Dynamic Convolution and Graph-Coupled Attention for Cross-Subject EEG-Vision Decoding

Tianyu Zhang[1]
tianyu.zhang2@durham.ac.uk
Fan Wan[†2]
fan.wan.uk@gmail.com
Kaili Sun[1]
kaili.sun@durham.ac.uk
Xingyu Miao[1]
xingyu.miao@durham.ac.uk
Yueming Sun[1]
yueming.sun@durham.ac.uk
Minye Shao[1]
minye.shao@durham.ac.uk
Yang Long[†1]
yang.long@durham.ac.uk

[1] Department of Computer Science
Durham University
Durham, UK

[2] Central Research Institute
Tongfang Knowledge Network Digital Technology Co., Ltd.
China National Nuclear Corporation
Beijing, China

Abstract

Electroencephalography (EEG) offers a non-invasive route to visual object decoding, but practical deployment is hampered by the signals' non-stationarity, low signal-to-noise ratio and pronounced inter-subject variability. Existing models employ fixed convolutional filters and therefore generalize poorly across subjects. We introduce *ECHO-Net*—an adaptive, hierarchically organised network that assembles dynamic convolutional kernels, conditioning them on each incoming EEG trial to capture transient neural dynamics. A cross-modal contrastive objective aligns the resulting representations with CLIP image embeddings, while a channel–filter attention mechanism emphasises task-relevant electrodes and time–frequency bands. To regularise spatial structure, an embedded EEG-GAT module propagates information over a fully connected electrode graph, producing more consistent cross-subject features. Evaluated on the 200-way THINGS-EEG benchmark, our method attains 18.5% top-1 and 44.1% top-5 cross-subject accuracy—surpassing the strongest prior approach by 2.9% top-1.

1 Introduction

Electroencephalography (EEG) offers a non-invasive window onto visual object representations and underpins a broad spectrum of brain–computer-interface (BCI) technologies

[†]: Corresponding authors.

[2, 3, 7, 22]. Beyond basic neuroscience, EEG-based object decoding is increasingly deployed for early neurological screening and adaptive human–computer interaction [27]. Yet EEG signals are marred by low signal-to-noise ratios, non-stationarity and pronounced inter-subject variability. Consequently, practical decoders must capture trial-specific dynamics without incurring the high cost of retraining for every new user.

Previous studies have explored domain-adaptation techniques, advanced neural filters and deep-learning–driven feature extractors to mitigate these issues, but each offers only partial relief [16, 21, 24]. Despite promising progress, current EEG decoding frameworks still suffer from three fundamental limitations. **First**, conventional decoding pipelines rely on handcrafted features and shallow classifiers such as Support Vector Machines and Linear Discriminant Analysis [1, 4]. While efficient, these methods lack the capacity to model EEG's non-linear and dynamic spatio-temporal patterns. Deep learning models [16, 23] improve feature extraction using convolutional hierarchies. Nevertheless, most methods rely on fixed convolutional kernels, which are learned once and applied uniformly to all test samples. Although effective for structured, stationary data (e.g., images), this approach is ill-suited for EEG, which is inherently non-stationary, subject-specific, and evolves across trials and sessions. **Second**, existing multimodal alignment methods typically apply uniform attention across EEG channels and time steps. This design assumes that all electrodes and temporal segments are equally informative, which struggles to account for the spatio-temporal selectivity of brain activity, such as NICE-SA [26]. In practice, task-relevant cognitive signals are concentrated in specific channels or time windows. Uniform weighting thus introduces redundancy and noise, degrading semantic alignment between EEG and visual features. **Third**, prior studies [26] have incorporated Graph Attention Networks (GATs) to encode EEG's spatial topology. However, GATs are often used as standalone spatial modules, loosely coupled with temporal and semantic processing components. This architectural decoupling limits their ability to model inter-electrode dependencies in coordination with dynamic signal characteristics, thereby weakening decoding consistency across subjects and electrode configurations.

To systematically address these challenges, we propose *ECHO-Net* (EEG Cross-subject Hierarchical Organised Network), a novel dynamic convolutional neural network tailored for adaptive EEG decoding. The framework integrates dynamic convolution, attention mechanisms, and graph-based modelling within a unified hierarchical structure to jointly capture trial-specific neural patterns, highlight task-relevant features, and model inter-channel spatial dependencies. To overcome the limitations of non-stationary signals and subject variability, we introduce dynamic convolution into EEG feature extraction. Inspired by recent successes in computer vision [5, 17, 25, 28, 31], our PatchEmbedding module with Dynamic Convolution, *PEDC*, aggregates convolutional kernels based on input-specific EEG patterns. This design enables trial-aware and subject-adaptive representation learning, improving generalisation under non-stationary conditions.

To mitigate the noise and redundancy introduced by uniform attention, we develop a multimodal contrastive learning framework enhanced by *Channel-Filter Attention*. Channel attention selectively amplifies spatially informative electrodes via global pooling and non-linear gating, while filter attention emphasises salient segments aligned with visual semantics.

To strengthen spatial relational modelling within the decoding pipeline, we integrate *EEG-GAT* into the *ECHO-Net* pipeline. Unlike prior decoupled GAT modules, *EEG-GAT* operates in tandem with dynamic convolution and attention mechanisms, enabling joint spatio-temporal-semantic modelling. This unified pipeline enables spatially consistent, se-

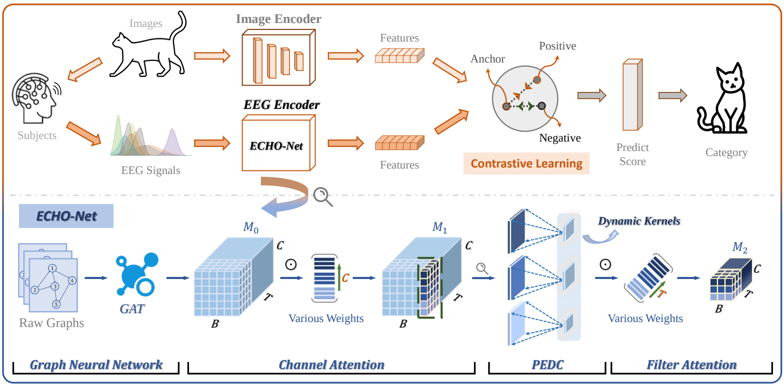

Figure 1: Overview of the proposed ECHO-Net framework, integrating dynamic convolution, attention, and contrastive learning. The architecture highlights modular interactions between spatial encoding, dynamic filtering, and multimodal alignment.

mantically aligned, and subject-invariant EEG representations for cross-modal decoding. The overall architecture is illustrated in Figure 1, depicting the end-to-end pipeline from raw EEG input to multimodal contrastive learning.

Our contributions are summarized as follows:

- **Dynamic-convolution backbone.** We propose a backbone that assembles dynamic convolution kernels, capturing EEG non-stationarity and generalizing across subjects without per-user fine-tuning.
- **Contrastive learning with channel–filter attention.** A multimodal contrastive objective, guided by channel–filter attention, focuses on the most informative electrodes and time–frequency bands, tightly aligning EEG features with visual embeddings.
- **Integrated EEG-GAT reasoning.** A lightweight graph attention layer interleaved with the backbone enables joint spatio-temporal semantic reasoning, generating structurally consistent representations.

2 Related Work

Dynamic Convolution for EEG. Deep learning has significantly advanced EEG decoding by capturing non-linear, non-stationary patterns in brain signals. Classical architectures [16, 23] demonstrated that compact CNNs can efficiently extract spatio-temporal EEG features. More recently, hybrid models like Local-Global Net [29] captured multi-scale dependencies by jointly modelling local oscillations and global brain dynamics. ADFR [20] leveraged deep CNNs and domain adaptation to enhance cross-subject generalization. Despite these advances, most existing methods utilise static convolutional kernels that struggle to model the intrinsic temporal variability and subject heterogeneity in EEG. To address this, dynamic convolution [5, 12, 17, 31] has emerged as a promising solution, enabling kernel weights to adapt based on input context. EEG-specific implementations. Our work introduces PEDC, a data-driven approach that dynamically generates convolutional filters conditioned on raw EEG inputs. PEDC improves robustness to cross-subject variation by capturing trial-specific transient dynamics in a learnable, end-to-end manner.

Contrastive Multimodal EEG-Vision Alignment. Contrastive learning has recently demonstrated strong potential in aligning EEG signals with semantic visual representations. Early efforts such as BENDR [14] applied transformer-based self-supervised pretraining to generalize across tasks. NICE-SA [26] employed dual encoders with attention for zero-shot EEG-image classification, while other recent works have begun exploring large-scale vision-language models (e.g., CLIP) to facilitate semantic consistency across modalities. EEG-Diffusion [18] further introduced generative modelling to reconstruct images from EEG embeddings. However, many of these approaches [8, 9, 13, 14, 18] rely on static attention or fixed fusion strategies, overlooking the fact that EEG signals exhibit time-varying, subject-specific, and non-stationary patterns. To address this, we propose a Multimodal Contrastive Attention framework incorporating channel-wise and temporal attention prior to contrastive alignment. This allows the network to selectively emphasise task-relevant EEG segments and channels, enhancing alignment with vision features while reducing semantic noise.

Graph Attention for Spatial EEG Modelling. The spatial organisation of EEG sensors reflects underlying cortical regions, motivating graph-based models that capture inter-electrode dependencies. Early efforts include EEG-GAT [6], which embedded graph attention into EEGNet for adaptive spatial filtering. STGATE [19] and STAFNet [11] combined GATs with transformers to model complex spatio-temporal interactions. Recent studies have also explored meta-learning strategies combined with adaptive graph structures for zero-shot cross-subject decoding, illustrating that flexible topology adjustment can significantly enhance generalization. Despite promising results [6, 19, 26], many graph-based methods require predefined connectivity or static graphs, limiting their adaptability across subjects or sessions. To overcome this, we integrate EEG-GAT with PEDC and channel-filter attention in a unified encoder. Our architecture implicitly learns graph structures from raw EEG and modulates spatial dependencies based on input characteristics.

3 Methodology

We treat each trial as a multivariate time series $X \in \mathbb{R}^{C\times T}$ with C electrodes and T samples. ECHO-Net processes X in four stages. First, **PEDC**: we split X into P non-overlapping temporal patches ($T' = T/P$) to obtain $X_{\text{patch}} \in \mathbb{R}^{C\times P\times T'}$ and assemble a trial-conditioned dynamic kernel $K(X) = \sum_{i=1}^{N} \alpha_i(X) K_i$. Unless otherwise stated, we use *2D dynamic convolution* with candidate kernels $K_i \in \mathbb{R}^{k_c\times k_t}$ operating jointly over the channel$\times$time axes of X_{patch}; for brevity, we denote only the spatial support $(k_c \times k_t)$ and omit the in/out-channel dimensions, which match those of the replaced `Conv2d` layer. Convolving $K(X)$ with each patch and projecting yields a compact representation $Z \in \mathbb{R}^{C\times D}$. Second, **Channel–filter attention**: we re-weight Z across electrodes and temporal filters to emphasise task-relevant components, producing Z'_{EEG}. Third, **EEG-GAT**: we perform attention-based message passing on a 64-node fully connected electrode graph (no self-loops), yielding $H \in \mathbb{R}^{C\times D}$. Finally, **Alignment and retrieval**: global pooling and a linear head g_{EEG} map H into the joint embedding space, where an InfoNCE loss aligns it to frozen CLIP image features $g_{\text{IMG}}(F_{\text{img}})$. At test time, we perform zero-shot retrieval by comparing EEG embeddings against a gallery of CLIP image embeddings (Top-k).

3.1 PEDC: Adaptive Feature Extraction

The non-stationary nature of EEG signals makes static convolutional kernels less effective in capturing the dynamic patterns of brain activity. To address this, PEDC adaptively adjusts convolutional kernels conditioned directly on the EEG input, enabling precise capture of trial-specific patterns.

Given an EEG signal represented as $\mathbf{X} \in \mathbb{R}^{C \times T}$, where C denotes the number of EEG channels and T is the time dimension, the signal is first divided into P non-overlapping patches, each of length $T' = T/P$:

$$\mathbf{X}_{\text{patch}} \in \mathbb{R}^{C \times P \times T'} = \text{PatchSplit}(\mathbf{X}), \quad \mathbf{Z} \in \mathbb{R}^{C \times D} = \text{Proj}(\text{Pool}_{T'}(\mathbf{X}_{\text{patch}})), \tag{1}$$

where D is the embedding dimension, T' is the length of each patch, and $\text{Proj}(\cdot)$ denotes a learnable linear projection that reduces redundancy while retaining discriminative EEG patterns relevant to downstream recognition. The segmented representation $\mathbf{X}_{\text{patch}} \in \mathbb{R}^{C \times P \times T'}$ is then pooled and projected into a compact feature space.

Unlike kernel generation methods that synthesize filters directly via a meta-network [30], our dynamic convolution adopts an attention-based aggregation mechanism. Specifically, we define N shared candidate kernels $K_1, \ldots, K_N$, and compute input-dependent weights $\alpha_1(\mathbf{X}), \ldots, \alpha_N(\mathbf{X})$ through a lightweight projection over the global average pooled EEG input. The final convolution kernel is then a convex combination of the candidate kernels, where the effective kernel is formulated as a weighted sum of N predefined kernels:

$$K(\mathbf{X}) = \sum_{i=1}^{N} \alpha_i(\mathbf{X}) K_i, \quad \text{s.t.} \quad \sum_{i=1}^{N} \alpha_i(\mathbf{X}) = 1, \tag{2}$$

This formulation allows the network to adaptively modulate its receptive field per input trial, capturing dynamic and non-stationary EEG patterns without increasing kernel parameters. Unless otherwise stated, we use *2D dynamic convolution* with candidate kernels $K_i \in \mathbb{R}^{k_c \times k_t}$ operating jointly on the channel×time axes of X_{patch}; for brevity, we denote only the spatial support $(k_c \times k_t)$ and omit the in/out-channel dimensions, which match those of the replaced `Conv2d` layer. Note that the kernel generation is conditioned solely on EEG features, without reliance on subject ID embedding or other external metadata. Here, $\alpha_i(\mathbf{X})$ are input-dependent weights computed via softmax:

$$\alpha_i(\mathbf{X}) = \text{softmax}(\mathbf{W}_i(\text{GAP}(\mathbf{X})) + b_i), \tag{3}$$

where $\mathbf{W}_i \in \mathbb{R}^{1 \times C}$ and b_i are learnable parameters. This adaptive mechanism improves feature extraction consistency and cross-subject generalization. This module enables context-aware kernel selection tailored to input variability, thereby enhancing the consistency and cross-subject generalization of EEG feature extraction.

3.2 Contrastive Learning: Semantic EEG-Image Alignment with Channel-Filter Attention

We map EEG and image embeddings into a shared space via two linear projections, g_{EEG} and g_{IMG}, and train them with the contrastive loss:

$$\mathcal{L} = -\frac{1}{N} \sum_i \log \frac{\exp(\text{sim}(g_{\mathsf{EEG}}(\mathbf{Z}_{\mathsf{EEG}}^{(i)}), g_{\mathsf{IMG}}(\mathbf{Z}_{\mathsf{IMG}}^{(i)}))/\tau)}{\sum_j \exp(\text{sim}(g_{\mathsf{EEG}}(\mathbf{Z}_{\mathsf{EEG}}^{(i)}), g_{\mathsf{IMG}}(\mathbf{Z}_{\mathsf{IMG}}^{(j)}))/\tau)}, \tag{4}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity and τ is a temperature scaling parameter. A lightweight channel–filter attention module is applied to $\mathbf{Z}_{\mathsf{EEG}}$ before projection, selectively amplifying spatial–temporal components that correlate with visual semantics. To improve the discriminative capacity of EEG features prior to contrastive alignment, we introduce two complementary attention mechanisms:

Channel Attention. This module emphasises spatially informative EEG electrodes. A global average pooling is applied across the temporal axis, followed by a two-layer MLP with non-linear activations to generate attention weights:

$$\boldsymbol{\alpha}_{\mathrm{c}} = \sigma(W_2\,\delta(W_1\,\mathrm{GAP}(\boldsymbol{Z}_{\mathsf{EEG}}))), \quad \boldsymbol{Z}'_{\mathsf{EEG}} = \boldsymbol{Z}_{\mathsf{EEG}} \odot \boldsymbol{\alpha}_{\mathrm{c}}, \tag{5}$$

where $\delta(\cdot)$ and $\sigma(\cdot)$ represent ReLU and sigmoid activation functions, respectively, and W_1, W_2 are learnable parameters.

Filter Attention. To further highlight temporally salient segments, we apply scaled dot-product attention over the temporal dimension:

$$\boldsymbol{\alpha}_{\mathrm{f}} = \mathrm{softmax}\left(\frac{\mathbf{Q}(\boldsymbol{Z}_{\mathsf{EEG}}) \cdot \mathbf{K}(\boldsymbol{Z}_{\mathsf{EEG}})^{\top}}{\sqrt{d}}\right), \quad \boldsymbol{Z}'_{\mathsf{EEG}} = \boldsymbol{Z}_{\mathsf{EEG}} \odot \boldsymbol{\alpha}_{\mathrm{f}}, \tag{6}$$

where $\mathbf{Q}(\cdot)$ and $\mathbf{K}(\cdot)$ denote learned linear projections, and d is the dimensionality of the feature vectors.

3.3 EEG-GAT: Modelling Inter-Electrode Spatial Dependencies

GATs are typically treated as standalone modules in EEG signal processing, loosely integrated with the broader architecture. Although prior work has used GATs to enhance spatial encoding, these modules remain decoupled from the temporal and semantic pipelines. In contrast, our proposed EEG-GAT is tightly coupled with both dynamic convolution and attention mechanisms, forming a unified architecture for spatially consistent and semantically enriched EEG decoding.

We represent the EEG electrode configuration as a graph $G = (V, E)$, where each node $i \in V$ corresponds to an EEG electrode. To model long-range dependencies and avoid reliance on predefined spatial priors, we adopt a fully connected topology: every electrode is connected to all others, i.e., $\mathcal{N}(i) = V \setminus \{i\}$. In our implementation, this translates into a fully connected, unweighted, and undirected graph with 64 nodes, excluding self-loops. The edge connectivity is fixed and consistent across trials, following the 10–20 system layout [26]. This design enables global attention-based message passing while allowing the model to flexibly learn cross-channel interactions without relying on anatomical assumptions. Each node feature $\mathbf{x}_i$ is derived from adaptively extracted spatio-temporal EEG representations. Node updates are computed via attention-weighted aggregation:

$$\mathbf{x}'_i = \sigma\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij} \mathbf{W}\mathbf{x}_j\right), \quad \alpha_{ij} = \frac{\exp\left(\mathrm{LeakyReLU}\left(\mathbf{a}^T [\mathbf{W}\mathbf{x}_i || \mathbf{W}\mathbf{x}_j]\right)\right)}{\sum_{k \in \mathcal{N}(i)} \exp\left(\mathrm{LeakyReLU}\left(\mathbf{a}^T [\mathbf{W}\mathbf{x}_i || \mathbf{W}\mathbf{x}_k]\right)\right)}. \tag{7}$$

where $\mathbf{W}$ is a learnable transformation and α_{ij} is the attention coefficient between electrodes i and j. The attention coefficients α_{ij} are dynamically computed for each input sample, as they depend on the input-specific node features $\mathbf{x}_i$ and $\mathbf{x}_j$. This ensures that the graph attention adapts on a per-trial basis, enabling the model to flexibly model inter-electrode dependencies that vary across different EEG samples.

4 Experiments

4.1 Experimental Setup

We evaluated the effectiveness and generalizability of our proposed framework on two large-scale public benchmarks: THINGS-EEG [26] and THINGS-MEG [10]. These datasets provide rich semantic diversity, and cover multiple neural modalities (EEG and MEG), making them ideal for assessing the model's ability to decode visual semantics. THINGS-EEG contains EEG recordings from 10 subjects using a Rapid Serial Visual Presentation (RSVP) paradigm, covering 1654 training and 200 test object categories. THINGS-MEG consists of MEG recordings from 4 subjects, involving 1854 training and 200 test concepts. EEG and MEG data underwent consistent preprocessing: band-pass filtering (0.1–100 Hz), baseline correction (200 ms pre-stimulus), downsampling (250 Hz), trial averaging for improved SNR, and z-score normalization. All experiments were conducted using PyTorch with reproducible random seeds (seed=42). Models were trained with Adam optimizer (2×10^{-4}, cosine annealing), batch size 512 for 200 epochs. We performed both subject-dependent (within-subject) and subject-independent (leave-one-subject-out cross-validation) analyses. Performance was measured using Top-1 and Top-5 accuracy, representing the exact prediction accuracy and semantic consistency, respectively. Notably, we follow a *zero-shot classification* setting, in which no class labels are used to supervise the EEG encoder during training. Instead, EEG and image features are aligned in a shared embedding space using contrastive learning. At test time, EEG embeddings are compared to a fixed set of image embeddings, and classification is performed via Top-k retrieval based on similarity. This enables semantic decoding without direct category-level supervision on EEG data. Statistical significance was assessed via paired t-tests, with significance thresholds clearly reported ($p < 0.05$, $p < 0.01$, $p < 0.001$). Our approach was benchmarked against several state-of-the-art baselines: BraVL [8], NICE [26], NICE-SA, and NICE-GA, representing competitive cross-modal neural decoding methods.

4.2 Overall Performance

The proposed model outperformed all baselines across both datasets and evaluation protocols. On THINGS-EEG, our framework achieved the highest subject-dependent accuracy, with an average Top-1 score of 18.5% and Top-5 score of 44.1% (Table 1). Compared to NICE-GA, which integrated graph attention for spatial modelling, our model yielded a 2.9% improvement in Top-1 accuracy ($p < 0.01$), highlighting the benefit of combining dynamic convolution, attention-guided refinement, and contrastive alignment. In the subject-independent setting, where inter-subject variability was a major challenge, our model achieved an average Top-1 accuracy of 9.1%, surpassing NICE (6.2%) and NICE-GA (5.9%). These results underscored the effectiveness of our architecture in learning robust, subject-invariant representations from EEG. To further assess the applicability of our model, we evaluated on the THINGS-MEG dataset. Despite significant modality differences, our model attained the highest subject-dependent performance with a Top-1 accuracy of 16.2% and Top-5 accuracy of 45.4% (Table 2). The improvements across both EEG and MEG suggested that our method effectively captured shared neural structures underpinning object-level visual semantics. This indicates the method's ability to learn representations suitable for either EEG or MEG, although training was performed independently on each modality.

Table 1: Zero-shot classification accuracy (Top-1/Top-5, %) across 10 subjects on THINGS-EEG under subject-dependent and subject-independent settings. We highlight the best, second-best, and third-best scores.

Method	Subject 1		Subject 2		Subject 3		Subject 4		Subject 5		Subject 6		Subject 7		Subject 8		Subject 9		Subject 10		Avg	
	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
Subject dependent - train and test on one subject																						
BraVL [[redacted]]	6.1	17.9	4.9	14.9	5.6	17.4	5.0	15.1	4.0	13.4	6.0	18.2	6.5	20.4	8.8	23.7	4.3	14.0	7.0	19.7	5.8	17.5
NICE [[redacted]]	12.3	36.6	10.4	33.9	13.1	39.0	16.4	47.0	8.0	26.9	14.1	40.6	15.2	42.1	20.0	49.9	13.3	37.1	14.9	41.9	13.8	39.5
NICE-SA [[redacted]]	13.3	40.2	12.1	36.1	15.3	39.6	15.9	49.0	9.8	34.4	14.2	42.4	17.9	43.6	18.2	50.2	14.4	38.7	16.0	42.8	14.7	41.7
NICE-GA [[redacted]]	15.2	40.1	13.9	40.1	14.7	42.7	17.6	48.9	9.0	29.7	16.4	44.4	19.1	43.1	20.3	52.1	14.1	39.7	19.6	46.7	15.6	42.8
ECHO-Net	17.0	43.0	20.5	41.0	16.5	45.0	19.0	49.5	14.0	32.5	20.0	44.0	20.5	47.5	24.5	51.5	14.5	38.5	18.0	48.5	18.5	44.1
Subject independent - leave one subject out for test																						
BraVL	2.3	8.0	1.5	6.3	1.4	5.9	1.7	6.7	1.5	5.6	1.8	7.2	2.1	8.1	2.2	7.6	1.6	6.4	2.3	8.5	1.8	7.0
NICE	7.6	22.8	5.9	20.5	6.0	22.3	6.3	20.7	4.4	18.3	5.6	22.2	5.6	19.7	6.3	22.0	5.7	17.6	8.4	28.3	6.2	21.4
NICE-SA	7.0	22.6	6.6	23.2	7.5	23.7	5.4	21.4	6.4	22.2	7.5	22.5	3.8	19.1	8.5	24.4	7.4	22.3	9.8	29.6	7.0	23.1
NICE-GA	5.9	21.4	6.4	22.7	5.5	20.1	6.1	21.0	4.7	19.5	6.2	22.5	5.9	19.1	7.3	25.3	4.8	18.3	6.2	26.3	5.9	21.6
ECHO-Net	7.3	24.7	8.0	24.9	8.7	26.4	6.6	23.8	7.6	23.9	9.0	24.3	5.0	21.1	9.9	26.2	8.8	24.8	11.5	32.4	9.1	25.3

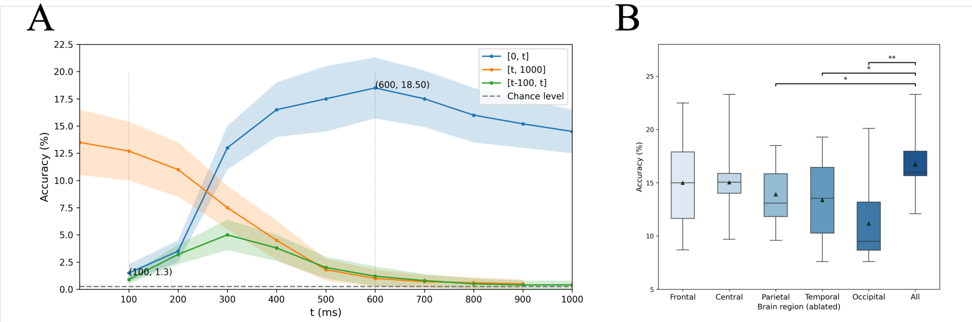

Figure 2: (A) Average decoding accuracy across subjects for different EEG time windows. Peak accuracy occurs at 600 ms. (B) Electrode-region ablation shows significant decreases (**$p < 0.01$) for occipital and temporal regions.

Mid Phase (300–600 ms): Temporal and frontal regions demonstrate marked activation increases, reflecting higher-order semantic and attentional processing (e.g., N400-like activity). This phase contributes most to decoding accuracy, as supported by the sharp accuracy peak around 600 ms in Figure 2 (A). The improvement is statistically significant across subjects ($p < 0.01$), suggesting that this phase captures semantic representations essential for object recognition.

Table 2: Subject-dependent 200-way zero-shot classification accuracy (%) on *THINGS-MEG*. best, second-best, third-best.

Method	Subject 1		Subject 2		Subject 3		Subject 4		Avg	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Intra-subject: train and test on one subject										
NICE	10.2	31.5	12.3	35.7	11.6	34.2	13.1	37.1	11.8	34.6
NICE-SA	9.6	28.4	17.8	45.5	10.4	36.2	12.8	27.5	12.7	34.4
NICE-GA	8.7	30.5	21.3	53.8	14.9	46.1	9.3	32.7	13.5	40.8
ECHO-Net	15.3	42.5	17.4	47.1	16.7	48.3	15.4	43.7	16.2	45.4

Cross-Subject Variability: Importantly, our spatially adaptive modules, EEG-GAT and Channel Attention, help mitigate these effects by learning personalized spatial priors and selectively amplifying informative electrodes, as further validated in the brain-region ablation analysis (Figure 2 (B). Overall, the combination of topographic visualization, temporal performance analysis, and spatial ablation offers strong empirical evidence for the interpretability and biological plausibility of ECHO-Net. These results validate the model's capacity to identify temporally and spatially discriminative neural signals, while remaining robust to inter-subject variability, a critical advantage for real-world BCI deployment.

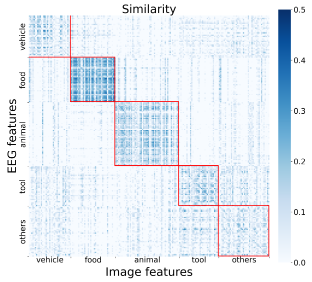

Figure 3: Representational Similarity Analysis (RSA) matrix between EEG and image features.

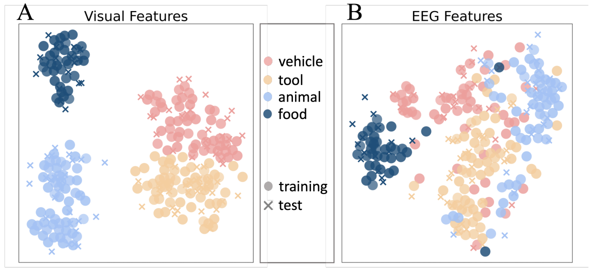

Figure 4: t-SNE visualization of four categories: animal, food, vehicle, and tool. (A) Visual feature distribution of the training and test sets. (B) EEG feature distribution of the training and test sets.

4.3 Semantic Similarity Analysis

To evaluate whether EEG embeddings encode high-level semantic information rather than low-level visual features, we conduct a Representational Similarity Analysis (RSA) [15]. The embedded elements derived from EEG are grouped into five semantic categories, *vehicles, food, animals, tools, and others*, and we compute pairwise cosine similarity between EEG and image features to assess their alignment with a structured semantic matrix.

Quantitative Findings. As shown in Figure 3, the similarity matrix reveals clear within-category clustering, indicating that EEG features are semantically organised. The RSA correlation between the EEG-based similarity matrix and the ground-truth semantic category matrix is $r = 0.334$ $(p < 0.001)$, confirming a statistically significant alignment. In contrast, a control experiment with randomly shuffled category labels yields a near-zero correlation of $r = 0.012$ $(p = 0.0145)$, suggesting that the observed structure is non-random and genuinely reflects semantic representations.

Qualitative Observations. We further analyze decoding outcomes by visualizing both the global feature distribution (Figure 4). The t-SNE plot illustrates that EEG embeddings cluster well within semantic categories, particularly for *animals*, *food*, and *vehicles*, indicating strong alignment between neural features and semantic representations.

5 Ablation Study

Kernel Diversity in Dynamic Convolution: Accuracy Gains and Limits. To further explore the role of dynamic convolution, we analyzed the effect of varying the number of convolutional kernels while keeping other components fixed. The results in Figure 5 reveal a clear trend: increasing the number of kernels initially improves accuracy but eventually leads to performance saturation and degradation. Using only a single kernel results in the lowest accuracy, as it limits the diversity of learned EEG patterns. As the kernel number increases from 1 to 64, Top-1 accuracy improves from 15.1% to 18.5%, demonstrating that a richer

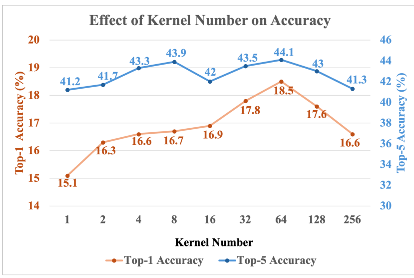

Figure 5: Effect of the number of dynamic convolution kernels on classification performance.

Table 3: Ablation study of the proposed EEG decoding framework, evaluating the contributions of Channel Attention (α_c), Filter Attention (α_f), and Graph Attention (α_g). Results show the incremental effect of each module on subject-dependent EEG classification (THINGS-EEG). Statistically significant differences compared to the baseline model, Static Conv (No attention), are marked: $^{*}(p < 0.05)$, $^{**}(p < 0.01)$, $^{***}(p < 0.001)$.

Model Variants	α_c	α_f	α_g	**Top-1 (%)**	**Top-5 (%)**
Static Conv (No attention)	-	-	-	14.2	38.4
Dynamic Conv Only	-	-	-	15.2 (+1.0)***	40.8 (+2.4)***
Dynamic (Channel Attention Only)	✓	-	-	15.7 (+1.5)**	41.5 (+3.1)***
Dynamic (Filter Attention Only)	-	✓	-	15.4 (+1.2)*	41.1 (+2.7)**
Dynamic (Graph Attention Only)	-	-	✓	15.8 (+1.6)***	42.9 (+4.5)***
Dynamic (Channel + Filter)	✓	✓	-	16.2 (+2.0)***	43.0 (+4.6)***
Dynamic (Channel + Graph)	✓	-	✓	16.9 (+2.7)***	43.7 (+5.3)***
Dynamic (Filter + Graph)	-	✓	✓	16.3 (+2.1)***	42.6 (+4.2)***
Dynamic (Channel + Filter + Graph)	✓	✓	✓	**18.5 (+4.3)*****	**44.1 (+5.7)*****

set of convolutional kernels enhances EEG feature extraction. However, beyond 64 kernels, performance begins to drop, due to overfitting and redundant representations. Notably, using 128 or 256 kernels results in lower accuracy than 64 kernels, confirming that excessive model complexity can harm generalization. These findings suggested that while kernel diversity improves EEG representation learning, excessive expansion can be counterproductive.

Ablation on Attention Modules: Disentangling Spatial and Temporal Contributions. We conducted ablation studies by selectively removing channel (α_c), filter (α_f), and graph (α_g) attention modules, both individually and in combination (Table 3). The results indicate that each component contributes meaningfully to overall performance. Notably, graph attention (α_g) yields the largest improvement, especially in Top-5 accuracy, highlighting the importance of modelling spatial dependencies among electrodes. Compared to the full model (18.5%), using only α_c or α_f results in a Top-1 accuracy of 15.7% and 15.4%, respectively, suggesting their complementary roles in spatial and temporal refinement.

6 Conclusion

In this paper, we proposed ECHO-Net, a dynamic convolutional framework for EEG-based object recognition, integrating adaptive spatio-temporal feature extraction, multimodal contrastive learning, and graph-based spatial modelling. Experiments on two large-scale benchmarks (THINGS-EEG and THINGS-MEG) demonstrate that our method achieves state-of-the-art performance in both subject-dependent and subject-independent settings. Ablation studies verify the importance of dynamic convolution in handling EEG's non-stationarity, while channel-filter and graph attention modules provide complementary gains via targeted spatio-temporal refinement. Temporal dynamics analysis identifies the 300–600 ms post-stimulus window as critical for decoding, consistent with neuroscience findings. In addition, semantic similarity results show that our EEG embeddings preserve high-level semantic structure, supporting effective multimodal alignment. Despite these advances, challenges remain regarding cross-dataset generalization and real-time efficiency. Future work should explore adaptive transfer strategies for real-world deployment, and investigate model compression and hardware-aware optimization for efficient inference. Overall, ECHO-Net offers a robust and scalable solution to neural decoding, with broad potential for neuroadaptive interfaces, clinical decision support, and human-centered AI systems. This unified design enables a trial-adaptive and semantically aligned decoding pipeline that, to our knowledge, has not been jointly explored in EEG decoding literature.

Acknowledgment

This work was supported by the Royal Society International Exchanges Scheme—Towards Collaborative Cloud–Edge Deep Learning Deployment (Grant No. IEC\NSFC\223523); the National Edge AI Hub for Real Data: Edge Intelligence for Cyber-disturbances and Data Quality (EP/Y028813/1); and the UK Medical Research Council (MRC) Innovation Fellowship (MR/S003916/2).

References

[1] Kai Keng Ang, Zheng Yang Chin, Haihong Zhang, and Cuntai Guan. Filter bank common spatial pattern (fbcsp) in brain-computer interface. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 2390–2397. IEEE, 2008.

[2] HS Anupama, NK Cauvery, and GM Lingaraju. Real-time eeg based object recognition system using brain computer interface. In *2014 International Conference on Contemporary Computing and Informatics (IC3I)*, pages 1046–1051. IEEE, 2014.

[3] Marie T Banich and Rebecca J Compton. *Cognitive neuroscience*. Cambridge University Press, 2018.

[4] Benjamin Blankertz, Steven Lemm, Matthias Treder, Stefan Haufe, and Klaus-Robert Müller. Single-trial analysis and classification of erp components—a tutorial. *NeuroImage*, 56(2):814–825, 2011.

[5] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic convolution: Attention over convolution kernels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11030–11039, 2020.

[6] Andac Demir, Toshiaki Koike-Akino, Ye Wang, and Deniz Erdoğmuş. Eeg-gat: graph attention networks for classification of electroencephalogram (eeg) signals. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 30–35. IEEE, 2022.

[7] Yi Ding, Chengxuan Tong, Shuailei Zhang, Muyun Jiang, Yong Li, Kevin JunLiang Lim, and Cuntai Guan. Emt: A novel transformer for generalized cross-subject eeg emotion recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

[8] Changde Du, Kaicheng Fu, Jinpeng Li, and Huiguang He. Decoding visual neural representations by multimodal learning of brain-visual-linguistic features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10760–10777, 2023.

[9] Yule Duan, Chuang Chen, Maixia Fu, Xiuwen Gong, Yingying Niu, and Fulin Luo. Gitanet: Group interactive threshold-based attention network for hyperspectral image classification. *IEEE Transactions on Multimedia*, 2025.

[10] Martin N Hebart, Oliver Contier, Lina Teichmann, Adam H Rockter, Charles Y Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I Baker. Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *Elife*, 12:e82580, 2023.

[11] Fo Hu, Kailun He, Mengyuan Qian, Xiaofeng Liu, Zukang Qiao, Lekai Zhang, and Junlong Xiong. Stafnet: an adaptive multi-feature learning network via spatiotemporal fusion for eeg-based emotion recognition. *Frontiers in Neuroscience*, 18:1519970, 2024.

[12] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[13] Yu Hu, Xiaobo Chen, Sheng Wang, Luyang Liu, Hengyang Shi, Lihong Fan, Jing Tian, and Jun Liang. Deformle cross-attention trnsformer for wekly aligned rgb–t pedestrin detection. *IEEE Transactions on Multimedia*, 2025.

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

[15] Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008.

[16] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.

[17] Chao Li, Aojun Zhou, and Anbang Yao. Omni-dimensional dynamic convolution. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=DmpCfq6Mg39.

[18] Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, and Quanying Liu. Visual decoding and reconstruction via EEG embeddings with guided diffusion. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=RxkcroC8qP.

[19] Jingcong Li, Weijian Pan, Haiyun Huang, Jiahui Pan, and Fei Wang. Stgate: Spatial-temporal graph attention network with a transformer encoder for eeg-based emotion recognition. *Frontiers in Human Neuroscience*, 17:1169949, 2023.

[20] Shuang Liang, Linzhe Li, Wei Zu, Wei Feng, and Wenlong Hang. Adaptive deep feature representation learning for cross-subject eeg decoding. *BMC bioinformatics*, 25(1):393, 2024.

[21] F. Lotte, L. Bougrain, A. Cichocki, et al. A review of classification algorithms for eeg-based brain–computer interfaces: a 10 year update. *Journal of neural engineering*, 15 (3):031005, 2018.

[22] Petra Ritter and Arno Villringer. simultaneous eeg–fmri. *Neuroscience & Biobehavioral Reviews*, 30(6):823–838, 2006.

[23] Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human brain mapping*, 38(11):5391–5420, 2017.

[24] Qingshan She, Xinsheng Shi, Feng Fang, Yuliang Ma, and Yingchun Zhang. Cross-subject eeg emotion recognition using multi-source domain manifold feature selection. *Computers in Biology and Medicine*, 159:106860, 2023.

[25] Amir Shirian, Subarna Tripathi, and Tanaya Guha. Dynamic emotion modeling with learnable graphs and graph inception network. *IEEE Transactions on Multimedia*, 24: 780–790, 2021.

[26] Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding natural images from EEG for object recognition. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=dhLIno8FmH.

[27] TJ Tsai, Andreas Stolcke, and Malcolm Slaney. A study of multimodal addressee detection in human-human-computer interaction. *IEEE Transactions on Multimedia*, 17(9):1550–1561, 2015.

[28] Peihao Wu, Wenqian Wang, Faliang Chang, Chunsheng Liu, and Bin Wang. Dss-net: Dynamic self-supervised network for video anomaly detection. *IEEE Transactions on Multimedia*, 26:2124–2136, 2023.

[29] Shuning Xue, Bu Jin, Jie Jiang, Longteng Guo, and Jing Liu. A hybrid local-global neural network for visual classification using raw eeg signals. *Scientific Reports*, 14(1): 27170, 2024.

[30] Brandon Yang, Gabriel Bender, Quoc V Le, and Jiquan Ngiam. Condconv: Conditionally parameterized convolutions for efficient inference. *Advances in Neural Information Processing Systems*, 32, 2019.

[31] Tianyu Zhang, Fan Wan, Haoran Duan, Kevin W Tong, Jingjing Deng, and Yang Long. Fmdconv: Fast multi-attention dynamic convolution via speed-accuracy trade-off. *Knowledge-Based Systems*, page 113393, 2025.