Tri-Domain Cross-Attention Transformers for Cross-Subject EEG Decoding of Cognitive States

Zhengde Li[1] **& Gaoyan Zhang**[1,*]
[1]School of Artificial Intelligence, Tianjin University, Tianjin, 300350, Tianjin, China
[*]Corresponding author, Email: zhanggaoyan@tju.edu.cn

Abstract

Mental fatigue and high cognitive loads reduce cognitive efficiency and task performance. Electroencephalogram (EEG) can track these states with high temporal resolution, but informative patterns span time, frequency, and distributed scalp topology. Most decoders handle these domains separately or fuse them only at the output, which can miss cross-domain dependencies. We propose TD-CAT, a tri-domain cross-attention Transformer that models interactions among temporal, spectral, and spatial representations. TD-CAT builds domain-specific features via temporal and spectral convolutions and a topology-aware graph module, then integrates the three streams with bidirectional cross-attention and low-rank fusion. Under leave-one-subject-out evaluation on SADT and SEED-VIG (driver fatigue) and MAT (cognitive load), TD-CAT reaches 80.7% and 92.6% accuracy on SADT and SEED-VIG, and 82.4% on MAT, demonstrating strong cross-subject performance across multiple datasets and task settings.

Keywords: EEG; mental fatigue; cognitive load; Transformer

Introduction

Mental fatigue and elevated cognitive load are two common human states in sustained task settings and human–machine interaction, and both can reduce the efficiency with which cognitive resources are deployed during task performance (Benelli et al., 2024; Longo et al., 2022; Sweller, 1988). Mental fatigue and cognitive load arise from different immediate conditionsSalihu et al., 2022. Mental fatigue is typically associated with prolonged time-on-task and reduced alertness, whereas high cognitive load is more directly associated with current task demands. Both states can nevertheless impair cognitive efficiency and task performancePaas and Van Merrienboer, 2020. When cognitive efficiency declines, sustained and safety-critical work becomes more vulnerable to lapses and adverse outcomes (Kunasegaran et al., 2023). Reliable detection of these states can therefore enable timely intervention before errors become costly. Electroencephalography (EEG) provides a direct, time-resolved measure of brain activity and is well suited to monitoring such state-dependent changes. In particular, EEG can track shifts in oscillatory activity and distributed neural interactions that vary with cognitive states, thereby supporting real-time assessment (Berka et al., 2007; Gao, Tang, et al., 2023).

Existing approaches decode fatigue and cognitive load from EEG using three main perspectives: temporal dynamics, spectral representations, and spatial structure. Convolutional and recurrent networks learn short- and mid-range patterns from EEG sequences (Gao, Li, et al., 2023; Ye et al., 2022). Spectral pipelines compute band-power features or time–frequency maps and feed them to deep classifiers, leveraging rhythmic signatures such as α and θ activity (Gao, Tang, et al., 2023; P. Wang et al., 2024). Graph-based models encode inter-electrode relations using geometry-based adjacency or learned connectivity (Hu et al., 2024). Attention-based architectures (e.g., Transformers) model longer-range dependencies in EEG sequences (J. Li et al., 2024). Hybrid designs such as Conformer integrate convolution and attention to combine local and global information (Song et al., 2022). Multi-view methods further fuse multiple domains, for example by combining graph spectral transforms with cross-stream aggregation (Khateeb et al., 2021; Sartipi et al., 2021).

Decoding cognitive states remains challenging because relevant EEG signatures are distributed across time, frequency, and space (Tran et al., 2020; H. Wang et al., 2021; Yue et al., 2022). Many models still process these domains independently or fuse them only shallowly, which can miss cross-domain dependencies and limit interpretability. It also remains unclear how temporal dynamics and oscillatory structure relate to distributed spatial organization under fatigue or high cognitive load. Spatial inductive biases pose a further trade-off: geometry-based graphs may miss task-relevant functional organization, whereas fully learned graphs can drift away from psychologically meaningful structure. In addition, most studies are validated within a single dataset, leaving evaluation across datasets and task contexts relatively limited. For mechanistic understanding, decoding evidence should map onto plausible organization across regions and networks (H. Wang et al., 2021; Yue et al., 2022). These gaps motivate a framework that learns complementary tri-domain features, models their interactions explicitly, and supports pathway-level interpretation relevant to cognitive states.

To address these issues, we introduce the **Tri-Domain Cross-Attention Transformer**, TD-CAT, for EEG decoding of cognitive states. TD-CAT frames fatigue and cognitive load as distinct but related human states that may both manifest as reduced cognitive efficiency during ongoing task performance. It first builds domain-specific embeddings with temporal convolution, frequency-domain convolution, and a cognition-inspired GCN over scalp topology. The GCN combines a cognitive prior graph with data-driven adaptation to preserve interpretability while fitting the data. TD-CAT then

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

applies bidirectional cross-attention between the temporal and spectral streams, the temporal and spatial streams, and the spectral and spatial streams. A channel-wise low-rank fusion module integrates the three streams and reduces redundant mixing.

We evaluate TD-CAT on SADT (Cao et al., 2019) and SEED-VIG (Zheng & Lu, 2017) for driver fatigue and on MAT (Zyma et al., 2019) for cognitive load under leave-one-subject-out protocols. We also conduct ablations and analyze learned spatial connections to identify region-level interaction patterns associated with efficiency decline.

The main contributions are as follows:

1. We propose TD-CAT, a tri-domain Transformer for EEG cognitive state decoding that learns temporal, spectral, and spatial representations through dedicated feature modules, including a topology-aware graph component for spatial modeling.
2. We introduce an explicit tri-domain integration strategy based on bidirectional cross-attention and low-rank fusion, enabling direct information exchange across domains rather than relying on late fusion.
3. We evaluate TD-CAT under leave-one-subject-out protocols on SADT, SEED-VIG, and MAT, covering fatigue and cognitive-load settings, and provide pathway-level interpretation by analyzing attention-derived spatial interactions that are broadly consistent with established neurocognitive markers.

Proposed Method

The proposed framework consists of three core components: a temporal-spatial-frequency multi-perspective encoder module, a tri-domain cross-attention Transformer, and a multi-domain low-rank feature fusion module. Initially, the temporal-spatial-frequency multi-perspective encoder module encodes the raw EEG signals into hidden representations across three distinct domains. Subsequently, cross-attention modules are employed between each pair of domains to learn and extract inter-domain pattern features. Following this, a channel-wise low-rank feature fusion module integrates features from all three perspectives. Furthermore, to fully leverage both shared and complementary properties across different domain features, the fused features from all three domains are concatenated and processed through a self-attention Transformer to obtain enhanced EEG embeddings. Finally, a classifier composed of fully connected layers with ReLU activation functions and dropout mechanisms predicts the corresponding labels from the fused features. The complete framework architecture is illustrated in Fig. 1.

Temporal-Spatial-Frequency Multi-Perspective Encoder

The temporal-spatial-frequency multi-perspective encoder consists of three sub-modules: a temporal convolution module, a frequency-domain convolution module, and a cognition-inspired graph convolution module.

In this study, let $X_i^t \in \mathbb{R}^{C\times T}$ denote the temporal samples of EEG signals, where C represents the number of channels and T represents the temporal length of each sample.

To extract time-related embedding representations from the samples, we first employ a 1×1 pointwise convolution W_{point} to fuse information across different channels, transforming the raw signals into higher-level features. This is followed by a $1 \times K$ temporal convolution W_{time}, where K denotes the kernel size. Subsequently, we apply the Gaussian Error Linear Unit (GELU) (Hendrycks & Gimpel, 2016) activation function for non-linear transformation, followed by batch normalization (BN) to mitigate covariate shift. The resulting temporal dimension embedding representation Z_i^t can be obtained through:

$$Z_i^t = \mathrm{BN}(\mathrm{GELU}(W_{time} * (W_{point} * X_i^t + b_{point}) + b_{time})) \quad (1)$$

For frequency-domain data, considering its inherent ability to separate noisy and beneficial frequency components, we maintain the frequency axis intact. We apply a similar 1×1 pointwise convolution operation W_{freq} to the frequency data $X_i^F \in \mathbb{R}^{C\times F}$, where $F = \lfloor \frac{T}{2} \rfloor + 1$. X_i^F is obtained by transforming the time-domain signal into the frequency domain via the Fast Fourier Transform (FFT). The extracted features then undergo GELU non-linear activation and BatchNorm operations:

$$Z_i^f = \mathrm{BN}(\mathrm{GELU}(W_{freq} * X_i^F + b_{freq})) \quad (2)$$

Given the multi-channel topological characteristics of EEG signals in the scalp spatial domain, the spatial distribution of electrodes naturally forms a biophysical graph structure. Graph Convolutional Networks (GCNs) can explicitly model neural interaction fields and functional connectivity patterns between electrodes. Traditional convolutional neural networks rely on strong assumptions about regular grids. In contrast, GCNs are better suited to capturing relational structure in EEG channel interactions. In the spatial branch, we adopt a dual-graph design that combines a global learnable adjacency matrix with a sample-specific functional connectivity graph. The learnable adjacency matrix serves as a shared prior graph, while the data-driven graph is constructed from the phase-locking value (PLV) of each EEG segment. PLV is computed by applying the Hilbert transform to obtain instantaneous phases and then measuring pairwise phase synchrony between channels. Using parameter sharing and spectral graph space alignment (C. Li et al., 2024), we process the two graph spaces with the same graph convolutional layer and fuse them at the feature level rather than at the adjacency level.

$$Z_i^{sc} = \mathrm{ReLU}\left(\sum_{k=0}^{K_g-1} \theta_k T_k(\widetilde{L}_c) X_i^t\right) \quad (3)$$

$$Z_i^{sd} = \mathrm{ReLU}\left(\sum_{k=0}^{K_g-1} \theta_k T_k(\widetilde{L}_d) X_i^t\right) \quad (4)$$

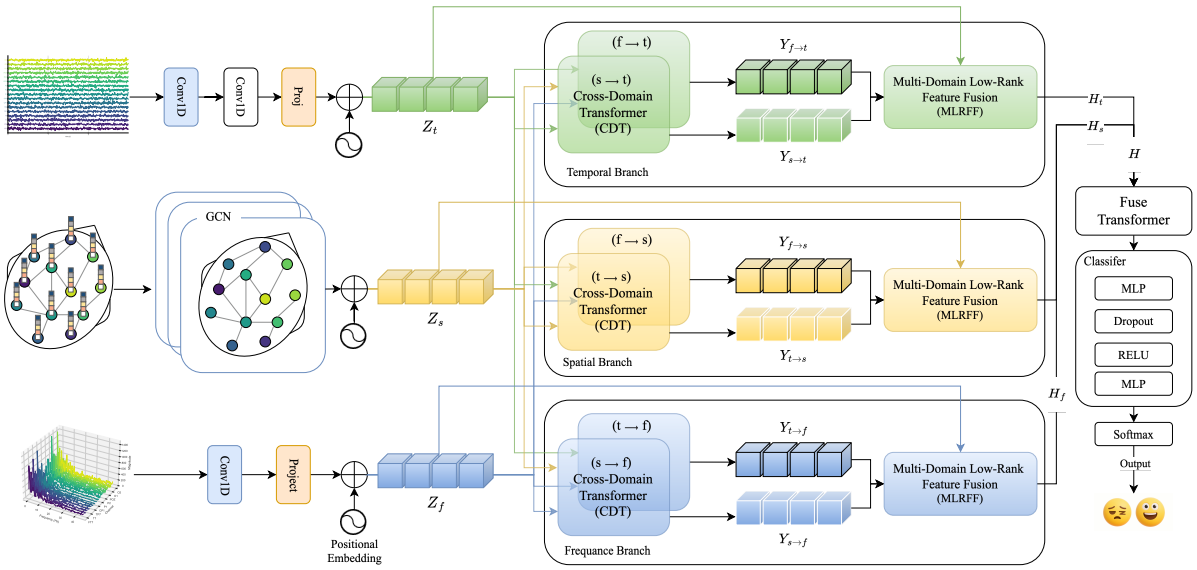

Figure 1: The overall architecture of our proposed framework.

where $\widetilde{L}_c$ and $\widetilde{L}_d$ represent the scaled Laplacian matrices of the learnable prior graph and the sample-specific PLV graph, respectively, T_k denotes the Chebyshev polynomial basis function, and θ_k represents learnable parameters. The final spatial representation is obtained by averaging the two graph-convolution outputs:

$$Z_i^s = \frac{Z_i^{sc} + Z_i^{sd}}{2} \tag{5}$$

Before proceeding with the next fusion operation, it is essential to ensure consistent hidden dimensions across all components. For both temporal and frequency convolution modules, we employ linear layers to map their data dimensions to the same attention hidden dimension h. Similarly, for the cognition-inspired graph convolution module, we set the output dimension of the graph convolutional network to h.

Tri-Domain Cross-Attention Transformer

EEG signals serve as a macroscopic representation of neuronal electrical activity. They exhibit strong coupling across temporal dynamics, frequency energy, and spatial topology. Traditional analysis methods struggle with these multi-domain links. They leave potential inter-dimensional correlations under-explored. To address this limitation, we introduce the Tri-Domain Cross-Attention Transformer (TD-CAT) module (Vaswani et al., 2017). TD-CAT builds deep interaction networks among the temporal, frequency, and spatial domains. This design enables fine-grained modeling of complex neural patterns in EEG signals. Consequently, the feature representations become more complete and discriminative.

The design of the TD-CAT module is inspired by research on multimodal Transformers (Tsai et al., 2019), which demonstrates the capability of multimodal Transformers to transfer information from source modalities to target modalities. The core of the TD-CAT module is the cross-attention mechanism, which operates by computing similarity matrices between queries and keys from different modalities to perform weighted aggregation of values.

The attention mechanism in Transformer models does not capture sequential order. Therefore, position encoding (PE) is added before feature processing to supply positional cues. PE methods can be learnable or fixed (Devlin et al., 2019; Vaswani et al., 2017). Because EEG token sequences are short and of fixed length, we select the fixed sinusoidal approach. For each position pos and dimension i, the encoding is computed as:

$$\begin{aligned} PE(pos, 2i) &= \sin\left(\frac{pos}{10000^{2i/h}}\right) \\ PE(pos, 2i+1) &= \cos\left(\frac{pos}{10000^{2i/h}}\right) \end{aligned} \tag{6}$$

The computed position encodings and Layer Normalization are applied to obtain normalized query, key, and value matrices:

$$\begin{aligned} \hat{Q}_s &= \text{LayerNorm}_Q(Z_s + PE) \\ \hat{K}_t &= \text{LayerNorm}_K(Z_t + PE) \\ \hat{V}_t &= \text{LayerNorm}_V(Z_t + PE) \end{aligned} \tag{7}$$

where Z_s, Z_t, and Z_f represent feature matrices from different domains, PE denotes positional encodings, and LayerNorm operations help standardize feature distributions across domains.

This approach effectively mitigates distribution inconsistency issues during cross-domain feature fusion and enhances the stability and effectiveness of the attention mechanism. The mathematical expression is as follows:

$$H_{t\to s} = \text{Softmax}\left(\frac{\hat{Q}_s \hat{K}_t^\top}{\sqrt{d_k}}\right)\hat{V}_t \tag{8}$$

where $H_{t\to s}$ represents the cross-attention output from domain t to domain s, d_k is the dimension of the key matrix, and $\sqrt{d_k}$ prevents numerical issues caused by inner product operations. In this module, we construct three bidirectional cross-attention sub-modules responsible for feature interactions between temporal-spatial, temporal-frequency, and spatial-frequency domains.

After completing cross-attention calculations across the three domains, each cross-attention output is further refined by a two-layer feed-forward network with a residual connection:

$$Y_{t\to s} = H_{t\to s} + W_2\,\text{GELU}\big(W_1\,\text{LayerNorm}(H_{t\to s}) + b_1\big) + b_2 \tag{9}$$

where W_1 and W_2 are weight matrices, and b_1 and b_2 are bias vectors. This operation enhances non-linear feature transformation while preserving the original cross-domain interaction signal through the residual path.

Finally, TD-CAT outputs the results of mutual transfers between the three domains, denoted as $Y_{t\to s}$, $Y_{t\to f}$, $Y_{s\to f}$, $Y_{s\to t}$, $Y_{f\to t}$, and $Y_{f\to s}$.

Multi-Domain Low-Rank Feature Fusion

To effectively integrate feature information from temporal, frequency and spatial domains while reducing cross-domain fusion redundancy, we propose a Multi-Domain Low-Rank Feature Fusion (MLRFF) module. This module is based on the low-rank tensor decomposition concept (Liu et al., 2018), achieving complementary integration of multi-domain features through channel-wise low-rank fusion operations. Specifically, for temporal features $Z_t \in \mathbb{R}^{C\times h}$, frequency features $Z_f \in \mathbb{R}^{C\times h}$, spatial features $Z_s \in \mathbb{R}^{C\times h}$ and their corresponding cross-attention outputs $Y_{s\to t}, Y_{f\to t}, Y_{s\to f}, Y_{t\to f}, Y_{t\to s}, Y_{f\to s}$, MLRFF adopts a channel-wise low-rank tensor decomposition strategy. The low-rank factorization provides a structured fusion mechanism that mitigates redundant cross-domain mixing while preserving discriminative information.

To integrate temporal, frequency, and spatial representations while limiting redundant cross-domain mixing, we propose a Multi-Domain Low-Rank Feature Fusion (MLRFF) module. Inspired by Low-rank Multimodal Fusion (LMF) (Liu et al., 2018), MLRFF performs channel-wise low-rank fusion over three inputs per branch: the intra-domain feature and two incoming cross-attention messages. Let $Z_t, Z_f, Z_s \in \mathbb{R}^{C\times h}$ denote the domain-specific features, and let $Y_{a\to b} \in \mathbb{R}^{C\times h}$ denote the cross-attention output transferred from domain a to domain b.

For each channel $c \in \{1,\dots,C\}$, the temporal fusion branch combines $\{Z_t^c, Y_{s\to t}^c, Y_{f\to t}^c\}$ using a rank-r low-rank fusion:

$$H_t^c = \sum_{i=1}^{r}\Big((U_{1,t}^{(i)}\hat{Z}_t^c)\odot(U_{2,t}^{(i)}\hat{Y}_{s\to t}^c)\odot(U_{3,t}^{(i)}\hat{Y}_{f\to t}^c)\Big), \tag{10}$$

where $\hat{x} = [x;1] \in \mathbb{R}^{h+1}$ appends a constant 1 to model bias, $\odot$ denotes element-wise multiplication, and $U_{m,t}^{(i)} \in \mathbb{R}^{h\times(h+1)}$ are modality-specific factor matrices for rank i. The frequency and spatial branches are computed analogously using inputs $\{Z_f^c, Y_{t\to f}^c, Y_{s\to f}^c\}$ and $\{Z_s^c, Y_{t\to s}^c, Y_{f\to s}^c\}$, yielding $H_f, H_s \in \mathbb{R}^{C\times h}$.

After low-rank fusion, we concatenate the three branches along channels to obtain $H = [H_t; H_f; H_s] \in \mathbb{R}^{3C\times h}$. A self-attention "fuse Transformer" is then applied for global aggregation, producing $\tilde{H}$ for downstream classification. Experimental results demonstrate that this module reduces model parameters and computational costs while preserving rich feature interactions.

Classifier Module

We adopt a lightweight two-layer MLP to map the aggregated fused representation to the predicted label distribution. Specifically, we flatten $\tilde{H}$ into a vector $h_{\text{cls}} = \text{Flatten}(\tilde{H})$, and obtain the final prediction $\hat{y}$ via Softmax:

$$\hat{y} = \text{Softmax}(W_2\,\text{ReLU}(\text{Dropout}(W_1 h_{\text{cls}} + b_1)) + b_2) \tag{11}$$

where W_1, W_2, b_1, and b_2 are learnable parameters, and $\hat{y}$ denotes the predicted class distribution.

Experiments

This section describes the datasets, experimental setup, preprocessing, and evaluation metrics, and then presents the results and discussion.

Datasets

SADT The Sustained-Attention Driving Task (SADT) dataset was collected from 27 participants aged 22–28 during a 90-minute driving simulator experiment. EEG signals were recorded using a 32-channel cap at 500 Hz. Following the protocol of Cui et al. (Cui et al., 2022), we applied a 1–50 Hz band-pass filter and performed artifact removal. We then segmented 3-second EEG epochs immediately preceding each lane-departure event and assigned an "alert" or "fatigued" label based on the driver's reaction time. Consistent with the same protocol, we retained only participants with sufficient usable artifact-free epochs and complete event annotations for LOSO evaluation. The resulting dataset contains 2,022 labeled trials from 11 participants, with each trial represented as a 30-channel × 384-time-point segment.

SEED-VIG The SEED-VIG dataset includes EEG data collected from 21 participants (average age 23.3) during a 118-minute simulated driving experiment in virtual reality. EEG signals were recorded using a 17-channel Neuroscan system at 1000 Hz, resampled to 200 Hz, and then downsampled to 128 Hz and filtered following Cui et al. (Cui et al., 2023). Three-second EEG samples were extracted and labeled using the PERCLOS index. Following the same methodology, we kept only subjects whose recordings yielded complete labeled

Table 1: Performance comparison on different datasets

Model	SADT		SEED-VIG		MAT	
	Accuracy	F1	Accuracy	F1	Accuracy	F1
SVM	0.661 ± 0.076	0.653 ± 0.078	0.751 ± 0.072	0.740 ± 0.082	0.625 ± 0.121	0.600 ± 0.142
EEGNet	0.730 ± 0.072	0.731 ± 0.067	0.818 ± 0.088	0.804 ± 0.112	0.646 ± 0.112	0.620 ± 0.192
Conformer	0.712 ± 0.112	0.731 ± 0.101	0.840 ± 0.087	0.831 ± 0.098	0.693 ± 0.146	0.670 ± 0.200
EEG-Deformer	0.751 ± 0.099	0.726 ± 0.131	0.822 ± 0.099	0.793 ± 0.153	0.703 ± 0.144	0.673 ± 0.207
TFormer	0.797 ± 0.061	0.805 ± 0.056	0.912 ± 0.044	0.911 ± 0.045	**0.825** ± 0.119	0.826 ± 0.130
MATRN	0.754 ± 0.061	0.747 ± 0.058	0.875 ± 0.075	0.864 ± 0.104	0.807 ± 0.135	0.797 ± 0.174
Ours	**0.807** ± 0.076	**0.814** ± 0.063	**0.926** ± 0.051	**0.925** ± 0.052	0.824 ± 0.125	**0.835** ± 0.111

samples after preprocessing and quality control. The resulting subset contains 4,566 labeled samples from 12 subjects.

MAT Dataset The Mental Arithmetic Task (MAT) dataset was collected from 36 participants aged 18–26 during a mental arithmetic task. EEG signals were recorded using a 19-channel Neurocom system at 500 Hz. Following the methodology of (Ding et al., 2024), the data were filtered and artifacts were removed. Each cognitive load trial lasted 60 seconds, and the trials were segmented using a 4-second sliding window. The dataset includes 2,088 samples.

Implementation Details

The experiments were conducted on an Apple Silicon M3 using PyTorch 2.7.0. We used the leave-one-subject-out (LOSO) cross-validation method on the three public datasets above. The cross-domain attention blocks used 8 attention heads, and the fusion self-attention block used 4 heads. The hidden size for cross-domain attention was set to h=128, the fused representation was projected to 192 dimensions before the final fusion attention stage, and the low-rank fusion rank was set to r=3. Dropout was set to 0.3 in the Transformer and GCN modules, 0.15 in the low-rank fusion module, and 0.5 in the classifier head. We optimized the model with cross-entropy loss and Adam. The learning rate was 5×10^{-4}, β_1 was 0.9, and β_2 was 0.99. The batch size was 32, and the model was trained for 60 epochs. We report the mean accuracy (ACC.) and binary F1-score together with their corresponding variances across LOSO folds.

Baselines

All baseline results were re-implemented under the same preprocessing pipeline and identical LOSO splits as TD-CAT. Table 1 contrasts the proposed TD-CAT model against five representative baselines across the SADT, SEED-VIG, and MAT datasets. On SADT, it attains 80.7% accuracy and 81.4% F1 score, surpassing the strongest baseline TFormer by 1.0% and 0.9%, respectively. On SEED-VIG, the method reaches 92.6% accuracy and 92.5% F1 score, leading TFormer by 1.4%. We find that 75% of participants achieve accuracy above 88%. The lowest accuracy among participants is 83%. On MAT, it reports 82.4% accuracy and 83.5% F1 score; the accuracy matches the state-of-the-art, while the F1 score improves by 0.9%. These results confirm the effectiveness and robustness of the tri-domain fusion approach.

Ablation Study

To verify the effectiveness of the proposed model, we conducted ablation studies on three public datasets (SADT, SEED-VIG, and MAT) under the same evaluation protocol. For brevity, we present detailed ablation results on SEED-VIG in this section. Results show the effectiveness of each module across datasets.

We first examine how each module contributes to the overall performance by progressively adding components to a basic classifier. Specifically, we designed four settings:

1. Initial model: a basic classification backbone without any proposed modules;
2. Combination 1: adding the temporal-spatial-frequency multi-perspective encoder;
3. Combination 2: further introducing the tri-domain cross-attention Transformer;
4. Full model: including all proposed modules, with feature fusion enabled.

Table 2: Performance comparison of different module combinations

Module Combination	Encoder	TD-CAT	Feature Fusion	SEED-VIG	
				Accuracy	Change
Initial	✗	✗	✗	0.892	-
Combination 1	✓	✗	✗	0.898	0.006
Combination 2	✓	✓	✗	0.915	0.017
Full Model	✓	✓	✓	0.926	0.011

The experimental results in Table 2 show clear improvements on SEED-VIG as modules are added stepwise. Compared with the initial model, adding only the temporal-spatial-frequency encoder raises accuracy by 0.6%. This gain confirms the importance of multi-perspective feature extraction. Adding the TD-CAT module to the encoder further improves accuracy by 1.7%, indicating that the tri-domain cross-attention mechanism effectively captures correlations between

domains. The full model, relative to the encoder-plus-TD-CAT version, achieves an additional 1.1% accuracy gain, validating the positive contribution of the feature fusion module. Overall, the ablation study confirms that each proposed module is essential. The modules also interact synergistically to enhance performance.

To further analyze the role of each domain in the input representation, we additionally performed a domain-level comparison by removing temporal, frequency, or spatial features. For a removed domain, we replaced its input with a zero vector of the same dimension. The results are summarized in Table 3.

Table 3: Performance comparison of different domain combinations

Time	Freq.	Spatial	SEED-VIG	
			Acc.	Change
✓	✗	✗	0.798	-0.128
✗	✓	✗	0.811	-0.115
✗	✗	✓	0.785	-0.141
✗	✓	✓	0.908	-0.018
✓	✗	✓	0.868	-0.058
✓	✓	✗	0.910	-0.016
✓	✓	✓	0.926	-

Single-modality settings are consistently inferior to multi-domain configurations. Frequency-only performs best (81.1%), followed by time-only (79.8%) and spatial-only (78.5%), highlighting the necessity of cross-domain fusion. When frequency-domain features are withheld, accuracy drops by 5.8%, from 92.6% to 86.8%, indicating that the model cannot fully reconstruct spectral information from raw temporal or spatial inputs. In contrast, the absence of temporal or spatial features incurs smaller yet still significant reductions: removal of temporal features lowers accuracy to 90.8%, while removal of spatial features lowers accuracy to 91.0%. This attenuated degradation is consistent with the fact that both the temporal convolution module and the GCN layer receive raw EEG channels as inputs and can therefore recover a portion of the missing temporal or spatial cues internally.

Visualization of Key Brain Regions

Since we employed the data-driven GCN Branch, we can visualize the learned adjacency matrix weights between channels to investigate the significance of brain regions corresponding to different electrodes. We first exported all LOSO adjacency matrices from the SADT dataset and performed normalization to constrain the weight values between -1 and 1. The visualization of the adjacency matrix weights heatmap is shown in Fig. 2. Through analysis of the visualization results, we identified several significant channel connections with absolute weight values exceeding 0.8, including FP1-F3, F4-CP4, FC4-O2, CPZ-T5, and FCZ-O1, encompassing a total of 8

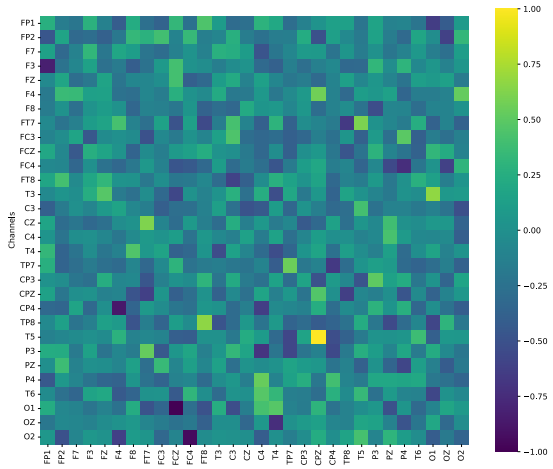

Figure 2: Adjacency Matrix Weight Heatmap

electrodes. These electrodes are primarily distributed across the frontal, occipital, parietal, and temporal lobes of the brain. This finding aligns with previous research establishing the crucial role of these brain regions in visual information processing and fatigue state transitions (Tran et al., 2020; H. Wang et al., 2021; Yue et al., 2022).

Conclusion

In this paper, we propose TD-CAT, a tri-domain cross-attention Transformer for decoding cognitive states from EEG. TD-CAT first constructs domain-specific representations from the temporal, spectral, and spatial perspectives. Temporal and spectral convolutions capture dynamic and oscillatory EEG patterns, while a cognition-inspired graph convolution encodes electrode relations defined by scalp topology. TD-CAT then applies tri-domain cross-attention to model cross-domain dependencies and uses channel-wise low-rank fusion to integrate complementary information with reduced redundancy. Under within-dataset leave-one-subject-out evaluation on SADT and SEED-VIG for driver fatigue and on MAT for cognitive load, TD-CAT exceeds previous strong Transformer baselines in F1-score. TD-CAT achieves 80.7% accuracy on SADT, 92.6% on SEED-VIG, and 82.4% on MAT. Ablation experiments further verify the contribution of each EEG domain. Analyses of learned spatial connections reveal region-level interaction patterns that are consistent with established signatures of fatigue and load. These results indicate that explicit tri-domain interaction modeling improves the accuracy and interpretability of cross-subject EEG decoding across multiple datasets and task settings, which supports practical EEG-based monitoring in sustained and safety-critical tasks.

References

Benelli, A., Memoli, C., Neri, F., Romanella, S. M., Cinti, A., Giannotta, A., Lomi, F., Scoccia, A., Pandit, S., Zambetta, R. M., et al. (2024). Reduction of cognitive fatigue and

improved performance at a vr-based driving simulator using trns. *Iscience*, *27*(9).

Berka, C., Levendowski, D. J., Lumicao, M. N., Yau, A., Davis, G., Zivkovic, V. T., Olmstead, R. E., Tremoulet, P. D., & Craven, P. L. (2007). Eeg correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation, space, and environmental medicine*, *78*(5), B231–B244.

Cao, Z., Chuang, C.-H., King, J.-K., & Lin, C.-T. (2019). Multi-channel eeg recordings during a sustained-attention driving task. *Scientific data*, *6*(1), 19.

Cui, J., Lan, Z., Liu, Y., Li, R., Li, F., Sourina, O., & Müller-Wittig, W. (2022). A compact and interpretable convolutional neural network for cross-subject driver drowsiness detection from single-channel eeg. *Methods*, *202*, 173–184.

Cui, J., Yuan, L., Li, R., Wang, Z., Yang, D., & Jiang, T. (2023). Benchmarking eeg-based cross-dataset driver drowsiness recognition with deep transfer learning. *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 1–6.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

Ding, Y., Li, Y., Sun, H., Liu, R., Tong, C., Liu, C., Zhou, X., & Guan, C. (2024). Eeg-deformer: A dense convolutional transformer for brain-computer interfaces. *IEEE Journal of Biomedical and Health Informatics*.

Gao, D., Li, P., Wang, M., Liang, Y., Liu, S., Zhou, J., Wang, L., & Zhang, Y. (2023). Csf-gtnet: A novel multi-dimensional feature fusion network based on convnext-gelu-bilstm for eeg-signals-enabled fatigue driving detection. *IEEE Journal of Biomedical and Health Informatics*, *28*(5), 2558–2568.

Gao, D., Tang, X., Wan, M., Huang, G., & Zhang, Y. (2023). Eeg driving fatigue detection based on log-mel spectrogram and convolutional recurrent neural networks. *Frontiers in Neuroscience*, *17*, 1136609.

Hendrycks, D., & Gimpel, K. (2016). Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Hu, F., Zhang, L., Yang, X., & Zhang, W.-A. (2024). Eeg-based driver fatigue detection using spatio-temporal fusion network with brain region partitioning strategy. *IEEE Transactions on Intelligent Transportation Systems*.

Khateeb, M., Anwar, S. M., & Alnowami, M. (2021). Multi-domain feature fusion for emotion classification using deap dataset. *Ieee Access*, *9*, 12134–12142.

Kunasegaran, K., Ismail, A. M. H., Ramasamy, S., Gnanou, J. V., Caszo, B. A., & Chen, P. L. (2023). Understanding mental fatigue and its detection: A comparative analysis of assessments and tools. *PeerJ*, *11*, e15744.

Li, C., Tang, T., Pan, Y., Yang, L., Zhang, S., Chen, Z., Li, P., Gao, D., Chen, H., Li, F., et al. (2024). An efficient graph learning system for emotion recognition inspired by the cognitive prior graph of eeg brain network. *IEEE Transactions on Neural Networks and Learning Systems*, *36*(4), 7130–7144.

Li, J., Chen, N., Zhu, H., Li, G., Xu, Z., & Chen, D. (2024). Incongruity-aware multimodal physiology signals fusion for emotion recognition. *Information Fusion*, *105*, 102220.

Liu, Z., Shen, Y., Lakshminarasimhan, V. B., Liang, P. P., Zadeh, A., & Morency, L.-P. (2018). Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*.

Longo, L., Wickens, C. D., Hancock, G., & Hancock, P. A. (2022). Human mental workload: A survey and a novel inclusive definition. *Frontiers in psychology*, *13*, 883321.

Paas, F., & Van Merrienboer, J. J. (2020). Cognitive-load theory: Methods to manage working memory load in the learning of complex tasks. *Current directions in psychological science*, *29*(4), 394–398.

Salihu, A. T., Hill, K. D., & Jaberzadeh, S. (2022). Neural mechanisms underlying state mental fatigue: A systematic review and activation likelihood estimation meta-analysis. *Reviews in the Neurosciences*, *33*(8), 889–917.

Sartipi, S., Torkamani-Azar, M., & Cetin, M. (2021). Eeg emotion recognition via graph-based spatio-temporal attention neural networks. *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 571–574. https://doi.org/10.1109/EMBC46164.2021.9629628

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

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive science*, *12*(2), 257–285.

Tran, Y., Craig, A., Craig, R., Chai, R., & Nguyen, H. (2020). The influence of mental fatigue on brain activity: Evidence from a systematic review with meta-analyses. *Psychophysiology*, *57*(5), e13554.

Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P., & Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. *Proceedings of the conference. Association for computational linguistics. Meeting*, *2019*, 6558.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wang, H., Xu, L., Bezerianos, A., Chen, C., & Zhang, Z. (2021). Linking attention-based multiscale cnn with dynamical gcn for driving fatigue detection. *IEEE Transactions on Instrumentation and Measurement*, *70*, 1–11. https://doi.org/10.1109/TIM.2020.3047502

Wang, P., Nam, J.-S., & Han, X. (2024). Development of a comprehensive fatigue detection model for beekeeping ac-

tivities based on deep learning and eeg signals. *Computers and Electronics in Agriculture*, *225*, 109265.

Ye, C., Yin, Z., Zhao, M., Tian, Y., & Sun, Z. (2022). Identification of mental fatigue levels in a language understanding task based on multi-domain eeg features and an ensemble convolutional neural network. *Biomedical Signal Processing and Control*, *72*, 103360.

Yue, K., Guo, M., Liu, Y., Hu, H., Lu, K., Chen, S., & Wang, D. (2022). Investigate the neuro mechanisms of stereoscopic visual fatigue. *IEEE Journal of Biomedical and Health Informatics*, *26*(7), 2963–2973. https://doi.org/10.1109/JBHI.2022.3161083

Zheng, W.-L., & Lu, B.-L. (2017). A multimodal approach to estimating vigilance using eeg and forehead eog. *Journal of Neural Engineering*, *14*(2), 026017. http://stacks.iop.org/1741-2552/14/i=2/a=026017

Zyma, I., Tukaev, S., Seleznov, I., Kiyono, K., Popov, A., Chernykh, M., & Shpenkov, O. (2019). Electroencephalograms during mental arithmetic task performance. *Data*, *4*(1), 14.