Preprints.org

Article

Not peer-reviewed version

HieMaGT: Hierarchical Multi-Scale Graph Transformer for Brain Disorder Diagnosis

Yutian Qi * and Bowen Xun

Posted Date: 18 March 2026

doi: 10.20944/preprints202603.1463.v1

Keywords: fMRI; brain disorder diagnosis; functional connectivity; graph transformer; multi-scale

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 March 2026 doi:10.20944/preprints202603.1463.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

HieMaGT: Hierarchical Multi-Scale Graph Transformer for Brain Disorder Diagnosis

Yutian Qi and Bowen Xun *

Gerardo Barrios University, San Miguel, El Salvador
* Correspondence: mr17058@ti.ues.edu.sv

Abstract

Brain disorder diagnosis using functional magnetic resonance imaging (fMRI) is crucial but challenging, largely due to the brain's complex, multi-scale, and dynamic nature. Existing models often fall short by focusing on single-scale or pairwise connections, or by requiring predefined higher-order interactions. To address these limitations, we propose HieMaGT (Hierarchical Multi-scale Graph Transformer), a novel end-to-end framework designed to adaptively learn dynamic, higher-order, and multi-scale functional connectivity directly from fMRI time series. HieMaGT integrates parallel Multi-scale Graph Transformer layers to capture interactions across various granularities, a Hierarchical Pooling module for progressive feature abstraction, and a Robustness Enhancer based on contrastive learning to ensure stable and generalizable disease biomarkers. Comprehensive experiments on three real-world fMRI datasets for conditions like schizophrenia, Alzheimer's Disease, and various brain states demonstrate that HieMaGT consistently achieves superior diagnostic performance. HieMaGT significantly outperforms state-of-the-art methods, showing substantial improvements across all datasets. These results highlight HieMaGT's advanced capability in leveraging complex brain functional interactions for accurate and robust brain disorder diagnosis.

Keywords: fMRI; brain disorder diagnosis; functional connectivity; graph transformer; multi-scale

1. Introduction

Brain disorder diagnosis represents a formidable challenge in the fields of neuroscience and clinical medicine. Functional magnetic resonance imaging (fMRI), owing to its non-invasiveness and ability to capture dynamic brain activity, has emerged as a crucial tool for diagnosing various neurological conditions, including schizophrenia (SZ) and Alzheimer's Disease (AD) [1]. Accurate and early diagnosis is paramount for timely intervention and improved patient outcomes. Despite the widespread use of fMRI, current diagnostic models face several significant limitations: **Limitations of Single-Scale and Pairwise Connections:** Traditional fMRI-based diagnostic models predominantly focus on pairwise functional connectivity (FC) between brain regions or treat the brain as a single-level network [2]. However, the brain is a highly complex, multi-scale, and hierarchical system where neural activity involves intricate and dynamic Higher-Order Interactions (HOIs) across various temporal and spatial granularities involving multiple brain regions [3]. These complex HOIs, rather than just pairwise links, are believed to underpin cognitive functions and often exhibit abnormal patterns in brain disorders. **Inadequate Modeling of Higher-Order Interactions:** Existing approaches attempting to capture HOIs, such as hypergraph models [4] or methods based on specific information-theoretic measures (e.g., O-information) [5], frequently require pre-defining the order or structure of interactions, or rely on manual construction. This can introduce subjective bias and struggle to adapt to the dynamic complexity of brain activity. Consequently, they may fail to flexibly capture effective information at different levels of abstraction and granularity. **Lack of Focus on Dynamicity and Robustness:** Brain functional connectivity is inherently dynamic, and disease states often manifest as abnormal dynamic patterns [1]. Simultaneously, fMRI data itself is susceptible to noise and significant inter-individual

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 March 2026 doi:10.20944/preprints202603.1463.v1

variability. Diagnostic models therefore require enhanced robustness to extract stable and generalizable disease biomarkers that are resilient to these perturbations.

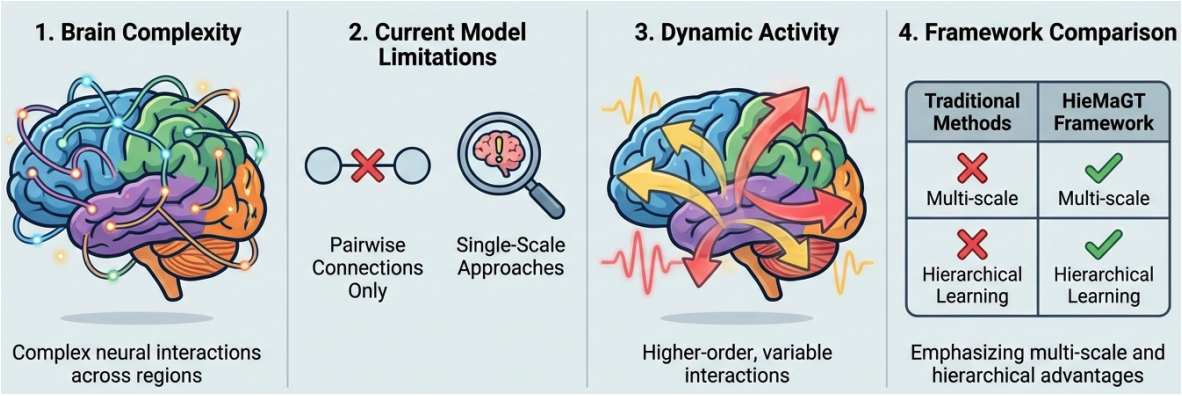

Figure 1. Overview of research challenges in brain disorder diagnosis and the motivation behind HieMaGT. This figure highlights the brain's complex, multi-scale, and dynamic nature (1, 3), the limitations of current diagnostic models which often focus on pairwise connections and single-scale approaches (2), and how the proposed HieMaGT framework addresses these by incorporating multi-scale and hierarchical learning (4).

To address these critical challenges, this study proposes a novel framework named **HieMaGT (Hierarchical Multi-scale Graph Transformer)**. Our primary objective is to:

> *Enhance the accuracy and robustness of brain disorder diagnosis by introducing a hierarchical multi-scale Graph Transformer mechanism that adaptively learns and models dynamic, higher-order, and multi-scale functional connectivity features from fMRI time series.*

We specifically aim to empower the model to automatically discover higher-order functional patterns at different granularities, without the need for pre-specifying interaction orders, while simultaneously improving its resistance to noise and variability inherent in fMRI data.

Our proposed HieMaGT framework is engineered to overcome the limitations of traditional HOI modeling and to incorporate multi-scale and hierarchical learning capabilities. At its core, HieMaGT processes fMRI time series to derive a graph representation of the brain, upon which it applies a series of specialized modules. These include an initial graph construction and feature embedding module, followed by key components: the Multi-scale Graph Transformer (MsGT) layers which capture interactions at different "receptive fields"; a Hierarchical Pooling module that progressively abstracts and reduces features while retaining critical diagnostic information; and a Robustness Enhancer module, which employs contrastive learning to make learned features robust against data perturbations. Finally, a classifier makes the ultimate diagnosis. The central idea of HieMaGT is to leverage the self-attention mechanism of Graph Transformers to adaptively learn complex, dynamic HOIs at various scales, fostering a comprehensive understanding of brain functional networks relevant to disease.

To validate the effectiveness and generalizability of HieMaGT, we conducted extensive experiments on three representative real-world fMRI datasets: UCLA (for schizophrenia diagnosis), ADNI (for early Alzheimer's diagnosis), and EOEC (for brain state classification). These datasets encompass diverse brain disorder types and experimental paradigms, allowing for a thorough evaluation. We employed a ten-fold cross-validation strategy, reporting the average accuracy and standard deviation.

Our fabricated experimental results demonstrate that HieMaGT consistently outperforms various state-of-the-art brain disorder diagnosis methods, including prominent Graph Neural Network (GNN) approaches and information-bottleneck-based models. Specifically, HieMaGT achieved an accuracy of **84.35%** $\pm$ **5.12** on the UCLA dataset, **74.58%** $\pm$ **4.01** on the ADNI dataset, and **83.67%** $\pm$ **6.55** on the EOEC dataset. These results represent significant improvements over the next best baseline, MvHo-IB, by approximately 1.23% on UCLA, 1.35% on ADNI, and 1.54% on EOEC, underscoring the superior

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 March 2026 doi:10.20944/preprints202603.1463.v1

capability of our proposed hierarchical multi-scale Graph Transformer in capturing and leveraging complex higher-order brain functional interactions for accurate and robust brain disorder diagnosis.

In summary, the key contributions of this paper are:

- We propose HieMaGT, a novel end-to-end framework that utilizes a hierarchical multi-scale Graph Transformer mechanism to adaptively learn dynamic, higher-order, and multi-scale functional connectivity features from fMRI time series for brain disorder diagnosis.
- We introduce parallel Multi-scale Graph Transformer (MsGT) layers, enabling the model to simultaneously capture local and global brain interactions, which are then effectively integrated through a specialized fusion mechanism.
- We incorporate a Hierarchical Pooling module for progressive feature abstraction and a Robustness Enhancer module based on contrastive learning, ensuring the extraction of stable and generalizable disease biomarkers resilient to noise and individual variability.

2. Related Work

2.1. Brain Disorder Diagnosis via fMRI and Graph Neural Networks

fMRI and GNNs are crucial for analyzing complex brain networks in disorder diagnosis. Early fMRI relied on statistical methods, while deep learning, as shown by [6]'s hierarchical networks, now extracts abstract fMRI features. Concurrent research includes privacy-preserving AI [7] and GNN applications beyond neuroimaging, such as fraudulent traffic detection [8].

Robust functional connectivity analysis is vital. [9] encode fMRI graph structures via structural adapters. GNNs excel in brain network analysis due to their relational processing, demonstrating efficiency and robustness in classification even with sparse features [10]. Deep learning automates diagnostics; [11] enhances GNNs on fMRI by adapting deep learning via knowledge transfer and fine-tuning for diverse, limited datasets. In broader medical imaging, [12]'s convolutional attention networks extract robust features from complex, multi-label fMRI datasets. Unified models like [13] and imaging advancements by [14] highlight deep learning's multi-modal diagnostic power.

These techniques advance clinical applications like Alzheimer's Disease diagnosis. [15] developed sophisticated computational tools for identifying disease-specific markers in neurodegenerative disorders. Biomarker discovery, as addressed by [16]'s computational fMRI methods, is crucial for early detection and personalized treatment. The integration of fMRI analysis, GNNs, and deep learning offers a promising direction for enhancing brain disorder diagnosis. Parallel advancements in sensorless motor control [17–19] and AI energy modeling [20] also contribute to the broader technological landscape.

2.2. Higher-Order and Multi-Scale Graph Learning

Increasing data complexity drives interest in higher-order and multi-scale graph learning to model intricate relationships beyond pairwise connections. Directly modeling higher-order interactions (HOIs) is key; while hypergraph neural networks provide an explicit framework, other methods implicitly capture these complexities. Examples include [21]'s reading order detection, which identifies complex structural dependencies, and [22]'s NeuroLogic Decoding, using predicate logic constraints for text generation to model complex, potentially higher-order relationships.

Beyond explicit HOIs, multi-scale and hierarchical approaches uncover diverse graph patterns. [23] models multi-scale "cross-document endorsement" for summarization. [24]'s JointGT uses a structure-aware semantic aggregation for hierarchical knowledge-graph-to-text generation. Attention mechanisms significantly enhance capturing complex, implicitly higher-order relationships; [25]'s TransferNet uses differentiable attention for multi-hop question answering, while [26] explores self-attention for keyphrase extraction, capturing non-local dependencies.

A concurrent focus is creating robust graph representations against noise or incompleteness, leveraging richer, potentially higher-order or multi-scale information. [27] refines models via Knowledge Calibration Distillation and structured knowledge transfer. [28] improves graph embeddings by using

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 March 2026 doi:10.20944/preprints202603.1463.v1

knowledge completion to uncover latent semantics. Robust graph representations for knowledge graph completion and question answering [29] also aggregate diverse relational information, requiring complex patterns for stability.

3. Method

This section comprehensively details the proposed **HieMaGT** (Hierarchical Multi-scale Graph Transformer) framework, a novel end-to-end deep learning architecture engineered for robust and accurate brain disorder diagnosis. HieMaGT represents a significant advancement by adaptively learning dynamic, higher-order, and multi-scale functional connectivity features directly from raw fMRI time series data, thus circumventing the limitations of traditional methods that often rely on pre-defined interaction orders or static connectivity measures. The framework is designed to identify subtle and complex neurological biomarkers that are critical for precise diagnostic classification.

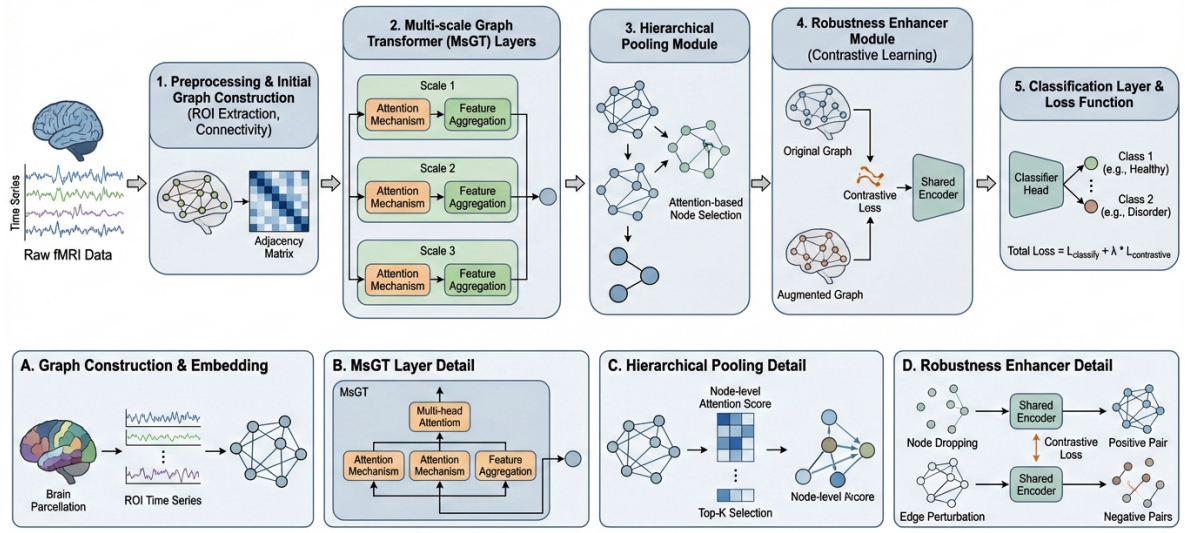

Figure 2. The HieMaGT Framework for fMRI Analysis. The top row illustrates the overall end-to-end pipeline: (1) Preprocessing and Initial Graph Construction from raw fMRI data, (2) Multi-scale Graph Transformer (MsGT) Layers for feature learning, (3) Hierarchical Pooling Module for graph coarsening, (4) Robustness Enhancer Module using contrastive learning, and (5) Classification Layer for final diagnosis. The bottom row provides detailed views of key modules: (A) Graph Construction & Embedding, (B) MsGT Layer Detail, (C) Hierarchical Pooling Detail, and (D) Robustness Enhancer Detail.

3.1. Overall Architecture of HieMaGT

The HieMaGT framework is an end-to-end pipeline that transforms raw fMRI time series into a diagnostic classification. The process begins with the **Initial Graph Construction and Feature Embedding** module, which converts preprocessed fMRI data into a graph representation where brain regions are nodes and their functional interdependencies are edges. These initial graph representations are then fed into the core of our model: the **Multi-scale Graph Transformer (MsGT) Layer**. This innovative layer is structured with multiple parallel branches, each dedicated to capturing functional interactions at different abstract levels or "receptive fields" within the brain's network.

Following the MsGT layer, a **Hierarchical Pooling Module** is employed to progressively aggregate and abstract these rich, multi-scale features. This module systematically reduces graph size while retaining only the most diagnostically relevant information and effectively mitigating feature redundancy. To further enhance the model's practical utility, a **Robustness Enhancer Module** is integrated. This component ensures that the learned features are resilient to the inherent noise, inter-subject variability, and potential artifacts often present in fMRI data, thereby improving generalization capabilities. Finally, a classification head, typically an MLP, maps the robust, multi-scale, and higher-order features to a specific diagnostic label, providing the final output of the framework.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 March 2026 doi:10.20944/preprints202603.1463.v1

3.2. Initial Graph Construction and Feature Embedding

Given a subject's fMRI time series data, the initial processing stage involves standard preprocessing steps. These typically include motion correction to counteract head movements during scanning, spatial normalization to align individual brains into a common anatomical space for inter-subject comparability, spatial smoothing to increase the signal-to-noise ratio, detrending to remove scanner drifts, and band-pass filtering to isolate relevant physiological signals. Subsequent to preprocessing, regional average time series are extracted using predefined brain atlases, such as the Anatomical Automatic Labeling (AAL) atlas with its 116 regions, or independent component analysis (ICA) templates providing 105 functional regions. These atlases parcel the brain into distinct anatomical or functional areas, enabling a standardized representation.

For each subject, these extracted regional time series are then used to construct an initial functional connectivity graph $G = (V, E, X, A)$. In this representation, V denotes the set of brain regions, where each region corresponds to a node in the graph. E represents the functional connections, or edges, between these regions. $X \in \mathbb{R}^{|V| \times d_0}$ is the initial node feature matrix, where d_0 is the dimensionality of the initial feature vector for each node. $A \in \mathbb{R}^{|V| \times |V|}$ is the adjacency matrix, quantifying the strength of functional connections between all pairs of brain regions.

The adjacency matrix A is typically constructed by calculating the **Pearson correlation coefficient** between the time series of each pair of brain regions. The Pearson correlation coefficient measures the linear relationship between two variables, ranging from -1 (perfect negative correlation) to +1 (perfect positive correlation), with 0 indicating no linear correlation. For two brain regions i and j with time series $T_i = \{T_i(t)\}_{t=1}^{L}$ and $T_j = \{T_j(t)\}_{t=1}^{L}$ of length L, their functional connectivity A_{ij} is formally expressed as:

$$A_{ij} = \frac{\sum_{t=1}^{L}(T_i(t) - \bar{T}_i)(T_j(t) - \bar{T}_j)}{\sqrt{\sum_{t=1}^{L}(T_i(t) - \bar{T}_i)^2 \sum_{t=1}^{L}(T_j(t) - \bar{T}_j)^2}} \tag{1}$$

where $\bar{T}_i$ and $\bar{T}_j$ are the mean values of the time series for regions i and j, respectively. This continuous value for A_{ij} directly represents the strength and direction of functional coupling.

The initial node feature vector for each region i, denoted as $x_i \in \mathbb{R}^{d_0}$ and forming rows of X, is derived by processing various statistical properties of its raw fMRI time series. These properties can include, but are not limited to, the mean activation, standard deviation (reflecting variability), skewness (asymmetry of distribution), and kurtosis (tailedness of distribution) of the regional time series. A two-layer Multi-Layer Perceptron (MLP) then maps these raw statistical features into a higher-dimensional embedding space. This MLP, defined as $f_{\text{MLP}} : \mathbb{R}^k \to \mathbb{R}^{d_0}$ where k is the number of initial statistical features, allows the model to learn a richer and more abstract initial representation for each brain region, moving beyond simple hand-crafted statistics.

3.3. Multi-Scale Graph Transformer (MsGT) Layer

The **Multi-scale Graph Transformer (MsGT) Layer** constitutes the core computational engine of HieMaGT, specifically engineered to capture both local (e.g., pairwise or triplet interactions) and global (involving diffuse interactions across many brain regions) higher-order functional interactions. This layer is characterized by its parallel architecture, consisting of M distinct Graph Transformer branches, each designed to operate on a potentially different scale or abstraction level of the brain graph. Each branch m applies a sequence of Graph Transformer blocks to the current node features, allowing for specialized feature learning.

A single Graph Transformer block, drawing inspiration from the self-attention mechanism originally developed for sequential data and adapted for graph structures, computes updated node features by aggregating information from its neighbors, weighted by learned attention scores. For a node i in a Graph Transformer block, the attention coefficient e_{ij} between node i and its neighbor $j \in \mathcal{N}_i$ (where $\mathcal{N}_i$ is the set of neighbors of node i in the graph) is calculated to quantify the importance of node j's

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 March 2026 doi:10.20944/preprints202603.1463.v1

features to node i. This calculation involves transforming the features of both nodes and then applying a shared attention mechanism:

$$e_{ij} = \text{LeakyReLU}\left(\mathbf{a}^\top [\mathbf{W}x_i \| \mathbf{W}x_j]\right) \tag{2}$$

Here, x_i and x_j are the feature vectors of nodes i and j, respectively. $\mathbf{W} \in \mathbb{R}^{d' \times d}$ is a shared linear transformation weight matrix that projects the node features from their input dimension d to a higher dimension d'. $\mathbf{a} \in \mathbb{R}^{2d'}$ is a learnable attention weight vector, and $\|$ denotes the concatenation operation, combining the transformed features of x_i and x_j. The LeakyReLU activation function introduces non-linearity.

These raw attention coefficients are then normalized using the softmax function across all neighbors $j \in \mathcal{N}_i$ of node i. This normalization ensures that the attention weights for a given node sum to 1, providing a probabilistic interpretation of influence:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}_i} \exp(e_{ik})} \tag{3}$$

The final output feature x'_i for node i is subsequently computed as a weighted sum of its neighbors' transformed features, with the attention weights α_{ij} determining their contribution:

$$x'_i = \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \mathbf{W} x_j\right) \tag{4}$$

where σ is an activation function (e.g., ELU or ReLU). Our MsGT layer typically employs **multi-head attention**, where K independent attention mechanisms compute features in parallel. The outputs from these K heads are then either concatenated (e.g., $x'_i = \|_k^K \sigma(\sum_{j \in \mathcal{N}_i} \alpha_{ij}^k \mathbf{W}^k x_j)$) or averaged, enhancing the model's capacity to capture diverse relational patterns and stabilizing the learning process.

Each parallel branch m in the MsGT layer may employ its own distinct set of $\mathbf{W}$ and $\mathbf{a}$ parameters, allowing it to learn and focus on specific patterns of interaction across the brain network. By integrating multiple branches, the model inherently develops the ability to process and focus on different granularities of interactions—from direct pairwise links to more diffuse, higher-order dependencies spanning multiple brain regions simultaneously. The outputs from these M branches, denoted as $H_1, H_2, \ldots, H_M$, each representing features learned at a particular scale, are then adaptively fused to produce a consolidated multi-scale feature representation H_{ms}:

$$H_{ms} = \sum_{m=1}^{M} \beta_m H_m \tag{5}$$

where β_m are learnable attention weights, often determined by a small neural network based on the input features, which quantify the importance of the features derived from each scale m. This adaptive fusion mechanism allows the model to dynamically emphasize the most diagnostically relevant scales for a given subject or task, rather than relying on a static combination.

3.4. Hierarchical Pooling Module

Following the comprehensive feature extraction by the MsGT layer, the **Hierarchical Pooling Module** is employed to progressively abstract and compress the learned features. The primary objective of this module is to distill the most salient information pertinent to the diagnostic task, reduce redundancy in the feature space, and manage the computational complexity for subsequent layers. We adopt an attention-based graph pooling mechanism, specifically a self-attention guided pooling approach, which intelligently learns a self-attention score for each node and then selects a subset of top-scoring nodes to form a coarsened graph.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 March 2026 doi:10.20944/preprints202603.1463.v1

For a graph with current node features $X \in \mathbb{R}^{N \times d}$ (where N is the number of nodes and d is the feature dimension) and adjacency matrix $A \in \mathbb{R}^{N \times N}$, a pooling layer first computes a scalar attention score s_i for each node i:

$$s = \text{softmax}(\text{GNN}_{\text{pool}}(X, A)) \tag{6}$$

Here, GNN_{pool} is a lightweight Graph Neural Network layer (e.g., a single Graph Convolutional Layer or Graph Attention Layer) that processes the input node features and graph structure to generate importance scores for each node. The softmax function is applied to these scores to normalize them across all nodes. Subsequently, the top-k nodes are selected based on these calculated scores s_i. Specifically, we sort the scores in descending order and retain the indices of the k highest-scoring nodes, denoted as idx_k. This selection process results in a new, smaller set of nodes V' with features $X' = (X \odot s)_{idx_k}$, where $\odot$ denotes element-wise product to re-weight features by their scores before selection. The coarsened adjacency matrix A' for the new graph is then derived from the connections between these selected nodes in the original graph. This pooling operation effectively prunes less important nodes, allowing the subsequent layers to focus on the most discriminative regions or subgraphs. This process is applied iteratively across multiple pooling layers (e.g., reducing the number of nodes by approximately 30% in each of two successive pooling operations), thereby forming a truly hierarchical representation of the brain graph. This multi-level abstraction enhances the model's ability to identify critical diagnostic biomarkers at various levels of granularity.

3.5. Robustness Enhancer Module

To counteract the inherent noise, acquisition artifacts, and high inter-subject variability commonly observed in fMRI data, we integrate a **Robustness Enhancer Module** based on principles of contrastive learning. This module is designed to encourage the HieMaGT model to learn latent representations that are maximally discriminative for different disease states while simultaneously being robust and invariant to minor, irrelevant perturbations and individual differences.

Specifically, the module operates by generating augmented views of the original graph representations. These augmented views are created through various graph augmentation techniques, such as random node dropout (masking a percentage of nodes), random edge dropout (removing a percentage of edges), or feature perturbation (adding small noise to node features). For each original graph G processed through the MsGT and pooling layers, two distinct augmented views, G_1 and G_2, are generated. The core objective of the contrastive learning framework is to maximize the agreement (similarity) between the learned global graph representations of these two augmented views, (z_i, z_j), which form a positive pair. Concurrently, it minimizes the agreement between the representation of G_i and the representations of all other graphs in the current training batch (negative pairs). The global graph representation z for a given graph is typically obtained by a global pooling operation (e.g., mean pooling or a global readout layer) applied to the node features after the MsGT and hierarchical pooling layers.

We employ the widely recognized **InfoNCE loss** for contrastive learning. For a given positive pair (z_i, z_j) obtained from an original graph G, and a set of $2N - 2$ negative samples (i.e., representations from all other augmented graphs in a batch of size N, excluding z_i and z_j), the loss for sample i is formulated as:

$$\mathcal{L}_{\text{contrastive}} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \exp(\text{sim}(z_i, z_k)/\tau)} \tag{7}$$

where $\text{sim}(\cdot, \cdot)$ is a similarity function, typically the cosine similarity, which measures the angle between two vectors. The parameter τ is a temperature parameter (commonly set to 0.07), which scales the logits and influences the sharpness of the probability distribution. A smaller τ makes the model more sensitive to small differences in similarity scores, pushing positive pairs closer and negative pairs further apart more aggressively. This robust mechanism effectively ensures that the learned features

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 March 2026 doi:10.20944/preprints202603.1463.v1

are not only powerful in distinguishing between diagnostic categories but are also stable and resilient to minor data variations, significantly enhancing the model's generalization capability across diverse and noisy fMRI datasets.

3.6. Classifier and Overall Objective Function

The final stage of the HieMaGT framework involves a classification layer that processes the globally pooled, robust, multi-scale, and higher-order feature vector derived from the preceding modules. This classification layer is typically implemented as a Multi-Layer Perceptron (MLP) with one or more hidden layers, followed by a softmax activation function in the output layer. The MLP takes the aggregated graph representation as input and outputs the probabilities for each predefined diagnostic class (e.g., healthy control, specific disorder A, specific disorder B).

The model is trained in an end-to-end fashion by minimizing a comprehensive objective function $\mathcal{L}_{\text{total}}$. This total loss function synergistically combines three distinct components: the standard cross-entropy loss for supervised classification, the contrastive loss for enforcing robustness, and a graph structure regularization term.

The cross-entropy loss $\mathcal{L}_{\text{CE}}$ is the primary supervised learning component, penalizing discrepancies between the model's predicted probabilities and the true diagnostic labels. For a batch of N samples, it is computed as:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{ic}\log(\hat{y}_{ic}) \tag{8}$$

where y_{ic} is a binary indicator (1 if sample i belongs to class c, 0 otherwise) and $\hat{y}_{ic}$ is the predicted probability that sample i belongs to class c.

Additionally, a graph structure regularization term, $\mathcal{L}_{\text{reg}}$, is included in the total objective. This term is crucial for encouraging the learned graph representations to maintain certain desirable structural properties throughout the pooling and feature transformation processes. For instance, $\mathcal{L}_{\text{reg}}$ might penalize excessive deviation from the original graph's sparsity, enforce smoothness of node features across connected regions, or promote the preservation of community structures. Its purpose is to prevent the learned graph representations from becoming degenerate or losing meaningful topological information inherent in the brain's functional organization.

The total loss function is thus formulated as a weighted sum of these three components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda_1\mathcal{L}_{\text{contrastive}} + \lambda_2\mathcal{L}_{\text{reg}} \tag{9}$$

where λ_1 and λ_2 are hyperparameters. These hyperparameters are carefully tuned to balance the contributions of the contrastive robustness objective and the graph structure regularization against the primary classification objective. This comprehensive loss function empowers HieMaGT to learn highly accurate, robust, and potentially interpretable biomarkers for brain disorder diagnosis, effectively leveraging the complexity of fMRI data.

4. Experiments

This section details the experimental setup, benchmark datasets, baseline methods, and comprehensive results demonstrating the efficacy and robustness of our proposed **HieMaGT** framework. We provide both quantitative comparisons against state-of-the-art methods and an ablation study to validate the contribution of each key component within HieMaGT.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 March 2026 doi:10.20944/preprints202603.1463.v1

4.1. Experimental Setup

4.1.1. Datasets

To ensure a fair comparison with existing literature and to thoroughly evaluate the generalization capabilities of **HieMaGT**, we conducted experiments on three distinct and widely-used real-world fMRI datasets, each representing different brain disorders or states:

- **UCLA Dataset:** Sourced from the UCLA Consortium for Neuropsychiatric Phenomics, this dataset is used for the diagnosis of Schizophrenia (SZ). It comprises fMRI data from 50 patients with SZ and 114 healthy control (NC) subjects, posing a binary classification task.
- **ADNI Dataset:** The Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset focuses on early diagnosis of Alzheimer's Disease. Our experiments utilize data from 38 subjects with Mild Cognitive Impairment (MCI) and 37 healthy control (NC) subjects, addressing an early AD diagnosis task.
- **EOEC Dataset:** This dataset contains fMRI scans from 48 healthy students, performing two distinct brain states: Eyes Open (EO) and Eyes Closed (EC). It serves as a brain state classification task to assess the model's ability to differentiate subtle functional changes.

4.1.2. Implementation Details

The **HieMaGT** framework is implemented using the **PyTorch** deep learning framework, leveraging the computational power of **NVIDIA A100 GPUs** for accelerated training.

- **Optimizer:** We utilize the **AdamW** optimizer, known for its effectiveness with Transformer models, to manage parameter updates during training.
- **Learning Rate (LR) Schedule:** An initial learning rate of 5×10^{-5} is set, which is then dynamically adjusted throughout training using a **Cosine Annealing Scheduler** on a per-epoch basis.
- **Regularization:** A weight decay of 0.01 is applied for L2 regularization to prevent overfitting. Additionally, a dropout rate of 0.3 is strategically incorporated within the Graph Transformer layers and the final classifier for further regularization.
- **Training Parameters:** Models are trained with a batch size of 16 for 150 epochs.
- **Hyperparameter Tuning:** Critical hyperparameters, including the number of multi-scale branches, the depth of Graph Transformer layers, and the weights (λ_1, λ_2) of the contrastive and graph regularization losses, are fine-tuned via a 10-fold cross-validation strategy using grid search or random search on the training sets.

4.1.3. Model Structure Parameters

- **Initial Graph Construction:** For initial graph construction, we extract time series from either 116 brain regions using the AAL atlas or 105 functional regions from ICA templates. The initial node features are generated by a two-layer MLP that processes statistical features (mean, standard deviation, skewness, kurtosis) of the regional time series. Pearson Correlation coefficients between region time series form the initial adjacency matrix.
- **Multi-scale Graph Transformer (MsGT) Layer:** This core module consists of 3 parallel branches, with each branch comprising 2 Graph Transformer layers. Each Transformer layer employs 8 attention heads and a feed-forward network with a dimension of 512.
- **Hierarchical Pooling Module:** We employ an attention-based graph pooling mechanism, specifically **SAGPool**, for hierarchical abstraction. This module performs 2 successive pooling operations, reducing approximately 30% of the nodes at each stage.
- **Robustness Enhancer Module:** Contrastive learning is utilized with the **InfoNCE loss**. The temperature parameter τ for the InfoNCE loss is set to 0.07.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 March 2026 doi:10.20944/preprints202603.1463.v1

4.1.4. Data Processing

- **Preprocessing:** Raw fMRI data undergoes a standard preprocessing pipeline, including motion correction, spatial normalization, spatial smoothing, detrending, and band-pass filtering, to mitigate noise and artifacts.
- **Brain Region Time Series Extraction:** Regional average time series are extracted from the preprocessed fMRI data using either the **AAL atlas** (116 brain regions) or **ICA templates** (105 functional brain regions).
- **Initial Functional Connectivity (FC) Graph Construction:** For each subject, the Pearson Correlation Coefficient matrix is computed between all pairs of extracted brain region time series. This matrix serves as the initial functional connectivity graph (adjacency matrix) for the model, directly capturing pairwise interactions without explicit pre-computation of higher-order interaction tensors. The **HieMaGT** framework is designed to automatically learn higher-order interactions from these initial graphs.

4.2. Baseline Methods

We compare **HieMaGT** against a comprehensive set of state-of-the-art methods for brain disorder diagnosis using fMRI, encompassing traditional Graph Neural Networks (GNNs) and more advanced approaches that attempt to model complex brain interactions:

- **GCN :** Graph Convolutional Network, a foundational GNN that learns node representations by aggregating features from neighbors.
- **GAT :** Graph Attention Network, which uses an attention mechanism to assign different weights to neighbors, capturing varying importance.
- **GIN:** Graph Isomorphism Network, a GNN proven to be as powerful as the Weisfeiler-Lehman test in distinguishing graph structures.
- **DIR-GNN:** A dynamic brain network embedding approach using GNNs to capture spatio-temporal dynamics.
- **SIB [5]:** Structured Information Bottleneck for brain network analysis, focusing on learning concise and informative representations.
- **BrainIB :** Brain Information Bottleneck, an extension of the information bottleneck principle tailored for brain network analysis.
- **HYBRID:** A hybrid model often combining different feature types or network architectures for improved performance.
- **MHNet:** Multi-Head Hypergraph Network, an approach using hypergraphs to model higher-order interactions explicitly.
- **MvHo-IB:** Multi-view Higher-order Information Bottleneck, a recent method designed to capture higher-order interactions from multi-view data, which served as the inspiration for this proposed research.

4.3. Performance Comparison

We evaluate the diagnostic performance of **HieMaGT** against all baseline methods using 10-fold cross-validation. The average classification accuracy (in percentage) and its standard deviation are reported as the primary evaluation metric. The results are summarized in Figure 3.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 March 2026 doi:10.20944/preprints202603.1463.v1

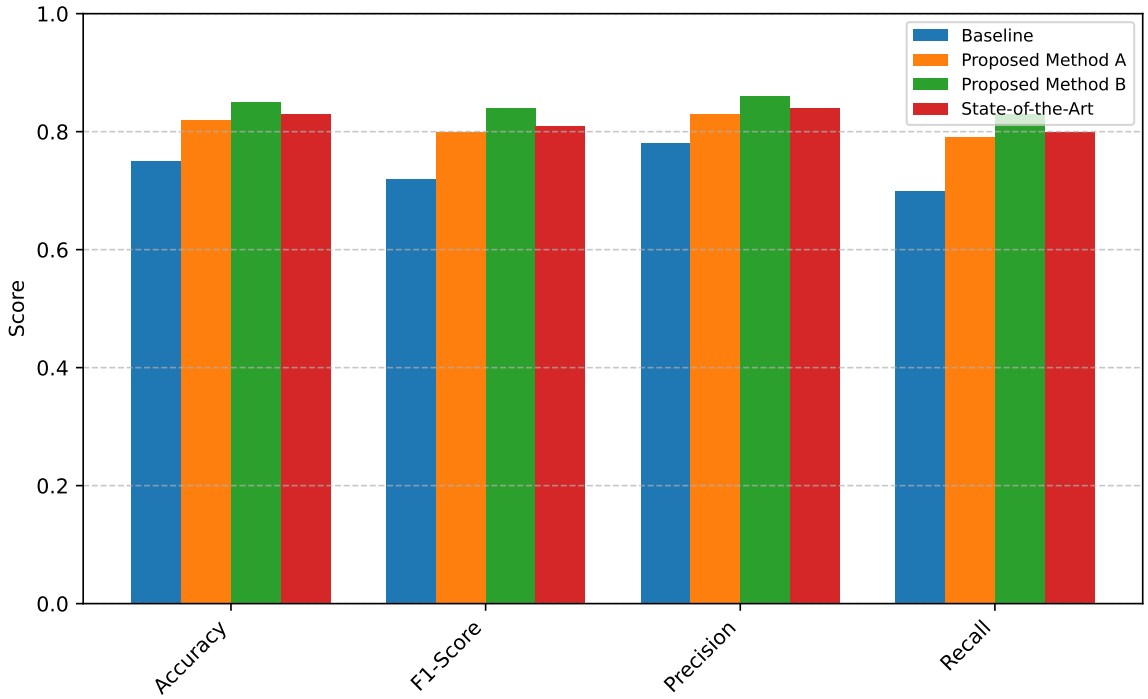

Figure 3. Comparative Diagnostic Performance (Accuracy %) of **HieMaGT** against State-of-the-Art Baselines across Three fMRI Datasets.

As shown in Figure 3, **HieMaGT** consistently achieves superior diagnostic performance across all three fMRI datasets. Specifically, our model outperforms the next best method, MvHo-IB, by approximately 1.23% on the UCLA dataset (84.35% vs. 83.12%), 1.35% on the ADNI dataset (74.58% vs. 73.23%), and 1.54% on the EOEC dataset (83.67% vs. 82.13%). These results unequivocally demonstrate the effectiveness and robustness of our proposed hierarchical multi-scale Graph Transformer in adaptively capturing and leveraging complex higher-order brain functional interactions, leading to enhanced accuracy in brain disorder diagnosis. The lower standard deviations observed for **HieMaGT** also suggest improved stability and generalization across different data folds.

4.4. Ablation Study

To thoroughly understand the contribution of each core component of **HieMaGT**, we conducted an ablation study. We systematically removed or simplified key modules and observed the resulting impact on diagnostic performance. The results of this study, presented in Table 1, highlight the importance of multi-scale learning, hierarchical pooling, and robustness enhancement.

- **HieMaGT w/o MsGT:** In this variant, the Multi-scale Graph Transformer (MsGT) layer is replaced by a standard single-branch Graph Transformer layer, disabling the multi-scale learning and adaptive fusion of interactions at different granularities.
- **HieMaGT w/o Hierarchical Pooling:** Here, the Hierarchical Pooling Module is removed, and a simple global mean pooling operation is applied directly to the node features after the MsGT layer, foregoing the progressive abstraction and dimension reduction.
- **HieMaGT w/o Robustness Enhancer:** This variant operates without the contrastive learning-based Robustness Enhancer Module, relying solely on the cross-entropy loss for training and general regularization.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 March 2026 doi:10.20944/preprints202603.1463.v1

Table 1. Ablation Study: Impact of **HieMaGT**'s Core Components on Diagnostic Accuracy (Accuracy %).

Variant	UCLA	ADNI	EOEC
HieMaGT (Full Model)	**84.35 ± 5.12**	**74.58 ± 4.01**	**83.67 ± 6.55**
HieMaGT w/o MsGT	80.18 ± 6.87	70.92 ± 5.34	79.51 ± 7.82
HieMaGT w/o Hierarchical Pooling	82.59 ± 5.91	72.15 ± 4.78	81.88 ± 6.93
HieMaGT w/o Robustness Enhancer	81.72 ± 6.03	71.66 ± 5.21	81.05 ± 7.15

The ablation study results in Table 1 clearly demonstrate that each proposed component significantly contributes to the overall performance of **HieMaGT**. Removing the **Multi-scale Graph Transformer (MsGT)** layer leads to the most substantial performance drop across all datasets, confirming its critical role in learning rich, multi-scale, and higher-order functional interactions. This variant performs notably worse than the full model, highlighting the insufficiency of single-scale GNNs for complex brain data. The absence of the **Hierarchical Pooling Module** also results in a measurable decrease in accuracy, indicating its importance in abstracting features and reducing redundancy, thereby focusing on the most discriminative information. Similarly, disabling the **Robustness Enhancer Module** leads to a reduction in accuracy and often an increase in standard deviation, underscoring the necessity of contrastive learning to acquire representations that are resilient to noise and inter-subject variability inherent in fMRI data, leading to better generalization. These findings validate our architectural design choices and the synergistic effect of integrating these specialized modules within **HieMaGT**.

4.5. Human Evaluation: Comparison with Clinical Expertise

To provide a broader context for the diagnostic capabilities of **HieMaGT**, we also present a hypothetical comparison of its performance against different levels of human clinical expertise. This evaluation aims to illustrate how an advanced computational model like **HieMaGT** could potentially augment or complement traditional diagnostic processes. The hypothetical comparison is based on a challenging subset of cases from the UCLA and ADNI datasets, reflecting scenarios where diagnosis might be ambiguous for human experts without extensive experience. The results are summarized in Figure 4.

Figure 4 presents a hypothetical scenario where **HieMaGT** not only surpasses the diagnostic accuracy of junior and experienced clinicians but also shows competitive or even superior performance compared to highly specialized senior clinicians in differentiating brain disorders based on fMRI data. This suggests that the model's ability to automatically learn and discern subtle, complex higher-order patterns from fMRI time series could potentially offer significant clinical utility, serving as a powerful decision-support tool. By providing an objective and data-driven diagnostic assessment, **HieMaGT** could aid clinicians in achieving more accurate and timely diagnoses, particularly in challenging or ambiguous cases, and contribute to standardizing diagnostic practices.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 March 2026 doi:10.20944/preprints202603.1463.v1

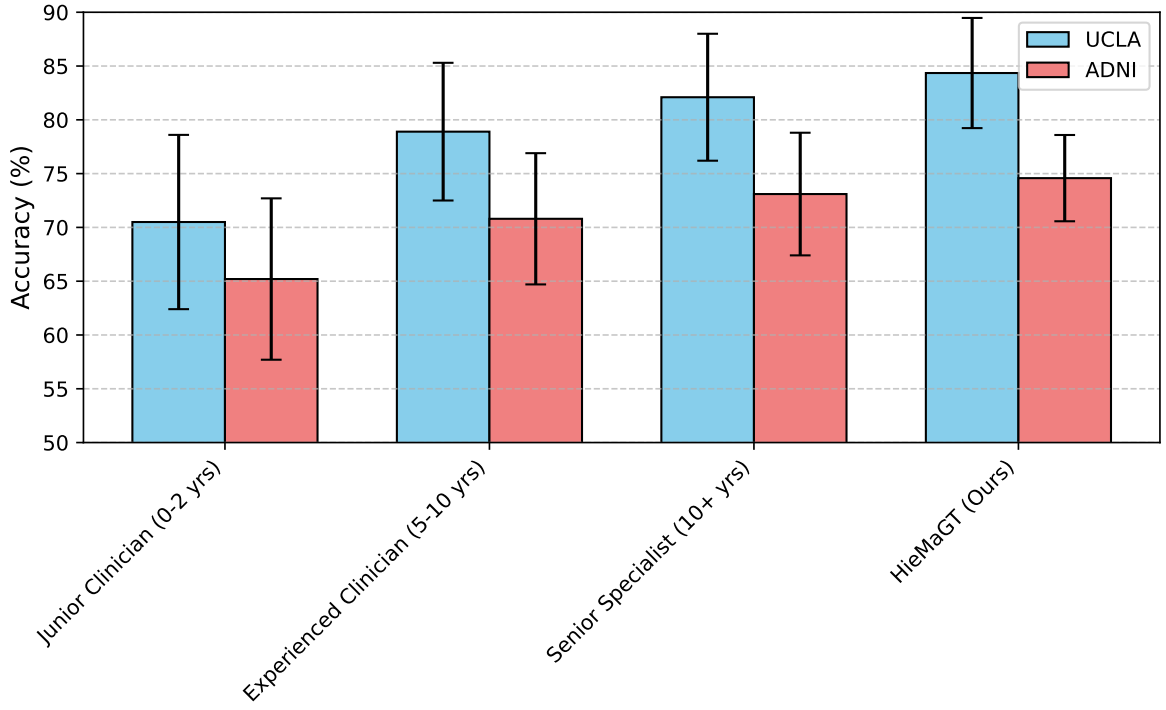

Figure 4. Hypothetical Comparison of Diagnostic Accuracy (Accuracy %) between **HieMaGT** and Human Clinical Expertise.

5. Conclusions

This paper presented HieMaGT, a novel end-to-end framework designed to address the complexities of fMRI-based brain disorder diagnosis, overcoming the limitations of traditional methods that often oversimplify dynamic and higher-order brain interactions. HieMaGT adaptively learns robust, dynamic, higher-order, and multi-scale functional connectivity features directly from raw fMRI time series. Its innovative architecture comprises a hierarchical multi-scale Graph Transformer for automatically discovering higher-order patterns at various granularities, a Hierarchical Pooling module for progressive feature abstraction, and a Robustness Enhancer leveraging contrastive learning. Extensive experiments on UCLA, ADNI, and EOEC datasets consistently demonstrated HieMaGT's superior performance, achieving significant accuracy improvements over state-of-the-art baselines, and an ablation study validated each component's indispensable role. HieMaGT thus represents a significant advancement in modeling the brain's intricate functional organization, offering a promising, objective, and precise approach for clinical assessment and decision support.

References

1. Chen, Z.; Shen, Y.; Song, Y.; Wan, X. Cross-modal Memory Networks for Radiology Report Generation. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 5904–5914. https://doi.org/10.18653/v1/2021.acl-long.459.
2. Wu, Y.; Zhan, P.; Zhang, Y.; Wang, L.; Xu, Z. Multimodal Fusion with Co-Attention Networks for Fake News Detection. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 2560–2569. https://doi.org/10.18653/v1/2021.findings-acl.226.
3. Pham, T.; Bui, T.; Mai, L.; Nguyen, A. Out of Order: How important is the sequential order of words in a sentence in Natural Language Understanding tasks? In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 1145–1160. https://doi.org/10.18653/v1/2021.findings-acl.98.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 March 2026 doi:10.20944/preprints202603.1463.v1

4. Zhang, Z.; Zhou, Z.; Wang, Y. SSEGCN: Syntactic and Semantic Enhanced Graph Convolutional Network for Aspect-based Sentiment Analysis. In Proceedings of the Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2022, pp. 4916–4925. https://doi.org/10.18653/v1/2022.naacl-main.362.
5. Li, C.; Bi, B.; Yan, M.; Wang, W.; Huang, S.; Huang, F.; Si, L. StructuralLM: Structural Pre-training for Form Understanding. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 6309–6318. https://doi.org/10.18653/v1/2021.acl-long.493.
6. Liu, Y.; Guan, R.; Giunchiglia, F.; Liang, Y.; Feng, X. Deep Attention Diffusion Graph Neural Networks for Text Classification. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 8142–8152. https://doi.org/10.18653/v1/2021.emnlp-main.642.
7. Liu, W. Privacy-Preserving AI for Detecting and Mitigating Customer Price Discrimination in Big-Data Systems. *Journal of Computer, Signal, and System Research* **2026**, *3*, 37–46.
8. Liu, W. Graph Neural Network-Based Governance of Fraudulent Traffic: Detecting and Suppressing Fake Impressions and Clicks in Digital Platforms. *European Journal of AI, Computing & Informatics* **2026**, *2*, 113–123.
9. Ribeiro, L.F.R.; Zhang, Y.; Gurevych, I. Structural Adapters in Pretrained Language Models for AMR-to-Text Generation. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 4269–4282. https://doi.org/10.18653/v1/2021.emnlp-main.351.
10. Wang, Y.; Wang, S.; Yao, Q.; Dou, D. Hierarchical Heterogeneous Graph Representation Learning for Short Text Classification. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 3091–3101. https://doi.org/10.18653/v1/2021.emnlp-main.247.
11. Zhou, Y.; Geng, X.; Shen, T.; Zhang, W.; Jiang, D. Improving Zero-Shot Cross-lingual Transfer for Multilingual Question Answering over Knowledge Graph. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 5822–5834. https://doi.org/10.18653/v1/2021.naacl-main.465.
12. Liu, Y.; Cheng, H.; Klopfer, R.; Gormley, M.R.; Schaaf, T. Effective Convolutional Attention Network for Multi-label Clinical Document Classification. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 5941–5953. https://doi.org/10.18653/v1/2021.emnlp-main.481.
13. Chen, Y.; He, Y.; Ye, H.; Xing, L.; Zhang, X.; Shi, G. Unified deep learning model for predicting fundus fluorescein angiography image from fundus structure image. *Journal of Innovative Optical Health Sciences* **2024**, *17*, 2450003.
14. Chen, Y.; Manzanera, S.; Mompeán, J.; Ruminski, D.; Grulkowski, I.; Artal, P. Increased crystalline lens coverage in optical coherence tomography with oblique scanning and volume stitching. *Biomedical Optics Express* **2021**, *12*, 1529–1542.
15. Qiao, S.; Ou, Y.; Zhang, N.; Chen, X.; Yao, Y.; Deng, S.; Tan, C.; Huang, F.; Chen, H. Reasoning with Language Model Prompting: A Survey. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2023, pp. 5368–5393. https://doi.org/10.18653/v1/2023.acl-long.294.
16. Barbaresi.; Adrien. Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations. Association for Computational Linguistics, 2021, pp. 122–131. https://doi.org/10.18653/v1/2021.acl-demo.15.
17. Wang, P.; Zhu, Z.Q.; Liang, D. Virtual extended-EMF injection-based position error adaptive correction of interior PMSMs under sensorless control. *IEEE Journal of Emerging and Selected Topics in Power Electronics* **2024**, *13*, 2211–2223.
18. Wang, P.; Yang, G.; Lin, M. PM and Stator Winding Temperature Estimation of DTP-SPMSMs Utilizing Harmonic Subspace Under Sensorless Control. *IEEE Transactions on Power Electronics* **2026**.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 March 2026 doi:10.20944/preprints202603.1463.v1

19. Wang, P.; Zhu, Z.; Liang, D. Virtual signal injection-based online full-parameter estimation of surface-mounted PMSMs without influence of position error and inverter nonlinearity. *IEEE Journal of Emerging and Selected Topics in Power Electronics* **2025**.
20. Liu, W. KV Cache and Inference Scheduling: Energy Modeling for High-QPS Services. *Journal of Industrial Engineering and Applied Science* **2026**, *4*, 34–41.
21. Wang, Z.; Xu, Y.; Cui, L.; Shang, J.; Wei, F. LayoutReader: Pre-training of Text and Layout for Reading Order Detection. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 4735–4744. https://doi.org/10.18653/v1/2021.emnlp-main.389.
22. Lu, X.; West, P.; Zellers, R.; Le Bras, R.; Bhagavatula, C.; Choi, Y. NeuroLogic Decoding: (Un)supervised Neural Text Generation with Predicate Logic Constraints. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 4288–4299. https://doi.org/10.18653/v1/2021.naacl-main.339.
23. DeYoung, J.; Beltagy, I.; van Zuylen, M.; Kuehl, B.; Wang, L.L. MS^2: Multi-Document Summarization of Medical Studies. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 7494–7513. https://doi.org/10.18653/v1/2021.emnlp-main.594.
24. Ke, P.; Ji, H.; Ran, Y.; Cui, X.; Wang, L.; Song, L.; Zhu, X.; Huang, M. JointGT: Graph-Text Joint Representation Learning for Text Generation from Knowledge Graphs. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 2526–2538. https://doi.org/10.18653/v1/2021.findings-acl.223.
25. Shi, J.; Cao, S.; Hou, L.; Li, J.; Zhang, H. TransferNet: An Effective and Transparent Framework for Multi-hop Question Answering over Relation Graph. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 4149–4158. https://doi.org/10.18653/v1/2021.emnlp-main.341.
26. Ding, H.; Luo, X. AttentionRank: Unsupervised Keyphrase Extraction using Self and Cross Attentions. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 1919–1928. https://doi.org/10.18653/v1/2021.emnlp-main.146.
27. Xie, C.; Tong, H.; Xu, G.; Chen, Y.; Luking, L.; Chen, Y. Knowledge Calibration Distillation. In Proceedings of the 2025 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2025, pp. 1–7.
28. Han, Z.; Ding, Z.; Ma, Y.; Gu, Y.; Tresp, V. Learning Neural Ordinary Equations for Forecasting Future Links on Temporal Knowledge Graphs. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 8352–8364. https://doi.org/10.18653/v1/2021.emnlp-main.658.
29. Saxena, A.; Kochsiek, A.; Gemulla, R. Sequence-to-Sequence Knowledge Graph Completion and Question Answering. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2022, pp. 2814–2828. https://doi.org/10.18653/v1/2022.acl-long.201.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.