$ and the ground-truth target representations/inputs $\{\mathbf{z}_t, \mathbf{x}_t\}$, which is formulated as the Input Reconstruction and Representation Alignment Loss:

$$\mathcal{L}_{\text{input}} = ||\hat{\mathbf{x}}_t - \mathbf{x}_t||_2^2 \quad \mathcal{L}_{\text{rep}} = ||\hat{\mathbf{z}}_t - \mathbf{z}_t||_2^2 \tag{9}$$

VIC Regularization. To further improve representation learning, we incorporate Variance-Invariance-Covariance Regularization (VICReg) (Bardes et al., 2022), which encourages decorrelated feature representations, preventing collapse into trivial solutions. VICReg introduces two key constraints: (1) Variance Regularization is a hinge function on the standard deviation of the representation $\mathbf{z}$ along the batch dimension, ensuring that all dimensions of the learned embedding space exhibit meaningful variation, enhancing feature diversity. (2) Covariance Regularization minimizes off-diagonal elements of the covariance matrix along the batch dimension, promoting statistical independence between learned features. It can be represented in a batch-wise format as:

$$\begin{aligned}
\mathcal{L}_{\text{reg}} &= \lambda_{\text{var}} * \mathcal{L}_{\text{var}} + \lambda_{\text{cov}} * \mathcal{L}_{\text{cov}} \\
\mathcal{L}_{\text{var}} &= \frac{1}{d}\sum_{j=1}^{d} \max(0, \gamma - \sqrt{\text{Var}(\mathbf{z}^j) + \epsilon}) \\
\mathcal{L}_{\text{cov}} &= \frac{1}{d}\sum_{i \neq j} [\frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{z}_i - \bar{\mathbf{z}})^T]_{i,j}^2
\end{aligned} \tag{10}$$

where $\bar{\mathbf{z}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_i$. $\mathbf{z}_i \in \mathbb{R}^d$ is the i^{th} vector in this batch, and $\mathbf{z}^j$ is the vector composed of each value at dimension j in all vectors in the batch. By enforcing these constraints, our method ensures that learned features remain informative, diverse, and uncorrelated, ultimately improving the generalization of the learned neural representations.

3.4. Adaptive Attention-Based Layer Refinement

The last layer of the pre-trained encoder might achieve low representation quality because of spliting the encoding capability for predicting and decoding. To enhance the effectiveness of our pre-trained encoder for downstream tasks, we introduce a post-pre-training refinement process, which fine-tunes intermediate representations rather than relying solely on the final encoder layer. This approach leverages an ensemble of multi-layer perceptron (MLP) heads attached to multiple layers near the end of the encoder, allowing the model to extract more abstract feature representations. By distributing refinement across multiple layers, the model improves adaptability and generalization in downstream applications. Unlike conventional methods that depend exclusively on the last encoder layer for feature extraction, our refinement strategy aggregates embeddings from multiple late-stage layers, mitigating potential limitations of a single late layer representation.

Given an input $\mathbf{x}$, let $f_l(\cdot;\theta)$ be the function of l consecutive blocks in the encoder, where the layer-wise representations are computed as:

$$\mathbf{z}_l = f_l(\mathbf{x};\theta), \quad l \in \{1, \ldots, L\} \tag{11}$$

where θ represents the encoder weights, and L is the total number of encoder layers. To refine the feature representations for downstream tasks, we aggregate representations from the final k layers of the encoder. We introduce **Adaptive Attention-Based Layer Refinement** to dynamically select important layers. For each layer, we can compute its attention score a_l as:

$$a_l = \text{softmax}\left(\frac{\mathbf{q}^\top W_q \text{Norm}(\mathbf{z}_l)}{\sqrt{d}}\right) \tag{12}$$

where W_q is a learnable query projection matrix, and $\mathbf{q}$ is a global learnable query vector. The final refined embedding $\mathbf{z}_{\text{ref}}$ is then adaptively aggregated over k layers as:

$$\mathbf{z}_{\text{ref}} = \sum_{l=L-k+1}^{L} \alpha_l \text{Norm}(\mathbf{z}_l) \tag{13}$$

which can be further attached to a classification head for downstream tasks. This adaptive refinement mechanism ensures that informative layers are prioritized while redundant layers are suppressed. The resulting refined embeddings provide a more stable and contextually meaningful representation for downstream classification, ensuring improved performance across neural cognitive and affective tasks.

4. Experiments

4.1. Experimental Setup

Datasets. We used three publicly available datasets for the EEG emotion recognition (ER) tasks: **SEED** (Zheng & Lu, 2015) for three-class valence classification; **SEED-IV** (Zheng et al., 2018) for four-class valence classification; and **DEAP** (Koelstra et al., 2011) for low-high valence/arousal binary classification. We used two datasets MSIT and ECR (Provenza et al., 2019) for the sEEG task engagement classification (TEC). We employed Subject-Dependent (SD) and Subject-Independent (SI) paradigms for SEED, SEED-IV, and DEAP, and we implemented SD paradigm for MSIT and ECR. A detailed description of datasets can be found in Appendix B.

Baselines. To evaluate classification performance on the aforementioned datasets, we implemented three key baselines using transformer architectures on all datasets: **ViT** (Dosovitskiy et al., 2021) without self-supervision; **MAE** (He et al., 2022) with masked self-supervision on the input space; and **I-JEPA** (Assran et al., 2023) with self-supervision on the representation space. Beyond that, we

Table 1. Classification performance on SEED and SEED-IV (Accuracy/F1 Score (%)). SD: Subject Dependent; SI: Subject Independent. **BOLD**: the best performance are bolded. UNDERLINE: the second-best performance.

MODEL	SSL	T	C	R	SEED		SEED-IV	
					SD	SI	SD	SI
DGCNN	×	×	√	×	75.62/76.19	72.17/71.10	68.89/68.17	67.42/67.74
RGNN	×	×	√	×	77.26/78.32	73.20/73.12	69.10/69.24	67.34/68.38
TSCEPTION	×	×	×	√	74.65/76.02	72.86/74.47	66.83/65.94	69.86/70.08
EEG CONFORMER	×	√	×	×	76.07/76.64	74.75/73.94	71.98/72.46	69.01/68.56
LGGNET	×	×	√	√	78.14/78.92	75.09/76.86	72.14/73.22	72.76/72.06
MMM	√	×	√	√	77.14/78.61	76.17/77.95	72.24/73.21	71.28/71.80
EPNNE	×	×	×	×	78.17/77.92	76.47/76.31	72.81/73.46	71.42/71.17
VIT	×	√	√	×	77.10/78.32	72.94/73.06	71.70/72.83	68.12/67.17
MAE	√	√	√	×	80.27/81.78	75.73/75.85	73.29/74.48	70.64/70.15
I-JEPA	√	√	√	×	80.81/81.30	76.93/77.84	**74.73/75.04**	72.24/71.47
RCT-MNM	√	√	√	√	**82.14/82.10**	**77.91/78.03**	74.46/74.81	**72.90/73.44**

Table 2. Fine-tuning classification AUROC (%) on MSIT/ECR. **BOLD**: the best performance are bolded. UNDERLINE: the second-best performance.

MODEL	SSL	T	C	R	MSIT	ECR
SVM	×	×	×	×	80.87	80.10
DGCNN	×	×	√	×	81.72	82.69
RGNN	×	×	√	×	82.39	81.76
LGG-NET	×	×	√	√	83.89	84.13
VIT	×	√	√	×	82.88	83.91
MAE	√	√	√	×	86.22	86.67
I-JEPA	√	√	√	×	87.17	86.59
RCT-MNM	√	√	√	√	**89.34**	**89.05**

implemented other baselines widely used in emotion recognition: **DGCNN** (Song et al., 2018), **RGNN** (Zhong et al., 2020), **TSception** (Ding et al., 2022), **EEG Conformer** (Song et al., 2022), **MMM** (Yi et al., 2024), **LGGNet** (Ding et al., 2023), **EPNNE** (Zhang et al., 2024), and task engagement classification: **SVM** (Provenza et al., 2019). We also employed **DGCNN**, **RGNN**, and **LGGNet** for task engagement classification, since they are graph neural network based models which rather match the inducitve bias of more distributed spatial representations in the MSIT, ECR datasets.

4.2. Experimental Results

Table 2 presents the fine-tuning AUROC results for the MSIT and ECR tasks. Traditional models such as SVM perform the worst, lacking the ability to capture complex neural dynamics. Graph-based models, including DGCNN, RGNN, and LGG-Net, improve performance by incorporating spatial connectivity but still fall short of transformer-based approaches. Among transformer-based models, ViT provides a baseline for self-attention mechanisms, while MAE and I-JEPA leverage self-supervised learning for improved representations. Our RCT-MNM model outperforms all baselines (89.34%, 89.05%), benefiting from spatial-temporal modeling and region-channel self-attention, which enables it to capture fine-grained sEEG representations crucial for cognitive conflict resolution. Overall, our model demonstrates its ability to effectively leverage self-supervised learning (SSL) to handle both subject-dependent and subject-independent classification tasks, achieving consistent improvements across datasets and metrics. The results highlight the robustness of our approach in capturing both intra- and inter-subject variability, making it a promising solution for emotion recognition and related applications. Table 1 presents classification results on SEED and SEED-IV under subject-dependent and subject-independent settings. Our model, RCT-MNM, consistently outperforms baselines. Table 3 further validates our approach on DEAP dataset. Across all evaluations, RCT-MNM surpasses both spatial modeling architecture (DGCNN, RGNN, LGG-Net) and purely temporal models (EEG Conformer) by effectively balancing temporal, regional, and channel interactions. The hierarchical temporal-first encoding ensures stable representations before spatial aggregation, while region-channel self-attention refines functional connectivity patterns. Compared with other self-supervised learning models, RCT-MNM further enhance the pre-training process with RCT-Mask with strong spatial inductive bias and adapt for downstream tasks using Adaptive Refinement process. Additionally, the model's ability to generalize across subjects suggests its effectiveness in handling inter-subject variability, a key challenge in EEG-based cognitive and emotion analysis. Overall, RCT-MNM advances neural representation learning by integrating hierarchical attention, self-supervised learning, and regional conditioning, achieving superior performance in emotion recognition and cognitive task engagement.

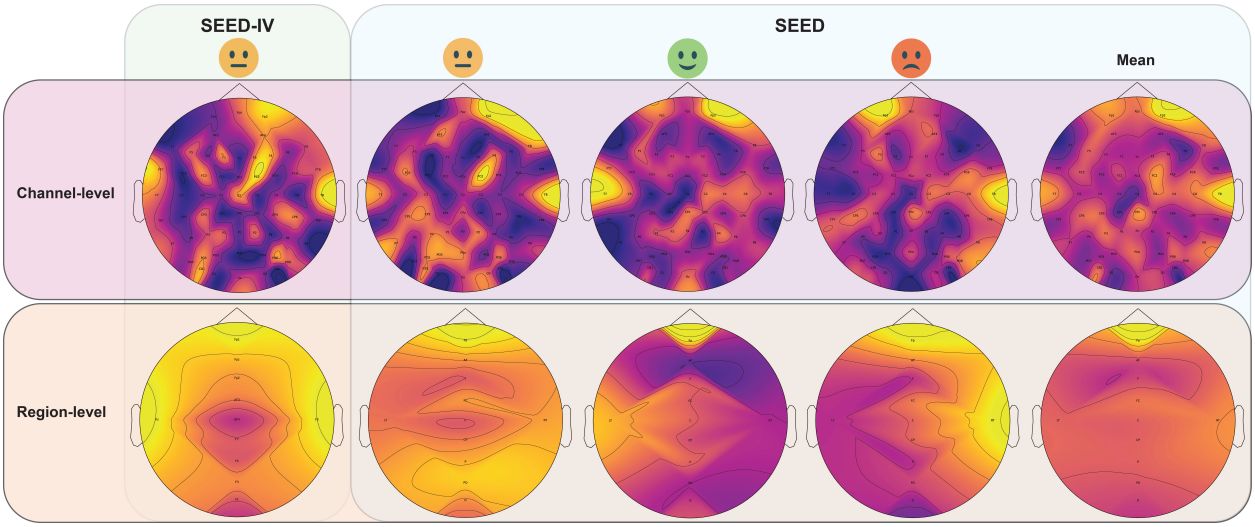

Figure 2. **Region- and Channel-Level Class Activation Maps.** This figure visualizes EEG activation patterns for different emotions in SEED and SEED-IV datasets, showing channel-level fine-grained activations and region-level aggregated representations for a holistic view of emotional processing.

4.3. Interpretations and Visualizations

Figure 2 presents class activation maps (Alarcao & Fonseca, 2017) at the channel and region levels for EEG-based emotion recognition across the SEED (Neutral, Positive, Negative) and SEED-IV (Neutral) datasets. The rightmost column (Mean) shows the average activation across conditions for SEED. The channel-level CAMs reveal distinct spatial distributions of neural activity across emotions. For SEED-IV, the model primarily focuses on frontal and central regions in the neutral condition. For SEED, neutral emotions exhibit similar activations as for SEED-IV. Positive emotions show diffused activation patterns particularly in parietal and occipital regions, and negative emotions elicit heightened activations in the frontal and temporal regions, aligning with findings that these areas are associated with negative affect and stress processing. At the region level, a more structured and spatially smooth representation of neural activity emerges, demonstrating the effectiveness of regional aggregation in our RCT-MNM framework. For SEED-IV (Neutral), activations are dominant in the frontal and central regions, reinforcing their role in neutral emotional states. In SEED, positive emotions show higher activations in the posterior regions, while negative emotions show stronger frontal dominance, consistent with the valence asymmetry hypothesis, which associates left-frontal activity with positive emotions and right-frontal activity with negative emotions. The region-level activations appear smoother and more interpretable compared to channel-level activations, suggesting that hierarchical spatial modeling enhances feature abstraction and generalization.

5. Conclusion

In this paper, we propose Region-Channel-Temporal Masked Neural Modeling (RCT-MNM), a self-supervised learning framework for universal EEG/sEEG representation learning. RCT-MNM leverages region-channel-temporal self-attention and masked neural modeling to effectively capture temporal and spatial dependencies in EEG and sEEG signals. By employing regional conditioned self-supervised learning (RC), the model learns robust neural representations that enhance both emotion recognition and psychiatric disorder analysis. We introduce a hierarchical structure, where local channel-level features are aggregated into region-level embeddings, improving both efficiency and interpretability. Experimental results on multiple EEG datasets for emotion recognition (SEED, SEED-IV, and DEAP) and sEEG datasets for task engagement classification (MSIT and ECR) demonstrate that RCT-MNM outperforms state-of-the-art models in both subject-dependent and subject-independent settings. The visualization of class activation maps and self-attention graphs further validates the model's ability to learn meaningful neurophysiological representations. Future work will explore real-time BCI applications, cross-domain transfer learning, and expanded pre-training datasets to further enhance model scalability and generalization.

Impact Statement

This work explores self-supervised learning for EEG/sEEG representation by integrating region-channel-temporal self-

attention and masked neural modeling to enhance feature extraction for emotion recognition and psychiatric analysis. Our approach improves the generalizability and interpretability of EEG/sEEG-based models, with potential applications in mental health monitoring and brain-computer interfaces (BCIs). There are no immediate negative societal impacts anticipated.

References

Alarcao, S. M. and Fonseca, M. J. Emotions recognition using eeg signals: A survey. *IEEE transactions on affective computing*, 10(3):374–393, 2017.

Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., and Ballas, N. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

Bardes, A., Ponce, J., and Lecun, Y. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *ICLR 2022-International Conference on Learning Representations*, 2022.

Chien, H.-Y. S., Goh, H., Sandino, C. M., and Cheng, J. Y. Maeeg: Masked auto-encoder for eeg representation learning. In *NeurIPS Workshop*, 2022. URL https://arxiv.org/abs/2211.02625.

Ding, Y., Robinson, N., Zhang, S., Zeng, Q., and Guan, C. Tsception: Capturing temporal dynamics and spatial asymmetry from eeg for emotion recognition. *IEEE Transactions on Affective Computing*, 14(3):2238–2250, 2022.

Ding, Y., Robinson, N., Tong, C., Zeng, Q., and Guan, C. Lggnet: Learning from local-global-graph representations for brain–computer interface. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Drane, D. L., Pedersen, N. P., Sabsevitz, D. S., Block, C., Dickey, A. S., Alwaki, A., and Kheder, A. Cognitive and emotional mapping with seeg. *Frontiers in Neurology*, 12:627981, 2021.

Gao, Z., Wang, X., Yang, Y., Mu, C., Cai, Q., Dang, W., and Zuo, S. Eeg-based spatio–temporal convolutional neural network for driver fatigue evaluation. *IEEE transactions on neural networks and learning systems*, 30(9):2755–2763, 2019.

Gevins, A., Leong, H., Smith, M. E., Le, J., and Du, R. Mapping cognitive brain function with modern high-resolution electroencephalography. *Trends in Neurosciences*, 18(10):429–436, 1995. ISSN 0166-2236. doi: https://doi.org/10.1016/0166-2236(95)94489-R. URL https://www.sciencedirect.com/science/article/pii/016622369594489R.

Goschke, T. Dysfunctions of decision-making and cognitive control as transdiagnostic mechanisms of mental disorders: Advances, gaps, and needs in current research. *International journal of methods in psychiatric research*, 23(S1):41–57, 2014.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Houssein, E. H., Hammad, A., and Ali, A. A. Human emotion recognition from eeg-based brain–computer interface using machine learning: a comprehensive review. *Neural Computing and Applications*, 34(15):12527–12557, 2022.

Jiang, W., Zhao, L., and Lu, B.-l. Large brain model for learning generic representations with tremendous eeg data in bci. In *The Twelfth International Conference on Learning Representations*, 2024.

Koelstra, S., Muhl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., and Patras, I. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, 3(1):18–31, 2011.

Kret, M. E. and Ploeger, A. Emotion processing deficits: a liability spectrum providing insight into comorbidity of mental disorders. *Neuroscience & Biobehavioral Reviews*, 52:153–171, 2015.

Lachaux, J., Rudrauf, D., and Kahane, P. Intracranial eeg and human brain mapping. *Journal of Physiology-Paris*, 97(4):613–628, 2003. ISSN 0928-4257. doi: https://doi.org/10.1016/j.jphysparis.2004.01.018. URL https://www.sciencedirect.com/science/article/pii/S0928425704000348. Neuroscience and Computation.

Liu, J., Younk, R., Drahos, L. M., Nagrale, S. S., Yadav, S., Widge, A. S., and Shoaran, M. Neural decoding and feature selection methods for closed-loop control of avoidance behavior. *Journal of Neural Engineering*, 21(5):056041, 2024.

Mentzelopoulos, G., Chatzipantazis, E., Ramayya, A. G., Hedlund, M., Buch, V., Daniilidis, K., Kording, K., and Vitale, F. Neural decoding from stereotactic eeg: accounting for electrode variability across subjects. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Ng, M. Y. and Weisz, J. R. Annual research review: Building a science of personalized intervention for youth mental health. *Journal of Child Psychology and Psychiatry*, 57 (3):216–236, 2016.

Organization, W. H. The world health report 2001: Mental health: new understanding, new hope. 2001.

Peled, N., Gholipour, T., Paulk, A., Felsenstein, O., Dougherty, D., Widge, A., Eskandar, E., Cash, S., Hamalainen, M., and Stufflebeam, S. Invasive electrodes identification and labeling. *GitHub Repos*, 10, 2017.

Provenza, N. R., Paulk, A. C., Peled, N., Restrepo, M. I., Cash, S. S., Dougherty, D. D., Eskandar, E. N., Borton, D. A., and Widge, A. S. Decoding task engagement from distributed network electrophysiology in humans. *Journal of neural engineering*, 16(5):056015, 2019.

Rafiei, M. H., Gauthier, L. V., Adeli, H., and Takabi, D. Self-supervised learning for electroencephalography. *IEEE Transactions on Neural Networks and Learning Systems*, 35(2):1457–1471, 2024. doi: 10.1109/TNNLS.2022.3190448.

Rouast, P. V., Adam, M. T., and Chiong, R. Deep learning for human affect recognition: Insights and new developments. *IEEE Transactions on Affective Computing*, 12 (2):524–543, 2019.

Saha, S. and Baumert, M. Intra- and inter-subject variability in eeg-based sensorimotor brain computer interface: A review. *Frontiers in Computational Neuroscience*, 13, 2020. ISSN 1662-5188. doi: 10.3389/fncom.2019.00087. URL https://www.frontiersin.org/journals/computational-neuroscience/articles/10.3389/fncom.2019.00087.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Shazeer, N. GLU variants improve transformer. *CoRR*, abs/2002.05202, 2020. URL https://arxiv.org/abs/2002.05202.

Song, T., Zheng, W., Song, P., and Cui, Z. Eeg emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, 11 (3):532–541, 2018.

Song, Y., Zheng, Q., Liu, B., and Gao, X. Eeg conformer: Convolutional transformer for eeg decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2022.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wang, G., Liu, W., He, Y., Xu, C., Ma, L., and Li, H. Eegpt: Pretrained transformer for universal and reliable representation of eeg signals. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Ye, M., Chen, C. P., and Zhang, T. Hierarchical dynamic graph convolutional network with interpretability for eeg-based emotion recognition. *IEEE transactions on neural networks and learning systems*, 2022.

Yi, K., Wang, Y., Ren, K., and Li, D. Learning topology-agnostic eeg representations with geometry-aware modeling. *Advances in Neural Information Processing Systems*, 36, 2024.

Zhang, H., Zuo, T., Chen, Z., Wang, X., and Sun, P. Z. Evolutionary ensemble learning for eeg-based cross-subject emotion recognition. *IEEE Journal of Biomedical and Health Informatics*, 28(7):3872–3881, 2024. doi: 10.1109/JBHI.2024.3384816.

Zheng, W.-L. and Lu, B.-L. Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. *IEEE Transactions on autonomous mental development*, 7(3):162–175, 2015.

Zheng, W.-L., Liu, W., Lu, Y., Lu, B.-L., and Cichocki, A. Emotionmeter: A multimodal framework for recognizing human emotions. *IEEE transactions on cybernetics*, 49 (3):1110–1122, 2018.

Zhong, P., Wang, D., and Miao, C. Eeg-based emotion recognition using regularized graph neural networks. *IEEE Transactions on Affective Computing*, 13(3):1290–1301, 2020.

A. Additional Classification Results

Table 3. Classification performance on DEAP (Valence/Arousal) (%). **BOLD**: the best performance are bolded. UNDERLINE: the second-best performance.

MODEL	SUBJECT-DEPENDENT		SUBJECT-INDEPENDENT	
	ACCURACY	F1 SCORE	ACCURACY	F1 SCORE
DGCNN	64.97/58.52	68.08/60.13	61.76/64.82	60.15/62.92
RGNN	65.65/67.54	64.70/58.74	63.45/65.26	62.28/66.11
TSCEPTION	64.19/66.26	63.86/65.47	62.83/64.78	63.86/**68.08**
EEG CONFORMER	62.48/65.02	64.09/62.60	62.00/60.16	61.76/57.84
LGGNET	67.28/66.61	67.14/66.69	66.17/65.85	64.54/67.44
MMM	65.18/67.01	64.96/68.97	66.03/66.16	65.58/65.64
EPNNE	67.20/64.23	65.06/66.17	65.58/63.31	65.00/63.88
VIT	66.40/63.27	65.83/64.42	65.15/64.67	64.67/63.52
MAE	67.20/68.17	66.01/67.16	65.58/64.52	65.30/65.27
I-JEPA	69.22/69.01	68.19/68.29	66.70/67.52	67.56/66.47
RCT-MNM	**71.27/70.92**	**71.78/70.60**	**67.47/67.77**	**68.15**/67.94

B. Dataset Details

MSIT/ECR. The MSIT and ECR datasets are utilized to study human cognitive and emotional conflict responses. These datasets consist of electrophysiological recordings from 17 participants with pharmaco-resistant complex partial seizures undergoing invasive monitoring. Participants performed two distinct conflict-based tasks: the Multi-Source Interference Task (MSIT) and the Emotional Conflict Resolution (ECR) task. The MSIT involved identifying a target number among distractors under varying levels of congruence, while the ECR task required resolving emotional conflict between facial expressions and overlaid text. Recordings were made using depth electrodes placed in up to 30 brain regions, capturing local field potentials (LFPs) at 2 kHz and then downsampled to 1 kHz. The number of electrodes used for recording varied across participants from 37 to 195. The LFP recordings from different channels were localized to standard brain regions using electrode mapping tools. To remove the low and high-frequency noise, a band-pass filter from 1.0-150Hz was applied to the original LFP. Line noise was removed using notch filters on 60 Hz and its harmonics. A data augmentation step by splitting each trial into smaller non-overlapping 4-second segments was applied. Then, the neural data $\mathbf{X} \in \mathbb{R}^{C\times 4000}$ was used to predict rest-state versus task-state as $y \in \{0, 1\}$, where C is the number of channels. We split each session into the training-validation-testing sets in a chronological manner with a 80%-10%-10% ratio, between them we left 4-second segments to make sure no adjacent segments will appear in both training and testing data. This dataset provides insights into task-specific and generalized cognitive processes by leveraging robust, multi-site neural recordings, with applications in psychiatric adaptive deep brain stimulation and beyond.

DEAP. The DEAP dataset is designed for analyzing human affective states using multimodal data. It includes EEG and peripheral physiological signals from 32 participants as they watched 40 one-minute music video clips. Participants rated the videos on arousal, valence, dominance, liking, and familiarity. EEG data was recorded using 32 electrodes at 512 Hz, downsampled to 128 Hz. The 3 seconds pre-trial baseline was removed for each trial. To remove the low and high-frequency noise, a band-pass filter from 4.0-45Hz was applied to the original EEG. The class label for each dimension is from 1 to 9, hence 5 was selected as a threshold to project the 9 discrete values into low and high classes in each dimension. A data augmentation step by splitting each trial into smaller non-overlapping 4-second segments was applied. Then, the neural data $\mathbf{X} \in \mathbb{R}^{32\times 512}$ was used to predict low valence/arousal versus high ones as $y \in \{0, 1\}$. There are two types of experiment settings in this paper: I) subject-dependent: trial-wise 10-fold cross-validation; II) subject-independent: leave-one-subject-out cross-validation. Randomly shuffling the segments among trials before the training-testing split of the data could make the adjacent segments be in training and testing data, which will give high classification results. But the accuracy will drop when the highly correlated segments are never seen by the model in the real-world situation. To get the more generalized evaluation, the folds are split among trials, which will make sure the adjacent segments in one trial will not appear in both training and testing data. The ratio of training to validation data is 80%-20%.

SEED. The SEED dataset is a discrete EEG emotion dataset collected through video stimuli. It includes EEG recordings from fifteen subjects, captured via 62 sensors while participants viewed multiple Chinese film clips, each eliciting one of

three emotional responses, namely positive, neutral, and negative. EEG data was recorded at 1 kHz, downsampled to 200 Hz. For each subject, recordings were conducted across three separate sessions, with each session containing EEG data from fifteen film clips (fifteen trials). Additionally, the SEED-IV dataset, an extended version of SEED, comprises EEG data from the same fifteen subjects using the same 62-channel setup. Each subject's data in SEED-IV spans three sessions, with each session including 24 trials, and each trial elcited one of four emotion responses, namely happy, neutral, sad, and fear. A data augmentation step by splitting each trial into smaller non-overlapping 4-second segments was applied. Then, the neural data $\mathbf{X} \in \mathbb{R}^{62\times800}$ was used to predict distinct emotion responses as $y \in \{0, 1, 2\}$ for SEED and $y \in \{0, 1, 2, 3\}$ for SEED-IV. There are two types of experiment I) subject-dependent: trial-wise 5-fold cross-validation; II) subject-independent: leave-one-subject-out cross-validation as DEAP. The ratio of training to validation data is 80%-20%.

C. Details of Baseline Models

DGCNN (Song et al., 2018) is a deep learning framework for EEG emotion recognition that utilizes dynamical graph convolutional neural networks. It dynamically constructs an adjacency matrix to capture the intrinsic relationships between EEG channels, enhancing feature extraction. The model applies graph convolution to extract localized spatial dependencies, followed by a 1×1 convolution layer for feature transformation. A fully connected layer is used for classification. Unlike traditional GCNNs, DGCNN learns the adjacency matrix adaptively during training, improving EEG emotion recognition performance.

RGNN (Zhong et al., 2020) is a deep learning model designed for EEG-based emotion recognition by incorporating biologically inspired graph structures. It models inter-channel relationships via an adjacency matrix, leveraging neuroscience principles to capture both local and global spatial dependencies. RGNN introduces two key regularizers: Node-wise Domain Adversarial Training (NodeDAT) to enhance cross-subject generalization and Emotion-Aware Distribution Learning (EmotionDL) to mitigate label noise. The model extends simple graph convolution networks (SGC) for efficient EEG representation learning.

TSception (Ding et al., 2022) is a multi-scale convolutional neural network designed for EEG-based emotion recognition. It consists of three key layers: a dynamic temporal layer, which employs multi-scale 1D convolutional kernels to capture temporal dynamics and frequency representations; an asymmetric spatial layer, which leverages the brain's hemispheric asymmetry to extract global and hemisphere-specific spatial patterns; and a high-level fusion layer, which integrates learned spatial and temporal representations. TSception effectively captures EEG temporal and spatial features, improving classification performance in emotion recognition tasks while maintaining a compact and efficient model architecture.

EEG Conformer (Song et al., 2022) is a hybrid convolutional-transformer model designed for EEG decoding and visualization. It consists of three key components: a convolution module for capturing low-level local temporal and spatial features, a self-attention module for modeling global temporal dependencies, and a classifier module based on fully connected layers for final prediction. The convolutional layers extract localized features using temporal and spatial convolutions, while an average pooling layer reduces redundant information. The self-attention mechanism is applied to the transformed feature maps to learn long-term dependencies.

MMM (Yi et al., 2024) is a self-supervised pre-training framework designed for EEG-based emotion recognition. It unifies EEG channel topologies using a Multi-dimensional position encoding, Multi-level channel hierarchy, and a Multi-stage pre-training strategy to obtain topology-agnostic representations. The model employs a masked autoencoder (MAE)-based approach, where EEG signals are mapped onto a unified topology that ensures cross-dataset generalization. Region-wise tokens are introduced to capture local spatial information, while a structured masking strategy (random and region-wise masking) enhances representation robustness.

LGG-Net (Ding et al., 2023) is a neurologically inspired graph neural network (GNN) designed for brain–computer interface (BCI) applications using EEG data. It integrates local-global-graph (LGG) representations to capture both localized and global brain activities. The model consists of a temporal learning block employing multi-scale 1D convolutions for dynamic EEG feature extraction and a graph learning block with local- and global-graph-filtering layers to model brain region interactions. LGG-Net applies neurophysiological priors, defining EEG as a structured LGG with functional subgraphs, and learns hierarchical spatial dependencies through an instance-specific global adjacency matrix.

EPNNE (Zhang et al., 2024) is an end-to-end framework designed for cross-subject EEG-based emotion recognition. It employs an evolutionary programming (EP)-based optimization strategy to automatically search for an optimal neural network architecture, mitigating inter-subject variability in EEG signals. The model integrates an ensemble of multilayer

perceptrons (MLPs) optimized through crossover and mutation operations, enhancing the generalizability of emotion recognition.

D. Metrics

We evaluated the performance of all baseline models and our model using the Area Under the Receiver Operating Characteristic Curve (AUROC) for task engagement classification on MSIT and ECR, and using Accuracy and Weighted F1-Score for emotion recognition on SEED, SEED-IV, and DEAP.

E. Additional Interpretations and Visualizations

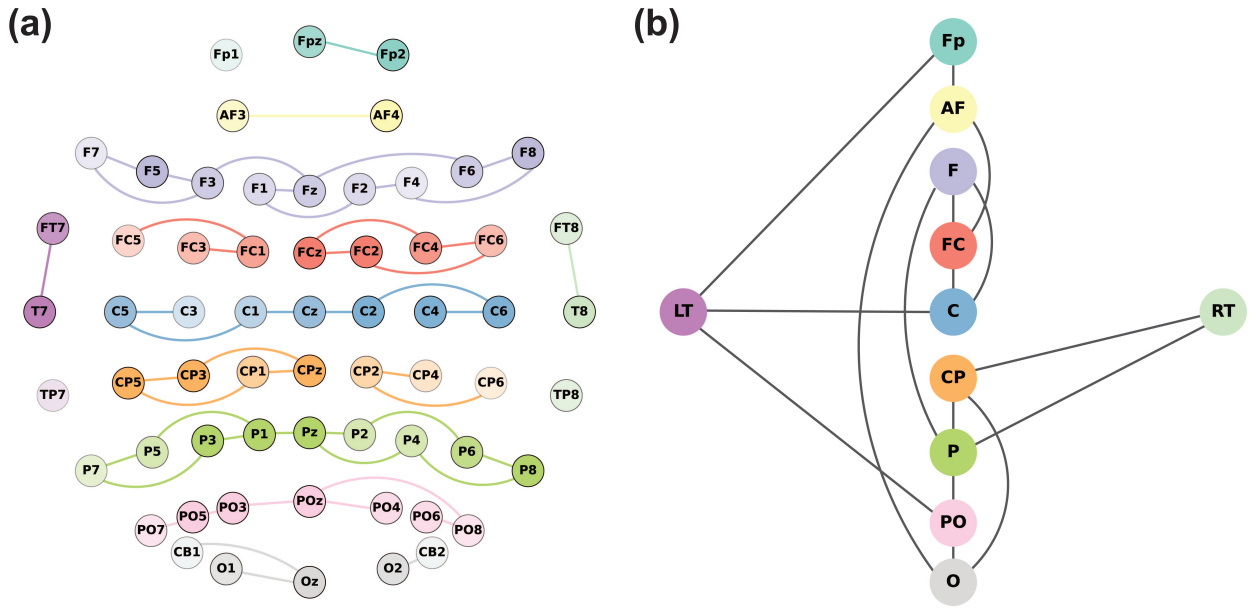

Figure 3. **Region-Channel Self-Attention Graphs.** (a) Localized EEG channel attention within regions. (b) Region-level attention dependencies, with nodes as aggregated regional embeddings. Connections highlight strong learned attention relationships between channels (colored) and regions (gray-scaled). The color intensity of channels in (a) indicates its level of attention to its corresponding region.

Figure 3 illustrates the Region-Channel Self-Attention Graphs for SEED dataset, providing insight into the structured spatial dependencies learned by the Region-Channel Transformer with Masked Neural Modeling framework. The channel connectivity graph in Figure 3(a) depicts the learned intra-regional, functional connections between channels measured using attention scores for local feature extraction. In Figure 3(b), each node of the channel connectivity graph represents a major brain region, with edges denoting the learned attention relationships between different regions. It is observed that there is stronger connectivity among adjacent and functionally related regions, such as frontal-central (FC-C) and parietal-occipital (P-PO) connections. Long-range dependencies between temporal (LT, RT) and pre-frontal (Fp) or parietal (P, PO) regions, highlight interactions crucial for cognitive and affective processes. Asymmetric attention distribution is observed which aligns with known emotional and neuropsychiatric functions.

Overall, the visualization results validate the RCT-MNM framework's ability to capture both local and global neural dependencies, leading to robust and interpretable emotion classification. The hierarchical self-attention mechanism effectively integrates channel- and region-level features, facilitating better generalization and alignment with functional brain networks involved in emotional and cognitive processes.

F. Additional Ablation Study

Table 4. Ablations to validate the effectiveness of MNM-Loss on MSIT/ECR (AUROC (%)) and SEED (Accuracy/F1 Score (%)).

$\mathcal{L}_{\text{input}}$	$\mathcal{L}_{\text{rep}}$	$\mathcal{L}_{\text{reg}}$	MSIT	ECR	SEED	
					DEPENDENT	INDEPENDENT
√	√	√	**89.34**	**89.05**	**82.14**/82.10	**77.91/78.03**
√	√	×	89.08	87.70	81.88/82.03	76.86/77.93
√	×	×	87.26	87.02	81.74/81.44	75.66/76.93
×	√	√	88.23	88.64	81.98/**82.57**	77.76/77.84

Table 5. Ablations to validate the effectiveness of Region-Channeel-Temporal Masking (RCT-Mask), Regioal Conditioning (RC), and Adaptive Attention-Based Refining on MSIT/ECR (AUROC (%)) and SEED (Accuracy/F1 Score (%)).

MODEL	MSIT	ECR	SEED	
			DEPENDENT	INDEPENDENT
OUR	**89.34**	**89.05**	**82.14**/82.10	**77.91/78.03**
W/O RC	88.37	87.77	81.85/**82.22**	77.24/77.21
RCT-MASK → RANDOM	88.24	87.12	81.05/81.38	77.09/76.14
W/O REFINE	87.71	87.20	81.47/81.11	76.64/76.81

We performed ablation studies to evaluate the effectiveness of different components in our proposed framework. Specifically, we examined the contributions of the loss components in the RCT-MNM in Table 4 and key architectural components, including Region-Channel-Temporal masking (RCT-Mask), Regional Conditioning (RC), and Adaptive Attention-Based Layer Refining in Table 5. The results highlight the importance of these design choices in achieving state-of-the-art performance across multiple datasets, including MSIT, ECR, and SEED.

Loss Component. We systematically remove one or more loss components. The complete model, utilizing all three losses, achieves the relatively higher performance across datasets. Removing the VICReg loss leads to a notable performance drop across all datasets except MSIT. It enforces a balanced representation space by penalizing covariance redundancy and maximizing feature invariance. This contributes to better generalization, especially for datasets with high inter-subject variability, such as SEED for subject-independent emotion recognition. When the embedding loss is excluded, the performance degradation is more pronounced. The representatioin alignment loss ensures consistency between masked embeddings and their corresponding latent representations, which is crucial for retaining task-relevant information. Its absence makes the model less robust, especially for SEED. Removing input reconstruction loss reduces Accuracy/F1 Score for SEED. This indicates that input loss aligns reconstructed inputs with the original signals, contributing to stability but is partially redundant when the embedding and VICReg losses are present.

Masking. Replacing the RCT-Mask with random masking significantly reduces performance. The RCT-Mask ensures structured masking along region, channel and temporal dimensions, which encourages the model to learn meaningful relationships across these domains. Random masking disrupts this structure, leading to degraded representations and poor generalization.

Regional Conditioning. Removing RC results in performance degradation. RC enables task-specific contextualization of target embeddings, which improves alignment with downstream objectives. While the drop is less severe than other ablations, it demonstrates the importance of context-aware feature learning.

Refinement. Removing the refinement step leads to consistent performance drops across all datasets. Refinement fine-tunes the intermediate-layer encoder representations to align with downstream tasks, significantly boosting performance. Its absence results in less effective task adaptation.

The results demonstrate the synergistic contributions of our proposed components. The **RCT-Mask** ensures the model learns meaningful spatial-temporal relationships, while **RC** improves feature contextualization. **Loss regularization**, particularly $\mathcal{L}_{\text{reg}}$, enhances the representation space's robustness, and **Refinement** aligns these representations with specific tasks. Collectively, these components enable our framework to achieve state-of-the-art performance on MSIT, ECR, and SEED

datasets, demonstrating its effectiveness for emotion recognition and psychiatric disorder prediction.

G. Addtional Figure

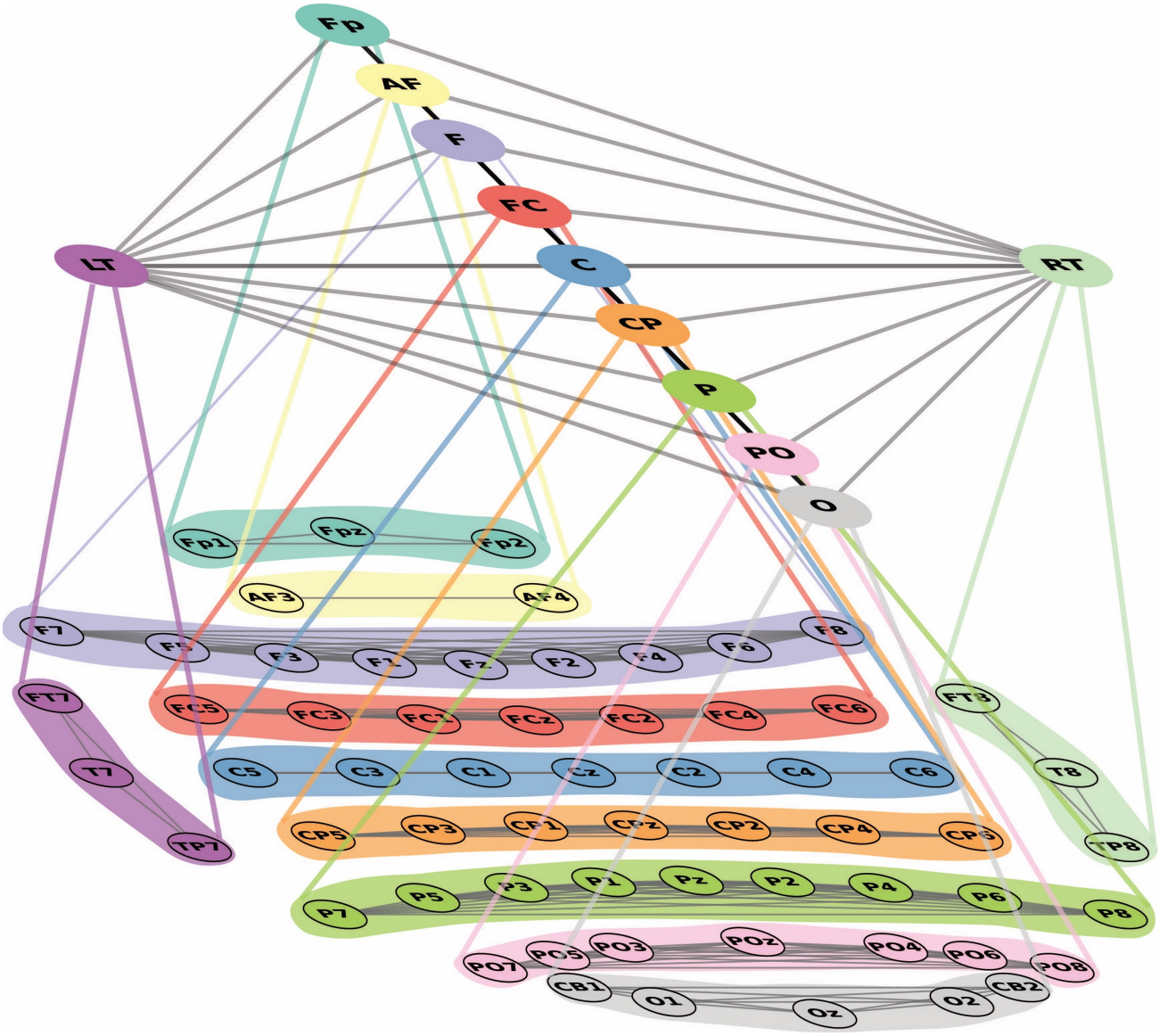

Figure 4. **Region-Channel Self-Attention Map.**