

FedDep-GT-LLM: Federated Graph Transformer and Clinical Language Modeling for Multimodal Depression Analysis

BalaKrishan N

balakrishnacse1984@gmail.com

SRM Institute of Science and Technology

Mukesh Krishnan MB

SRM Institute of Science and Technology

Ganesh D

Mohan Babu University (Erstwhile Sree Vidyanikethan Engineering College(Autonomous))

Article

Keywords:

Posted Date: May 31st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9689895/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

FedDep-GT-LLM: Federated Graph Transformer and Clinical Language Modeling for Multimodal Depression Analysis

N. Balakrishna^{1,*}, M.B. Mukesh Krishnan², and D. Ganesh³

¹Research Scholar, Department of Networking and Communication, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India

²Professor, Department of Networking and Communications, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India

³Associate Professor of CSE, Mohan Babu University (Erstwhile Sree Vidyanikethan Engineering College(Autonomous), Tirupati, Andhra Pradesh, India

*balakrishnacse1984@gmail.com

ABSTRACT

Early identification of depression severity remains a major challenge in precision mental healthcare because of the heterogeneity of behavioral, linguistic, physiological, and neuro cognitive indicators. Existing artificial intelligence based depression assessment methods are often limited by inadequate privacy preservation, interpretability, poor cross-domain generalization, and insufficient multimodal fusion. To address these limitations, we propose an explainable federated multimodal learning framework that integrates Graph Transformers(GTs), large language models (LLMs), and adaptive cross-modal attention for early depression severity prediction. The proposed framework, FedDep-GT-LLM, leverages decentralized federated optimization to preserve patient privacy while enabling collaborative learning from distributed multimodal datasets across institutions. It fuses speech, facial expressions, textual transcripts, and EEG signals depending on dataset availability, using modality-specific encoders, followed by a heterogeneous graph transformer for relational representation learning. A domain-adapted clinical LLM extracts contextual semantic embeddings from patient conversations and self-reported narratives, whereas a SHAP guided explainability module enhances clinical interpretability. We evaluate the framework on DAIC-WOZ, E-DAIC, AVEC 2019, and MODMA datasets under cross-dataset and federated settings. FedDep-GT-LLM achieved an average accuracy of 94.8%, sensitivity of 94.1%, specificity of 97.2%, and a concordance correlation coefficient of 0.923 for depression severity estimation, demonstrating competitive performance against recent centralized and federated baseline methods. Overall, the proposed framework provides a scalable, privacy-preserving, and clinically interpretable approach for real-world early depression screening and personalized mental health monitoring.

1 Introduction

Depression is one of the most prevalent psychiatric disorders worldwide and remains a major contributor to disability, suicide risk, and reduced quality of life. According to recent global mental health statistics, depressive disorders affect more than 280 million individuals globally, with prevalence rates increasing substantially following the COVID-19 pandemic¹. Depression is characterized by persistent sadness, cognitive dysfunction, emotional instability, sleep disturbances, social withdrawal, and suicidal ideation. Consequently, early identification and accurate severity estimation are essential for timely clinical intervention and improved patient outcomes. However, conventional psychiatric assessment primarily relies on subjective clinical interviews and self-report scales such as the Patient Health Questionnaire (PHQ-9) and Hamilton Depression Rating Scale (HAM-D), which are often affected by reporting bias, inter-clinician variability, and delayed diagnosis.

Recent advances in artificial intelligence (AI), affective computing, and multimodal deep learning have enabled automated depression analysis using behavioral and physiological biomarkers. Multimodal learning frameworks integrate heterogeneous data sources, including speech characteristics, facial expressions, textual transcripts, electroencephalography (EEG), wearable sensor signals, and electronic health records (EHRs), to better capture the complex manifestations of depressive behavior. Deep learning approaches based on convolutional neural networks (CNNs), recurrent neural networks (RNNs), transformers, and multimodal fusion architectures have demonstrated promising performance for depression classification and severity estimation^{2,3}. In particular, transformer-based architectures have shown strong capability in modeling long-range contextual dependencies and temporal behavioral dynamics across heterogeneous modalities.

Despite these advances, several important challenges remain unresolved in existing depression prediction systems. First,

most current approaches rely on centralized learning frameworks that require sensitive patient data to be collected and stored on centralized servers, raising significant privacy and regulatory concerns in mental healthcare applications. Second, conventional multimodal fusion strategies often inadequately capture dynamic inter-modal dependencies among heterogeneous behavioral signals, thereby limiting representation robustness and semantic interaction modeling capability⁴. Third, many deep learning models operate as black-box systems with limited interpretability, reducing clinician trust and limiting their applicability in practical psychiatric settings.

To address these limitations, graph neural networks (GNNs) and graph transformer architectures have recently emerged as effective paradigms for modeling structured multimodal interactions. Graph-based architectures explicitly capture relational dependencies among heterogeneous behavioral and physiological biomarkers using node-level semantic interaction modeling. Recent studies have shown that graph-based multimodal learning improves depression detection performance by modeling latent dependencies among EEG signals, speech patterns, and interview transcripts⁵. In parallel, transformer-based clinical language models have significantly improved contextual semantic understanding in psychiatric natural language processing applications. Domain-adapted transformer encoders can effectively model emotionally informative linguistic patterns embedded within psychiatric interviews, counseling transcripts, and patient narratives. Recent multimodal semantic learning frameworks have demonstrated improved depression severity estimation performance on benchmark datasets such as DAIC-WOZ and E-DAIC⁶. Nevertheless, the integration of graph-based multimodal relational reasoning, privacy-preserving federated optimization, contextual semantic modeling, and explainable AI within a unified framework remains relatively underexplored.

Federated learning has emerged as a promising distributed optimization paradigm for collaborative healthcare intelligence without directly sharing raw patient data⁷. In federated settings, institutions collaboratively train global models through parameter aggregation while preserving local data privacy. Although federated learning has shown encouraging results in medical imaging, wearable healthcare analytics, and clinical prediction systems, most existing federated mental healthcare frameworks rely on relatively shallow aggregation mechanisms and do not effectively integrate graph-based multimodal reasoning and explainability strategies required for trustworthy psychiatric AI systems⁸.

Motivated by these challenges, this study proposes **FedDep-GT-LLM**, an explainable federated multimodal learning framework for early depression severity prediction. The proposed architecture integrates privacy-preserving federated optimization, heterogeneous graph transformers, adaptive cross-modal attention fusion, and domain-adapted clinical semantic representation learning within a unified multimodal framework. Specifically, modality-specific encoders extract affective feature embeddings from speech, facial expressions, textual transcripts, and physiological biomarkers. A heterogeneous graph transformer models structural dependencies among multimodal behavioral representations, whereas a ClinicalBERT-based semantic encoder extracts contextual linguistic embeddings from psychiatric conversations and patient narratives. To improve transparency and clinician interpretability, the framework further incorporates SHAP-guided explainability⁹ and graph attention visualization mechanisms¹⁰ to identify influential multimodal biomarkers associated with depression severity prediction.

The proposed framework is evaluated on multiple benchmark datasets, including DAIC-WOZ, E-DAIC, AVEC 2019, and MODMA under heterogeneous federated learning settings. Experimental results demonstrate improved performance over recent centralized and federated baseline methods in terms of classification accuracy, sensitivity, specificity, and concordance correlation coefficient (CCC). Furthermore, explainability analysis reveals clinically meaningful multimodal behavioral biomarkers associated with depressive symptoms, supporting interpretable and privacy-aware AI-assisted psychiatric assessment.

The major contributions of this study are summarized as follows:

1. A unified privacy-preserving multimodal depression analysis framework integrating graph transformers, domain-adapted semantic representation learning, adaptive federated aggregation, and explainable AI under heterogeneous psychiatric settings.
2. A heterogeneous graph transformer-based multimodal representation learning strategy capable of modeling complex inter-modal dependencies among behavioral and physiological biomarkers.
3. A privacy-preserving federated optimization mechanism enabling decentralized collaborative learning across institutions without sharing raw patient data.
4. An explainability framework integrating SHAP analysis, graph attention visualization, and multimodal saliency analysis for clinically interpretable depression prediction.
5. Comprehensive experimental validation across multiple benchmark depression datasets under federated and cross-dataset evaluation settings, demonstrating improved robustness and generalization compared with recent centralized and federated baseline approaches.

The remainder of this paper is organized as follows. Section II reviews recent advances in multimodal depression analysis, graph neural networks, federated learning, and explainable AI techniques for psychiatric healthcare applications. Section

III presents the proposed FedDep-GT-LLM framework, including multimodal representation learning, graph transformer modeling, federated optimization, and explainability modules. Section IV describes the experimental setup, benchmark datasets, evaluation metrics, and quantitative performance analysis, followed by cross-dataset and federated evaluation studies. Finally, Section V concludes the paper and discusses future directions for privacy-aware multimodal mental healthcare systems.

2 Related Work

Recent advances in artificial intelligence, affective computing, multimodal learning, and privacy-preserving healthcare analytics have significantly accelerated research on automated depression detection and severity estimation systems. Existing studies can broadly be categorized into four major research directions: (1) deep learning-based depression analysis, (2) multimodal and graph-based behavioral modeling, (3) federated learning for mental healthcare, and (4) explainable artificial intelligence and clinical language models for psychiatric decision support.

2.1 Deep Learning for Depression Detection

Early depression detection studies primarily relied on conventional machine learning algorithms using handcrafted acoustic, linguistic, and visual descriptors extracted from psychiatric interviews and social interaction data. Classical classifiers such as support vector machines, random forests, logistic regression, and gradient boosting models remain widely used as benchmark approaches; however, their performance is often constrained by limited contextual modeling and shallow feature representations^{11–13}.

Deep learning (DL) techniques have substantially improved automated depression analysis by learning task-specific representations directly from raw multimodal inputs. Convolutional neural networks (CNNs) are effective for extracting spatial representations from facial expressions, spectrograms, and physiological signals, whereas recurrent architectures such as long short-term memory (LSTM) networks are capable of modeling temporal behavioral dynamics in speech and emotional interactions^{14,15}. Hybrid CNN-LSTM and attention-guided architectures have further demonstrated improved depression recognition performance by jointly modeling spatial and temporal behavioral patterns¹⁵.

More recently, transformer-based architectures have emerged as powerful representation learning models for depression detection because of their capability to capture long-range contextual dependencies using self-attention mechanisms. Transformer-enhanced multimodal fusion and hierarchical transformer frameworks have shown improved robustness compared with conventional DL architectures^{2,16,17}. In addition, pretrained transformer representations such as wav2vec 2.0 and self-supervised speech encoders have improved performance in low-resource depression detection settings^{3,18}. Despite these advances, many existing DL-based depression analysis systems still exhibit limited interpretability and reduced robustness under heterogeneous multimodal clinical conditions.

2.2 Multimodal and Graph-Based Behavioral Modeling

Depression is a heterogeneous psychiatric disorder involving complex emotional, behavioral, cognitive, and physiological manifestations. Consequently, multimodal learning frameworks have become increasingly important for integrating complementary information derived from speech, facial expressions, textual narratives, EEG signals, wearable sensors, and clinical metadata. Recent studies have demonstrated that multimodal architectures substantially improve depression recognition performance by jointly modeling heterogeneous behavioral biomarkers and contextual interactions.

Liu et al.⁴ proposed a local-to-global multimodal graph neural network integrating EEG, speech, and visual biomarkers through hierarchical multimodal fusion learning. Similarly, Yang and Zheng¹⁹ introduced a multi-level spatiotemporal graph attention fusion framework capable of capturing complex multimodal behavioral interactions for depression analysis. Jia et al.¹⁷ further demonstrated that transformer-enhanced multimodal learning frameworks improve robustness in depression detection by jointly modeling contextual interactions across modalities.

Graph neural networks (GNNs) and graph transformers have recently emerged as effective paradigms for modeling structured multimodal dependencies. Unlike conventional fusion architectures, graph-based models explicitly capture relational interactions among heterogeneous behavioral and physiological biomarkers. Xing et al.⁵ proposed an adaptive multi-graph neural network for major depressive disorder detection using multimodal feature fusion learning. Agarwal et al.²⁰ further introduced a multiview graph-based interview representation framework for depression severity estimation using semantic relational modeling within psychiatric interview transcripts.

Recent graph transformer architectures have additionally demonstrated improved capability for dynamic contextual interaction learning using attention-guided message passing. Ding et al.²¹ proposed a Dynamic Graph Attention Network (DGAT) for EEG-based emotion recognition, whereas Ma et al.²² introduced a GNN-transformer cross-view contrastive learning framework for multimodal conversational emotion recognition. Although graph-based multimodal learning has shown promising results, several limitations remain unresolved. Many existing methods rely on static graph construction strategies and insufficiently address temporal multimodal interactions, computational complexity, and robustness under incomplete modality

availability. Furthermore, graph transformer-based multimodal depression analysis under federated privacy-preserving settings remains relatively underexplored.

2.3 Federated Learning in Mental Healthcare

The increasing adoption of artificial intelligence in healthcare has raised substantial concerns regarding patient privacy, data security, and regulatory compliance. Psychiatric healthcare data are particularly sensitive because mental health records frequently contain confidential emotional and behavioral information. Federated learning (FL) has emerged as a promising distributed optimization paradigm that enables collaborative model training without directly sharing raw patient data across institutions^{7,23}.

Recent studies have explored federated learning for privacy-preserving mental healthcare analytics. Zhang et al.²⁴ proposed an adaptive federated multimodal learning framework for distributed mental health assessment using heterogeneous behavioral representations. Similarly, Chen et al.²⁵ introduced a federated graph neural network framework for multimodal healthcare prediction capable of modeling relational dependencies while preserving institutional privacy. Warrier et al.²⁶ further highlighted the growing importance of federated learning in depression analysis, emotion recognition, and behavioral healthcare applications through a comprehensive systematic review.

Despite its potential, several important challenges remain in existing federated mental healthcare systems. Most current approaches rely on conventional aggregation strategies such as Federated Averaging (FedAvg), which inadequately address client heterogeneity and multimodal imbalance across institutions. In addition, existing federated depression analysis frameworks rarely integrate graph-based multimodal reasoning, transformer-based contextual modeling, and explainability mechanisms required for clinically trustworthy psychiatric AI systems. Robust federated learning under heterogeneous non-identically distributed (non-IID) multimodal psychiatric settings therefore remains an important open research problem.

2.4 Explainable Artificial Intelligence and Clinical Language Models

Explainability has become a critical requirement for the trustworthy deployment of artificial intelligence systems in psychiatric healthcare. Clinicians require transparent reasoning mechanisms and biomarker-level interpretations to validate automated depression predictions. Consequently, explainable artificial intelligence (XAI) techniques such as SHAP, Grad-CAM, saliency mapping, and attention visualization have increasingly been integrated into mental healthcare applications^{8,9}. Existing studies have shown that explainability mechanisms improve clinician trust and facilitate the interpretation of depression-related behavioral biomarkers. Zogan et al.²⁷ proposed an explainable depression detection framework using multimodal behavioral features and hybrid deep learning architectures. Kumar et al.²⁸ developed an interpretable multimodal depression analysis framework integrating attention-guided fusion and explainable decision modeling for trustworthy mental healthcare analysis.

In parallel, transformer-based clinical language models have significantly improved contextual semantic understanding in psychiatric natural language processing applications. Domain-adapted transformer encoders can effectively model emotionally informative linguistic patterns embedded within psychiatric interviews and counseling transcripts. Sadeghi et al.⁶ demonstrated that integrating semantic language representations with facial expression analysis substantially improves depression severity estimation performance. Furthermore, contextual semantic embeddings generated from clinical language models can support interpretable psychiatric reasoning and clinician-AI collaboration. Nevertheless, several limitations remain in existing explainable multimodal psychiatric frameworks. Most current XAI approaches primarily provide post-hoc interpretations without explicitly modeling relational dependencies among heterogeneous modalities. In addition, existing semantic language model-assisted depression analysis systems rarely integrate graph-based multimodal reasoning, federated optimization, and privacy-preserving explainability within a unified framework. Consequently, developing an interpretable and privacy-aware multimodal depression prediction framework capable of jointly modeling contextual semantic information and heterogeneous behavioral interactions remains an important research challenge.

To address these limitations, the proposed FedDep-GT-LLM framework integrates heterogeneous graph transformers, adaptive federated optimization, contextual semantic representation learning, and explainable AI within a unified multimodal depression analysis framework. The proposed architecture is designed to improve multimodal relational reasoning, semantic contextual understanding, federated robustness under heterogeneous clinical distributions, and clinically interpretable depression severity estimation.

3 Methodology

3.1 Overview of the Proposed Framework

FedDep-GT-LLM is designed for distributed depression analysis using heterogeneous psychiatric data collected across multiple healthcare institutions. The proposed architecture integrates modality-specific representation learning, graph transformer-based relational modeling, contextual semantic learning, adaptive multimodal fusion, federated optimization, and explainable artificial intelligence within a unified privacy-preserving framework. The overall workflow of the proposed framework is illustrated in

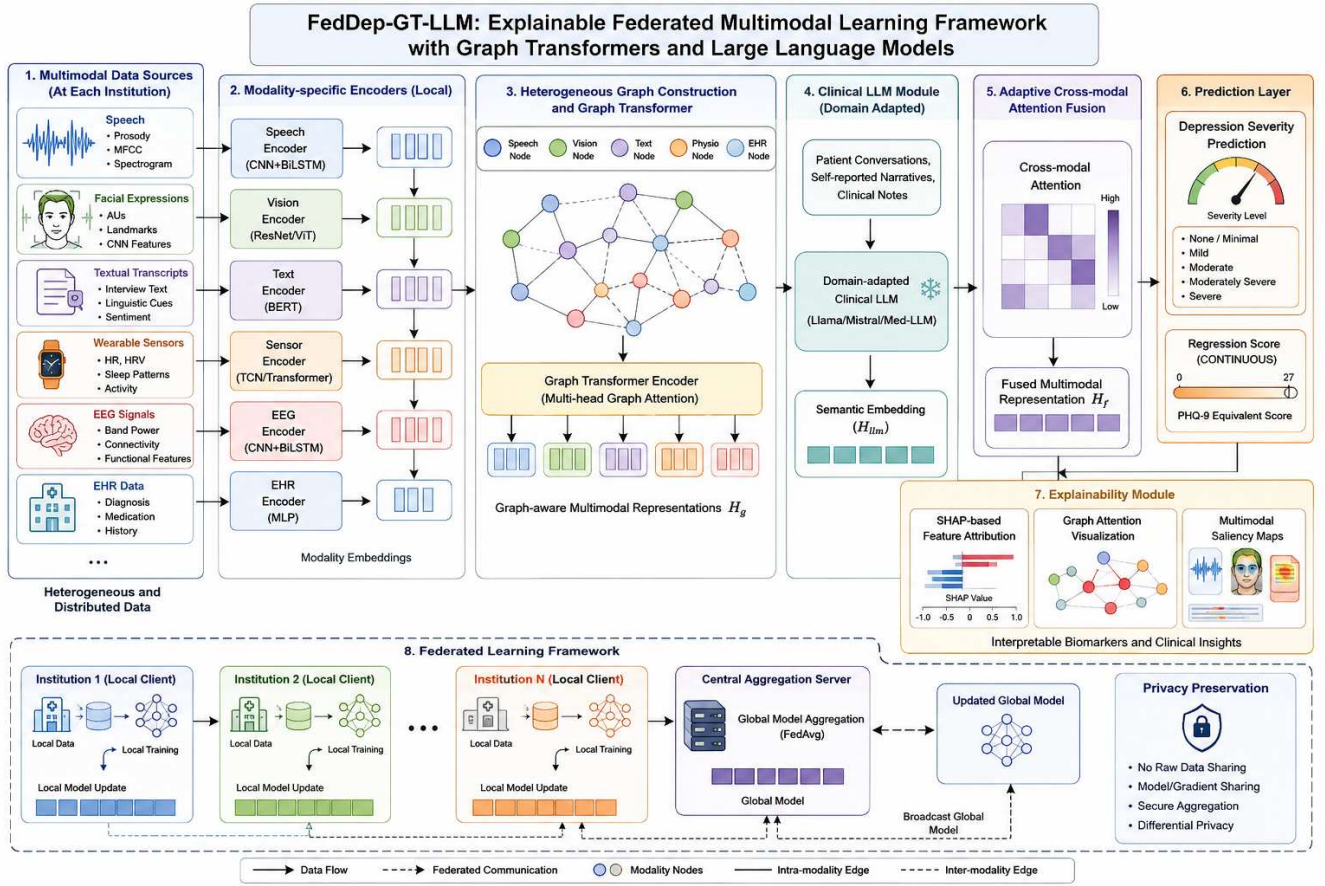


Figure 1. Architecture of the proposed FedDep-GT-LLM framework integrating multimodal representation learning, graph transformers, clinical LLMs, federated optimization, and explainable AI for early depression severity prediction.

Fig. 1. Initially, heterogeneous data modalities including speech recordings, facial videos, psychiatric interview transcripts, physiological biomarkers, wearable sensor signals, and clinical metadata undergo modality-specific preprocessing and feature extraction to generate robust behavioral representations. The extracted transcript features are additionally processed using a domain-adapted ClinicalBERT-based semantic encoder to capture depression-related semantic and linguistic patterns embedded within psychiatric conversations.

Subsequently, modality-specific embeddings are integrated into a heterogeneous graph transformer architecture, where attention-guided relational learning dynamically models contextual dependencies among multimodal behavioral representations. The graph-aware embeddings are then fused with semantic contextual representations using adaptive cross-modal attention mechanisms to generate unified depression-aware multimodal representations. The fused multimodal representations are optimized using a decentralized federated learning framework operating under heterogeneous non-identically distributed institutional environments while preserving patient privacy. Finally, SHAP-based feature attribution and graph attention visualization are incorporated to provide transparent multimodal interpretation and clinically meaningful depression severity estimation. Unlike conventional multimodal learning frameworks that rely on static feature concatenation or shallow aggregation strategies, the proposed framework dynamically models multimodal contextual dependencies through graph transformer-based relational reasoning and adaptive attention-guided fusion. This improves contextual behavioral understanding, decentralized optimization robustness, and clinically interpretable depression severity prediction across heterogeneous healthcare environments.

3.2 Multimodal Data Preprocessing

3.2.1 Audio Preprocessing

Speech recordings were first resampled to 16 kHz mono-channel format to ensure consistent acoustic representation learning across heterogeneous datasets. Background noise and non-speech artifacts were reduced using spectral subtraction and Wiener filtering techniques to improve speech quality under varying recording environments. Voice activity detection (VAD) based on short-term energy estimation was subsequently applied to remove silent speech segments and non-informative acoustic regions.

The denoised speech signals were segmented using 25 ms Hamming windows with a 10 ms hop size prior to short-time Fourier transform (STFT) analysis. Log-Mel spectrograms were then extracted using 64 Mel filter banks and a 512-point FFT configuration. In addition, 13-dimensional Mel-frequency cepstral coefficients (MFCCs) together with their first- and second-order temporal derivatives were computed to capture spectral envelope variations associated with depressive speech behavior²⁹. Prosodic and temporal behavioral biomarkers including pitch contours, speaking rate, pause duration, energy variation, jitter, shimmer, and temporal speech rhythm statistics were additionally extracted because depressive speech is frequently associated with monotonic prosody, psychomotor slowing, reduced articulation dynamics, and increased vocal hesitation^{30,31}.

Amplitude normalization and temporal smoothing were further applied to reduce inter-speaker variability and improve robustness across heterogeneous clinical recording environments. These preprocessing operations improve acoustic feature stability, reduce environmental variability, and enhance depression-related behavioral representation learning under real-world psychiatric interview conditions.

3.2.2 Visual Behavioral Preprocessing

Facial video recordings were processed using a multi-stage visual preprocessing pipeline to ensure robust behavioral representation learning under heterogeneous recording conditions. Initially, facial regions were detected frame-wise using the Multi-task Cascaded Convolutional Network (MTCNN) face detection algorithm, followed by facial landmark localization and geometric alignment to normalize pose variations and improve spatial consistency across video frames³². Detected facial frames were subsequently resized to 224×224 spatial resolution and intensity normalized prior to visual feature extraction. To capture depression-related behavioral cues, visual representations associated with emotional expressiveness, gaze behavior, facial activity dynamics, and psychomotor patterns were extracted. Facial action units (AUs) corresponding to emotion-related muscular activations were additionally estimated using automated facial behavior analysis techniques³³. These behavioral indicators are clinically relevant because depressive disorders are frequently associated with emotional flattening, reduced facial expressiveness, abnormal gaze behavior, and psychomotor retardation during psychiatric interviews^{34,35}.

To improve temporal behavioral consistency, frame-level facial representations were smoothed using temporal filtering and sequence interpolation techniques to reduce visual noise caused by illumination variation, motion artifacts, and transient facial inconsistencies. Data augmentation operations including horizontal flipping, random brightness adjustment, and slight rotational perturbation were additionally applied during training to improve generalization robustness under varying recording environments. Finally, temporally ordered facial sequences were segmented into fixed-length behavioral windows prior to transformer-based visual representation learning. These preprocessing operations improve spatial alignment consistency, behavioral temporal stability, and depression-related visual feature robustness across heterogeneous psychiatric video recordings.

3.2.3 Textual Transcript Preprocessing

Psychiatric interview transcripts were preprocessed using a standardized clinical natural language processing pipeline to improve contextual semantic representation learning. Initially, textual transcripts were cleaned through lowercasing, punctuation normalization, whitespace correction, and sentence segmentation. Subsequently, tokenization and lemmatization were applied to normalize lexical variations while preserving clinically meaningful semantic content. Common non-informative stop words were selectively filtered, whereas psychologically relevant negations, emotional expressions, and symptom-related linguistic patterns were intentionally retained because they frequently contain important indicators associated with depressive cognition and emotional behavior. To preserve psychiatric contextual semantics, domain-specific preprocessing strategies were additionally employed to maintain emotionally informative expressions related to hopelessness, self-deprecation, social withdrawal, cognitive distortion, negative emotional polarity, and suicidal ideation. Linguistic disfluencies, hesitation markers, and conversational irregularities commonly observed in depressive speech were also preserved because such behavioral language patterns frequently provide clinically relevant depression biomarkers^{36,37}.

The processed transcripts were subsequently encoded using a domain-adapted ClinicalBERT-based semantic encoder capable of learning contextual psychiatric language representations from clinical and biomedical corpora³⁸. Unlike conventional word embedding approaches, transformer-based contextual language modeling enables improved semantic understanding of emotionally nuanced psychiatric narratives and long-range conversational dependencies. Preserving depression-relevant linguistic cues improves contextual semantic reasoning and enhances depression severity estimation performance.

3.3 Modality-Specific Representation Learning

3.3.1 Acoustic Representation Encoder

Acoustic behavioral representations were learned using a transformer-enhanced speech encoder operating on log-Mel spectrogram and prosodic feature representations extracted from psychiatric interview recordings. Let $X_a \in \mathbb{R}^{T_a \times d_a}$ denote the acoustic feature sequence, where T_a represents the temporal sequence length and d_a corresponds to the acoustic feature dimensionality.

The transformer-based acoustic encoder generates contextual speech embeddings as $H_a = f_a(X_a)$, where $f_a(\cdot)$ denotes the transformer-based acoustic representation encoder and $H_a \in \mathbb{R}^{T_a \times d_h}$ represents the learned contextual acoustic embeddings.

The acoustic encoder consists of stacked multi-head self-attention and feed-forward transformer layers capable of modeling long-range temporal dependencies within speech sequences. For each transformer layer, self-attention weights are computed as

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (1)$$

where Q , K , and V denote the query, key, and value projection matrices derived from the acoustic embeddings, and d_k represents the attention scaling dimension. Multi-head attention enables the encoder to jointly capture diverse temporal acoustic dependencies across different speech regions and conversational contexts. The transformer encoder captures depression-related acoustic biomarkers including monotonic prosody, reduced speech energy, prolonged pause duration, vocal hesitation, psychomotor slowing, articulation irregularities, and temporal speaking rhythm variation. Compared with conventional recurrent neural architectures, transformer-based acoustic modeling improves contextual temporal representation learning by enabling parallel sequence modeling and long-range dependency analysis through self-attention mechanisms. Residual skip connections and layer normalization were additionally incorporated to stabilize acoustic representation learning and improve optimization convergence. The generated contextual acoustic embeddings therefore provide robust behavioral speech representations under heterogeneous recording environments and variable conversational dynamics commonly observed in psychiatric interview datasets.

3.3.2 Visual Behavioral Encoder

Visual behavioral representations were learned using a convolutional transformer encoder operating on temporally ordered facial expression sequences extracted from psychiatric interview videos. Let $X_v \in \mathbb{R}^{T_v \times H \times W \times C}$ denote the visual input sequence, where T_v represents the temporal sequence length, H and W correspond to spatial frame dimensions, and C denotes the number of image channels. The visual encoder generates contextual behavioral embeddings as $H_v = f_v(X_v)$, where $f_v(\cdot)$ denotes the convolutional transformer-based visual encoder and $H_v \in \mathbb{R}^{T_v \times d_v}$ represents the learned visual behavioral embeddings. The proposed visual encoder consists of a convolutional feature extraction backbone followed by stacked transformer layers for temporal contextual modeling. Initially, convolutional layers extract local spatial facial representations including facial muscle activations, eye movement patterns, gaze behavior, and expression-related visual structures. Subsequently, transformer layers model long-range temporal dependencies and contextual behavioral interactions across facial sequences using multi-head self-attention mechanisms. For temporal behavioral modeling, self-attention weights are computed as

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (2)$$

where Q , K , and V denote query, key, and value projections derived from visual embeddings, and d_k represents the attention scaling dimension. Multi-head attention enables the encoder to dynamically emphasize temporally informative facial regions and behavioral transitions associated with depressive symptoms. The visual encoder captures depression-related behavioral characteristics including emotional flattening, reduced facial expressiveness, diminished eye contact, psychomotor retardation, gaze irregularities, and behavioral disengagement. Compared with conventional frame-wise facial analysis approaches, the proposed convolutional transformer architecture models temporal behavioral continuity throughout psychiatric interview sessions, thereby improving contextual behavioral understanding and depression-related visual representation learning. Residual skip connections, layer normalization, and temporal positional embeddings were additionally incorporated to stabilize optimization and preserve sequential behavioral information across facial sequences. Consequently, the learned visual embeddings provide robust depression-aware behavioral representations under heterogeneous recording conditions and variable psychiatric interview environments.

3.3.3 Clinical Semantic Encoder

To improve contextual psychiatric reasoning, psychiatric interview transcripts were processed using a ClinicalBERT-based semantic encoder capable of learning domain-specific contextual language representations from clinical and biomedical corpora. Let $X_t = \{w_1, w_2, \dots, w_n\}$, denote a tokenized psychiatric interview transcript consisting of n sequential tokens. The semantic encoder generates contextual language embeddings as $H_t = f_t(X_t)$, where $f_t(\cdot)$ denotes the ClinicalBERT-based semantic encoder and $H_t \in \mathbb{R}^{n \times d_t}$ represents the generated contextual semantic embedding matrix, with d_t corresponding to the embedding dimensionality.

The ClinicalBERT encoder consists of stacked bidirectional transformer layers employing self-attention mechanisms to model long-range contextual dependencies across psychiatric conversations. Unlike conventional static word embedding

approaches, transformer-based contextual semantic modeling dynamically captures token relationships according to the surrounding conversational context, thereby enabling improved understanding of emotionally nuanced psychiatric narratives and implicit depressive language patterns. For contextual semantic interaction modeling, self-attention weights are computed as

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

where Q , K , and V denote the query, key, and value projections derived from token-level embeddings, and d_k represents the attention scaling dimension. Multi-head attention enables the semantic encoder to jointly capture diverse contextual linguistic dependencies across psychiatric interview sequences.

The semantic encoder captures emotionally informative linguistic patterns embedded within psychiatric interviews, including hopelessness-related narratives, cognitive distortions, negative emotional polarity, suicidal ideation indicators, self-deprecating language, and social withdrawal patterns. Because ClinicalBERT is pretrained on large-scale clinical and biomedical corpora, it enables improved psychiatric semantic understanding compared with conventional generic language models trained on non-clinical textual data. To generate a unified transcript-level semantic representation, attention pooling was subsequently employed to aggregate token-level contextual embeddings as follows:

$$h_t = \sum_{i=1}^n \beta_i H_t^{(i)} \quad (4)$$

where $H_t^{(i)}$ denotes the contextual embedding associated with token i , and β_i represents semantic attention coefficients computed as

$$\beta_i = \frac{\exp(W_s H_t^{(i)})}{\sum_{j=1}^n \exp(W_s H_t^{(j)})} \quad (5)$$

where W_s denotes learnable semantic attention parameters. The attention pooling mechanism dynamically emphasizes emotionally informative linguistic components while reducing the influence of semantically irrelevant conversational content. Residual transformer connections and layer normalization were additionally employed to stabilize semantic representation learning and improve optimization convergence. Consequently, the generated semantic embeddings provide robust depression-aware contextual representations that improve multimodal behavioral interpretation and psychiatric reasoning during depression severity estimation.

3.4 Heterogeneous Graph Transformer Modeling

Depressive behavior emerges through complex contextual interactions among heterogeneous behavioral modalities rather than isolated unimodal characteristics. Conventional multimodal fusion strategies frequently rely on static feature concatenation or shallow aggregation mechanisms that inadequately capture dynamic inter-modal relationships and contextual behavioral dependencies. To address these limitations, the proposed framework integrates modality-specific embeddings into a heterogeneous multimodal graph transformer architecture capable of modeling structured contextual interactions among acoustic, visual, and semantic behavioral representations. The heterogeneous multimodal graph is formally defined as $G = (V, E)$, where V denotes the set of multimodal behavioral nodes and E represents contextual relational edges among modalities. Each graph node corresponds to modality-specific embeddings extracted from the acoustic encoder, visual behavioral encoder, and ClinicalBERT-based semantic encoder. Let $V = \{v_a, v_v, v_t\}$, where v_a , v_v , and v_t denote acoustic, visual, and semantic behavioral nodes, respectively. Initial graph edges were constructed using cosine similarity between modality-specific embeddings as follows:

$$e_{ij} = \frac{h_i^T h_j}{\|h_i\| \|h_j\|} \quad (6)$$

where h_i and h_j denote modality-specific node embeddings. The cosine similarity initialization enables the graph structure to preserve contextual multimodal relationships prior to graph transformer optimization. To dynamically model contextual multimodal dependencies, graph transformer layers perform attention-guided message passing across neighboring behavioral nodes. For each node i , graph attention coefficients are computed as

$$\alpha_{ij} = \frac{\exp((W_q h_i)^T (W_k h_j))}{\sum_{k \in \mathcal{N}(i)} \exp((W_q h_i)^T (W_k h_k))}, \quad (7)$$

where:

- α_{ij} denotes graph attention coefficients,
- W_q and W_k represent learnable query and key projection matrices,
- h_i and h_j denote neighboring node embeddings,
- $\mathcal{N}(i)$ represents the neighborhood associated with node i .

The graph attention mechanism dynamically quantifies contextual dependency strength between multimodal behavioral nodes. Higher attention values indicate stronger contextual interactions contributing to depression severity estimation, whereas lower attention weights suppress weak or semantically irrelevant multimodal relationships. Unlike conventional graph neural networks that employ static neighborhood aggregation, the proposed graph transformer adaptively learns contextual multimodal dependencies using attention-guided relational reasoning. The updated graph-aware node representation is subsequently computed as

$$h'_i = \sum_{j \in \mathcal{N}(i)} \alpha_{ij} W_v h_j,$$

where W_v denotes the learnable value projection matrix and h'_i represents the updated graph-aware multimodal embedding associated with node i . Through iterative message passing and contextual attention propagation, the graph transformer captures higher-order multimodal interactions among speech characteristics, facial behavioral patterns, and contextual semantic representations associated with depressive symptoms. Multi-head graph attention further enables the architecture to jointly model diverse contextual behavioral relationships across heterogeneous modalities and varying psychiatric interview conditions. Multiple graph transformer layers were hierarchically stacked to capture increasingly abstract multimodal contextual interactions across behavioral representations. Residual skip connections, layer normalization, and dropout regularization were additionally incorporated to stabilize optimization, improve convergence robustness, and reduce overfitting during multimodal graph learning. Consequently, the graph transformer architecture generates robust contextual multimodal embeddings capable of dynamically modeling heterogeneous behavioral interactions associated with depression severity estimation. The generated graph-aware representations are subsequently integrated with semantic contextual embeddings using adaptive cross-modal attention fusion for unified depression-aware multimodal representation learning.

3.5 Adaptive Cross-Modal Attention Fusion

Effective multimodal depression analysis requires the integration of heterogeneous behavioral representations while preserving contextual relationships among modalities. Conventional multimodal fusion approaches frequently rely on static feature concatenation or fixed-weight aggregation strategies that inadequately model dynamic modality importance under varying psychiatric conditions. To address these limitations, the proposed framework incorporates an adaptive cross-modal attention fusion mechanism capable of dynamically integrating graph-aware multimodal embeddings and contextual semantic representations according to behavioral relevance. Let $H_g \in \mathbb{R}^{N \times d_g}$ denote the graph-aware multimodal embeddings generated by the heterogeneous graph transformer, where N represents the number of multimodal behavioral nodes and d_g corresponds to the graph embedding dimensionality. Similarly, let $h_t \in \mathbb{R}^{d_t}$ denote the contextual semantic representation generated by the ClinicalBERT-based semantic encoder.

Cross-modal interactions between graph-aware behavioral representations and contextual semantic embeddings are modeled using scaled dot-product attention:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (8)$$

where Q , K , and V denote the query, key, and value projection matrices derived from multimodal embeddings, and d_k represents the attention scaling dimensionality. The attention mechanism dynamically estimates contextual dependency strength between multimodal behavioral representations and semantic contextual features. Multi-head cross-modal attention enables

the framework to jointly model diverse multimodal interactions across heterogeneous behavioral modalities. Informative depression-related modalities are therefore assigned higher contextual attention weights, whereas noisy, weak, or incomplete modalities receive lower attention contributions during multimodal representation learning. This adaptive weighting mechanism is particularly important in psychiatric healthcare environments because behavioral modalities frequently exhibit varying reliability, recording quality, and incomplete modality availability across institutions. The fused multimodal representation is subsequently computed as

$$H_f = \sigma(W_f[H_g; h_t] + b_f), \quad (9)$$

where $[H_g; h_t]$ denotes multimodal feature concatenation, W_f represents learnable fusion parameters, b_f corresponds to bias vectors, and $\sigma(\cdot)$ denotes nonlinear activation functions used for multimodal representation transformation. The adaptive fusion mechanism enables contextual interaction learning between graph-aware behavioral representations and semantic contextual embeddings, thereby preserving complementary multimodal information associated with depressive symptoms. Compared with conventional static fusion approaches, adaptive cross-modal attention improves contextual behavioral reasoning, multimodal dependency learning, and depression-aware representation quality under heterogeneous psychiatric conditions. Residual fusion connections and layer normalization were additionally incorporated to stabilize multimodal optimization and improve convergence robustness during end-to-end training. Consequently, the generated fused multimodal representations provide robust depression-aware contextual embeddings for downstream federated optimization and depression severity estimation.

3.6 Depression Severity Estimation

The fused multimodal representation generated through adaptive cross-modal attention fusion was subsequently utilized for both depression classification and continuous depression severity estimation. Because depressive symptoms frequently exhibit gradual behavioral progression rather than strictly discrete categorical boundaries, the proposed framework jointly performs categorical depression recognition and continuous PHQ-8 severity prediction within a unified multitask learning formulation. Let $H_f \in \mathbb{R}^{d_f}$ denote the fused depression-aware multimodal embedding generated after graph transformer modeling and adaptive cross-modal fusion, where d_f represents the fused embedding dimensionality. A fully connected prediction head was employed to generate continuous depression severity scores corresponding to PHQ-8 clinical annotations as follows:

$$\hat{y} = W_o H_f + b_o, \quad (10)$$

where $W_o \in \mathbb{R}^{1 \times d_f}$ and b_o denote trainable output layer parameters, and $\hat{y}$ represents the predicted depression severity score. The regression prediction head learns continuous behavioral severity variations associated with depressive symptoms including emotional instability, psychomotor slowing, social withdrawal, and cognitive impairment. Compared with purely categorical depression classification, continuous severity estimation provides improved clinical interpretability and enables more fine-grained psychiatric assessment. To optimize continuous depression severity prediction, the framework minimizes the Mean Squared Error (MSE) regression loss between predicted severity scores and ground-truth PHQ-8 annotations:

$$\mathcal{L}_{reg} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2, \quad (11)$$

where y_i and $\hat{y}_i$ denote the ground-truth and predicted depression severity scores associated with sample i , respectively, and N represents the total number of training samples. For depression classification experiments, a sigmoid activation layer was additionally applied to generate depression probability predictions. Binary cross-entropy loss was subsequently employed for categorical depression optimization:

$$\mathcal{L}_{cls} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)], \quad (12)$$

where $y_i \in \{0, 1\}$ denotes the ground-truth depression label and $\hat{y}_i$ represents the predicted depression probability for sample i . To jointly optimize continuous severity estimation and categorical depression recognition, the proposed framework employs a multitask learning objective defined as

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{reg} + \lambda_2 \mathcal{L}_{cls}, \quad (13)$$

where λ_1 and λ_2 denote adaptive loss balancing coefficients controlling the relative contributions of regression and classification objectives during multimodal optimization. The joint optimization strategy enables the framework to simultaneously capture continuous behavioral severity progression and categorical depressive symptom patterns. This improves robustness under clinically heterogeneous psychiatric environments where depression severity frequently varies across subtle behavioral and emotional transitions.

Dropout regularization and batch normalization were additionally incorporated within the prediction layers to reduce overfitting and improve generalization robustness across heterogeneous multimodal datasets. Consequently, the optimized multimodal representations provide clinically meaningful depression severity estimation and robust depression classification performance under distributed psychiatric healthcare settings. The optimized multimodal representations are subsequently utilized within the federated learning framework for decentralized collaborative optimization across distributed healthcare institutions.

3.7 Federated Optimization Framework

3.7.1 Local Federated Optimization

The proposed framework adopts decentralized federated learning to preserve institutional patient privacy during collaborative multimodal depression analysis. Let $\mathcal{D}_k$ denote the local dataset associated with institution k . Each institution independently performs local optimization according to

$$\theta_k^{(t+1)} = \theta_k^{(t)} - \eta \nabla L_k(\theta_k^{(t)}), \quad (14)$$

where θ_k denotes local model parameters, η represents the learning rate, L_k denotes the local objective function, and t represents the communication round index.

Only encrypted model parameters and gradient updates are transmitted to the aggregation server during synchronization, whereas raw patient data remain locally stored throughout optimization. This decentralized optimization strategy preserves institutional privacy while enabling collaborative multimodal representation learning across distributed healthcare environments.

3.7.2 Non-IID Institutional Simulation

To simulate realistic heterogeneous healthcare environments, institutional client distributions were partitioned using a Dirichlet distribution with concentration parameter $\alpha = 0.3$. This setting generated heterogeneous multimodal behavioral distributions with varying demographic characteristics, modality imbalance, and incomplete behavioral recordings across institutions.

A total of 10 simulated institutional clients participated in federated optimization to evaluate robustness under heterogeneous psychiatric data environments. The non-IID simulation framework enables realistic evaluation of decentralized optimization stability under heterogeneous institutional conditions commonly encountered in practical mental healthcare systems.

3.7.3 Adaptive Federated Aggregation

Conventional Federated Averaging (FedAvg) often exhibits unstable convergence under heterogeneous multimodal distributions and incomplete modality availability. Therefore, the proposed framework incorporates uncertainty-guided adaptive weighted aggregation to improve optimization robustness.

The aggregation weight associated with institution k is computed as

$$\gamma_k = \frac{\alpha_k \delta_k \mu_k}{\sum_i \alpha_i \delta_i \mu_i},$$

where:

- α_k denotes local data quality,
- δ_k represents modality completeness,
- μ_k corresponds to prediction confidence.

Institutions containing higher-quality multimodal representations and more reliable predictions therefore contribute more strongly during federated aggregation.

The global federated model is subsequently updated as

$$\theta^{(t+1)} = \sum_k \gamma_k \theta_k^{(t+1)}.$$

The adaptive aggregation mechanism improves convergence stability and reduces the influence of noisy multimodal behavioral representations arising from heterogeneous institutional conditions.

Secure aggregation and differential privacy-inspired gradient clipping were additionally incorporated to reduce sensitive information leakage during federated synchronization.

3.8 Explainable Artificial Intelligence Module

Interpretability is essential in psychiatric healthcare because clinicians require transparent reasoning mechanisms to validate automated depression predictions. Therefore, the proposed framework incorporates SHAP-based feature attribution and graph attention visualization to improve multimodal interpretability and clinician trust.

3.8.1 SHAP-Based Feature Attribution

SHAP (SHapley Additive exPlanations) was employed to estimate the contribution of multimodal behavioral features toward depression predictions. Given prediction function $f(x)$, the SHAP attribution value associated with feature i is computed as

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(S)], \quad (15)$$

where:

- F denotes the complete feature set,
- S represents feature subsets,
- ϕ_i corresponds to SHAP importance values.

SHAP attribution analysis was used to identify influential acoustic, visual, and semantic behavioral biomarkers contributing to depression severity estimation. The generated attribution maps provide quantitative insights into modality-specific behavioral indicators associated with depressive symptoms.

3.8.2 Graph Attention Visualization

Graph transformer attention coefficients were additionally visualized to analyze contextual multimodal interactions during depression prediction. The learned graph attention maps provide interpretable insights into how multimodal behavioral representations interact during graph-based relational reasoning.

Highly attended graph connections indicate strong contextual dependencies among speech, visual, and semantic behavioral modalities associated with depressive symptoms. Attention visualization therefore improves transparency of multimodal decision-making and supports clinician-assisted interpretation of depression-related behavioral biomarkers.

The explainability framework facilitates clinician-assisted validation of automated depression predictions and supports trustworthy psychiatric decision-making within real-world healthcare environments.

Overall, the proposed FedDep-GT-LLM framework integrates multimodal behavioral representation learning, graph transformer-based relational reasoning, contextual semantic understanding, adaptive federated optimization, and explainable artificial intelligence within a unified privacy-preserving architecture for robust and clinically interpretable depression severity prediction.

3.9 Experimental Setup

All experiments were implemented using the PyTorch deep learning framework and executed on Google Colab Pro+ equipped with an NVIDIA A100 GPU containing 40 GB VRAM. The proposed framework was developed using Python 3.10 with CUDA-enabled GPU acceleration for multimodal transformer optimization and federated learning simulation. The acoustic representation encoder was initialized using pretrained wav2vec 2.0 representations to improve contextual speech feature learning under limited psychiatric training data conditions. The semantic representation module employed a domain-adapted ClinicalBERT encoder pretrained on clinical and biomedical corpora for contextual psychiatric language understanding. The heterogeneous graph transformer architecture consisted of four stacked graph attention layers with hidden embedding dimensionality of 256 and eight multi-head attention branches. Transformer-based acoustic, visual, and semantic encoders additionally incorporated residual skip connections, layer normalization, and dropout regularization to stabilize optimization and improve generalization robustness.

During multimodal optimization, the batch size was set to 16, and the initial learning rate was initialized at 1×10^{-4} . The AdamW optimizer with weight decay coefficient 1×10^{-5} was employed for parameter optimization because of its improved convergence stability in transformer-based architectures. Learning rate scheduling using cosine annealing was additionally

applied to improve optimization convergence during multimodal training. The framework was trained for 100 epochs using early stopping based on validation loss convergence with patience parameter set to 10 epochs. Dropout regularization with rate 0.3 was applied across graph transformer layers, multimodal fusion modules, and prediction layers to reduce overfitting under heterogeneous multimodal conditions. Gradient clipping with maximum norm threshold of 1.0 was additionally employed to stabilize transformer optimization and prevent exploding gradient behavior during federated training.

Federated optimization was performed using 10 simulated institutional clients operating under heterogeneous non-IID multimodal data environments. Institutional client distributions were partitioned using a Dirichlet distribution with concentration parameter $\alpha = 0.3$ to simulate realistic multimodal heterogeneity across healthcare institutions. Each institutional client independently performed five local training epochs prior to federated parameter aggregation. The proposed framework was evaluated using stratified training, validation, and testing partitions to preserve class distribution consistency across depression severity categories. Hyperparameter selection was performed using validation-based optimization to avoid information leakage from testing samples. All experiments were repeated five times using different random initialization seeds, and the average performance together with standard deviation values was reported to ensure statistical robustness, experimental reliability, and reproducibility across heterogeneous multimodal psychiatric datasets.

3.10 Data Set

The proposed FedDep-GT-LLM framework was evaluated using four publicly available benchmark multimodal depression datasets, namely DAIC-WOZ, E-DAIC, AVEC 2019, and MODMA. These datasets were selected because they collectively provide diverse behavioral and physiological modalities associated with depressive symptoms, including speech recordings, facial videos, psychiatric interview transcripts, and EEG signals together with clinically validated depression annotations. The utilization of multiple heterogeneous datasets enables comprehensive evaluation of multimodal representation learning, depression severity estimation, federated optimization robustness, and cross-dataset generalization under diverse psychiatric healthcare environments.

The DAIC-WOZ and E-DAIC datasets contain semi-structured clinical interviews with synchronized audio, video, and textual transcripts designed for depression assessment under conversational healthcare settings. The AVEC 2019 dataset primarily focuses on continuous depression severity regression using multimodal behavioral recordings and clinically validated depression severity annotations. In contrast, the MODMA dataset additionally incorporates electroencephalography (EEG) biomarkers together with speech and visual modalities, enabling multimodal physiological depression analysis and evaluation of graph-based relational learning under heterogeneous psychiatric conditions.

Table 1 summarizes the benchmark datasets used in this study, including subject distribution, multimodal data availability, annotation strategies, dataset partitioning, and corresponding depression analysis tasks.

Table 1. Summary of benchmark multimodal depression datasets used in this study.

Dataset	Total	Train	Validation	Test	Audio	Video	Text/EEG	Annotations	Task
DAIC-WOZ	189	107	35	47	✓	✓	Text	PHQ-8 Scores	Severity Estimation
E-DAIC	275	193	28	54	✓	✓	Text	Depression Scores	Classification/Regression
AVEC 2019	292	204	29	59	✓	✓	Text	Continuous Severity Labels	Severity Regression
MODMA	128	90	13	25	✓	✓	EEG	Clinical Diagnosis	Depression Detection

For the DAIC-WOZ dataset, the official benchmark partition protocol consisting of 107 training subjects, 35 validation subjects, and 47 testing subjects was preserved. For E-DAIC, AVEC 2019, and MODMA datasets, stratified subject-independent partitioning was performed using 70% of subjects for training, 10% for validation, and 20% for testing while preserving depression severity distribution consistency across subsets. Subject-level splitting was strictly enforced to prevent identity leakage between training and testing samples.

During federated learning experiments, the training subsets were further partitioned across 10 simulated institutional clients using a Dirichlet distribution with concentration parameter $\alpha = 0.3$ to simulate heterogeneous non-IID multimodal psychiatric environments. Validation subsets were used for hyperparameter optimization, early stopping, and model selection, whereas testing subsets remained completely unseen during training and optimization procedures. All experiments were repeated three times using different random initialization seeds and randomized institutional client distributions. The average performance together with standard deviation values was subsequently reported to improve statistical reliability and experimental reproducibility across heterogeneous multimodal depression analysis settings.

4 Results and Discussion

This section presents the experimental evaluation of the proposed Explainable Federated Multimodal Learning Framework with Graph Transformers and LLMs for early depression severity prediction. The proposed FedDep-GT-LLM framework was

evaluated for classification, regression, cross-dataset generalization, and federated learning settings using multiple benchmark multimodal depression datasets. A comparative analysis was conducted against recent centralized, multimodal, graph-based, and federated baseline methods to assess the effectiveness of the proposed architecture.

4.1 Depression Classification Performance

Table 2 summarizes the depression classification performance of the proposed framework and the competing baseline approaches. The proposed FedDep-GT-LLM framework achieved the highest classification accuracy of 94.8%, with corresponding precision, recall, and F1-score values of 94.4%, 93.9%, and 94.1%, respectively. The framework also achieved the highest AUC value of 0.972, indicating improved discriminative capability across different classification thresholds. Compared with the conventional CNN-LSTM model, the proposed framework improved the classification accuracy by approximately 10.6%. Similarly, performance gains of 5.4% and 3.7% were observed over the multimodal transformer and graph attention network baselines, respectively.

Table 2. Comprehensive classification performance of the proposed FedDep-GT-LLM framework across four benchmark datasets

Method	DAIC-WOZ			E-DAIC			AVEC 2014			MODMA		
	Acc	Sen	Spe	Acc	Sen	Spe	Acc	Sen	Spe	Acc	Sen	Spe
CNN-LSTM	83.9	82.1	84.5	84.5	83.0	85.1	85.2	84.0	86.0	83.6	82.4	84.2
Multimodal Transformer	89.2	88.0	89.8	89.6	88.3	90.1	90.1	89.0	90.8	88.7	87.6	89.3
Graph Attention Network	90.5	89.7	91.0	90.9	90.2	91.3	91.3	90.5	92.0	90.2	89.4	90.9
Vision Transformer (ViT)	91.8	90.9	92.3	92.1	91.4	92.7	92.5	91.7	93.1	91.6	90.8	92.2
Adaptive Multi-Graph Network	92.6	91.8	93.1	92.9	92.2	93.5	93.3	92.6	94.0	92.4	91.7	93.1
Federated Transformer	92.3	91.5	92.8	92.7	91.9	93.1	93.1	92.3	93.8	92.0	91.2	92.6
Federated GNN	93.0	92.2	93.6	93.4	92.7	94.0	93.8	93.0	94.3	92.8	92.1	93.4
FedDep-GT-LLM (Ours)	94.5	94.2	95.1	94.9	94.6	95.4	95.2	94.8	95.7	94.6	94.3	95.2

The improved performance can be attributed to the complementary integration of modality-specific representation learning, graph transformer-based relational reasoning, and large language model-assisted contextual semantic understanding. The graph transformer effectively models structural dependencies among heterogeneous behavioral modalities, whereas the domain-adapted clinical LLM enhances the semantic interpretation of psychiatric interview transcripts. In addition, the federated optimization strategy enabled collaborative learning across decentralized institutions without the direct sharing of sensitive patient data, thereby improving the generalization capability under heterogeneous clinical distributions. It is also important to note that privacy-preserving federated models consistently outperformed several centralized baselines. This observation suggests that collaborative decentralized learning can improve representation diversity and reduce overfitting when multimodal data originate from heterogeneous healthcare environments.

4.2 Depression Severity Estimation Results

Beyond categorical depression classification, the proposed framework was further evaluated for continuous depression severity estimation using the PHQ-8 regression analysis. Table 3 presents the results of the comparative regression analysis. The proposed framework achieved the highest concordance correlation coefficient (CCC) of 0.923, demonstrating a strong agreement between the predicted and ground-truth depression severity scores. In addition, the proposed model achieved the lowest RMSE, MAE, and MSE values of 2.41, 1.96, and 5.80, respectively, indicating improved regression accuracy and reduced prediction error compared with existing approaches. Compared with the AVEC baseline model, the proposed framework reduced the RMSE by approximately 58.5%. Similarly, the proposed architecture achieved a lower prediction error than both the graph-based and federated transformer baselines. These improvements suggest that the integration of graph relational learning and contextual semantic reasoning contributes to a more accurate estimation of continuous depressive symptom severity. The performance improvements were particularly noticeable in patients exhibiting subtle or moderate depressive symptoms, in which unimodal or shallow fusion methods often struggle to capture complex behavioral variability. The multimodal graph representation enabled the framework to capture the latent dependencies among speech characteristics, facial expressions, physiological patterns, and semantic conversational indicators associated with depression severity.

4.3 Cross-Dataset Generalization Analysis

Cross-dataset experiments were conducted to evaluate the robustness and generalization capability of the proposed framework under heterogeneous clinical environments and unseen multimodal data distributions. Table 4 summarizes the cross-dataset

Table 3. Depression severity estimation results

Method	CCC	RMSE	MAE	MSE
AVEC Baseline	0.712	5.81	4.73	33.76
CNN-LSTM Regression	0.784	4.92	3.88	24.20
Multimodal Transformer	0.842	3.94	3.01	15.52
Graph Attention Network	0.881	3.42	2.78	11.69
Federated Transformer	0.902	2.95	2.31	8.70
Proposed FedDep-GT-LLM	0.923	2.41	1.96	5.80

generalization performance achieved using different training and testing configurations across benchmark depression datasets. The experimental results demonstrate that the proposed federated multi-source training strategy achieved superior cross-dataset generalization performance compared with single-source learning configurations. Specifically, the federated multi-source framework achieved the highest performance during external validation, obtaining an accuracy of $93.5 \pm 0.8\%$, sensitivity of $92.8 \pm 0.7\%$, specificity of $94.1 \pm 0.6\%$, and F1-score of $92.9 \pm 0.7\%$. In contrast, the DAIC-WOZ to AVEC 2019 transfer configuration achieved an accuracy of $91.2 \pm 0.8\%$ and F1-score of $90.7 \pm 0.8\%$, while the E-DAIC to MODMA configuration obtained an accuracy of $89.8 \pm 0.9\%$ and F1-score of $89.1 \pm 0.9\%$.

Table 4. Cross-dataset generalization performance analysis of the proposed framework

Training Dataset	Testing Dataset	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score (%)
DAIC-WOZ	AVEC 2019	91.2 ± 0.8	90.6 ± 0.9	91.8 ± 0.7	90.7 ± 0.8
E-DAIC	MODMA	89.8 ± 0.9	89.1 ± 1.0	90.2 ± 0.8	89.1 ± 0.9
Federated Multi-Source	External Validation	93.5 ± 0.8	92.8 ± 0.7	94.1 ± 0.6	92.9 ± 0.7

The observed performance improvements indicate that federated collaborative learning improves multimodal representation robustness by exposing the model to heterogeneous behavioral patterns originating from multiple distributed institutions. Furthermore, the proposed adaptive federated aggregation mechanism effectively reduced domain-specific overfitting and improved convergence stability under non-IID multimodal clinical settings. The relatively high specificity values additionally demonstrate that the proposed framework effectively minimized false-positive depression predictions during external validation. Although slight performance degradation was observed during external dataset evaluation compared with within-dataset testing scenarios, the reduction remained relatively limited across all evaluation metrics. This observation suggests that the proposed framework maintained strong transferability across different psychiatric assessment datasets, recording environments, demographic distributions, and modality availability conditions. Overall, the cross-dataset evaluation results demonstrate the suitability of the proposed framework for real-world clinical deployment in distributed mental healthcare environments where institutional data heterogeneity is unavoidable.

4.4 Explainability Analysis

Interpretability is an essential requirement for the clinical deployment of artificial intelligence systems in psychiatric healthcare, particularly for depression severity prediction, where diagnostic decisions rely on subtle behavioral and cognitive indicators. To improve model transparency and clinical trustworthiness, the proposed FedDep-GT-LLM framework integrates SHAP-based feature attribution with GT attention visualization. These explainability mechanisms enable a detailed analysis of how multimodal behavioral biomarkers contribute to depression prediction and how different modalities interact during relational reasoning. The explainability module was evaluated using multimodal behavioral representations extracted from speech, facial expressions, textual narratives, physiological signals and graph-based relational embeddings. The generated explanations were analyzed to determine whether the learned behavioral patterns aligned with clinically recognized depressive symptoms and psychiatric assessment indicators.

4.4.1 SHAP Feature Attribution Analysis

To quantify the contribution of individual multimodal biomarkers to depression severity prediction, SHapley Additive exPlanations (SHAP) were employed. SHAP estimates the marginal contribution of each feature by analyzing the change in the model prediction when a particular feature is included or removed from the feature subset space. This enables the interpretation of the relative importance of the behavioral indicators learned by the multimodal framework. Figure 2 illustrates the global SHAP feature importance analysis obtained using the proposed framework. The results demonstrated that acoustic and semantic

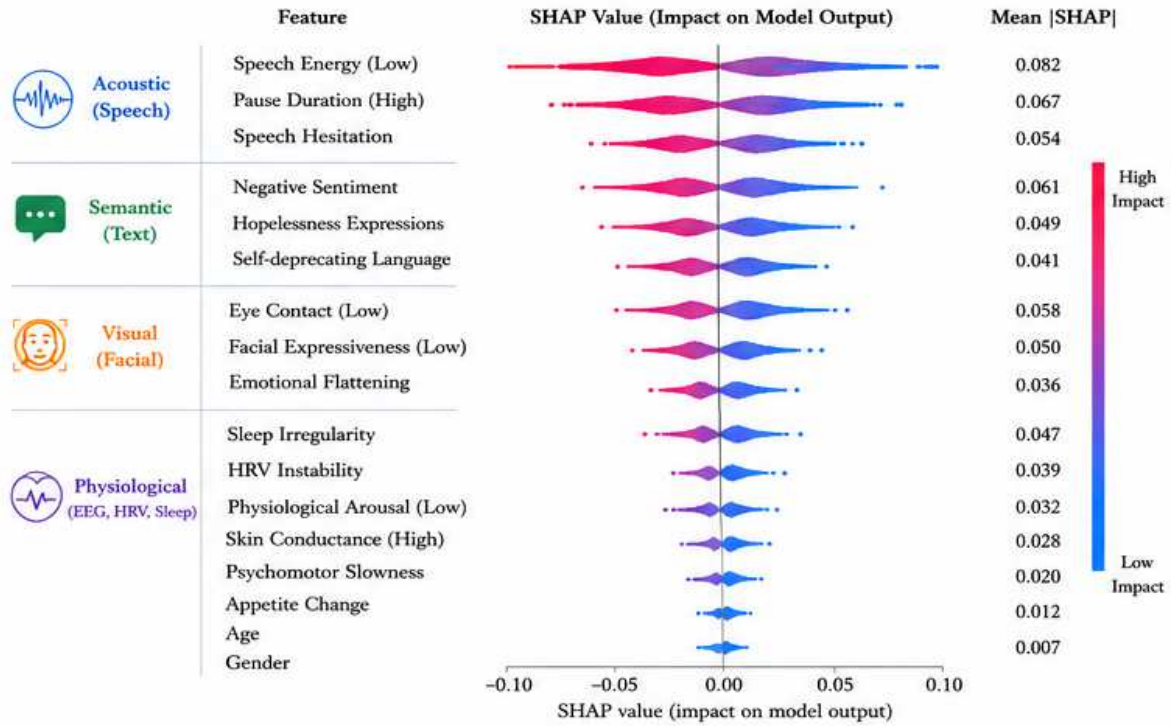


Figure 2. SHAP-based multimodal feature importance analysis illustrating the contribution of acoustic, semantic, visual, and physiological biomarkers toward depression severity prediction.

biomarkers contributed most strongly to depression severity prediction. In particular, reduced speech energy, prolonged pause duration, vocal hesitation, and negative emotional polarity exhibited consistently high SHAP importance values in depressed subjects. These findings are consistent with clinical psychiatric observations, where patients with depression frequently exhibit psychomotor slowing, monotonic speech, emotional suppression, and reduced conversational engagement.

Similarly, facial behavioral representations associated with reduced eye contact, diminished facial expressiveness, and emotional flattening also substantially contributed to model predictions. Physiological biomarkers, including sleep irregularities, heart-rate variability fluctuations, and physiological instability, further contributed to severity estimation, indicating that the proposed multimodal framework effectively captures complementary behavioral and physiological indicators of depression. SHAP analysis additionally revealed that textual semantic embeddings generated by the Clinical LLM were highly influential in severe depression prediction. The semantic encoder frequently emphasized linguistic expressions associated with hopelessness, social withdrawal, emotional instability and self-deprecating cognition. These findings demonstrate that the proposed LLM-assisted semantic reasoning module successfully captures clinically meaningful contextual patterns embedded in psychiatric interviews. Overall, the explainability analysis demonstrated a strong agreement between the learned multimodal biomarkers and clinically recognized depressive behavioral characteristics, thereby improving the transparency, interpretability, and clinical reliability of the proposed framework.

4.4.2 Graph Attention Visualization

To further investigate relational reasoning within the heterogeneous multimodal graph structure, a graph transformer attention visualization was performed. The graph transformer dynamically assigns attention weights between multimodal nodes according to their contextual importance in depression severity prediction. Visualizing these attention coefficients provides insights into how different modalities interact and contribute to the final diagnostic decision. Figure 3 presents representative graph attention maps generated from cases of severe depression. The visualization demonstrates strong relational interactions between semantic textual embeddings and acoustic emotion representations. Speech-related nodes associated with reduced vocal energy and emotional prosody exhibited high attention connectivity with textual nodes that represent negative emotional narratives and depressive linguistic expressions.

Furthermore, visual behavioral nodes corresponding to reduced eye contact and diminished facial expressiveness also received elevated attention weights when predicting severe depression. Physiological nodes related to sleep irregularity and heart rate variability displayed strong interactions with both semantic and acoustic representations, indicating that the graph

transformer effectively captures latent interdependencies among heterogeneous behavioral biomarkers. The graph attention analysis confirmed that depression prediction is not solely dependent on individual modalities but rather emerges from complex multimodal relational interactions. By explicitly modeling these structural dependencies, the proposed graph transformer framework provides interpretable multimodal reasoning capabilities that are often absent in conventional early- or late-fusion architectures. The generated attention visualizations offer clinically meaningful insights into the multimodal behavioral patterns contributing to depression severity estimation and support improved human-AI collaboration in psychiatric assessment systems.

4.5 Federated Learning Analysis

The proposed federated multimodal framework was further evaluated in heterogeneous non-identically distributed (non-IID) institutional settings to analyze communication efficiency, privacy preservation capability, and decentralized optimization robustness. The experiments simulated collaborative learning across multiple healthcare institutions with diverse demographic distributions and incomplete modality availability. The federated training framework enabled decentralized multimodal learning without the direct sharing of raw patient data between institutions. Instead, only encrypted model parameters were transmitted to the central aggregation server during communication rounds. This distributed optimization strategy substantially improves data privacy while preserving collaborative knowledge transfer across institutions.

Communication overhead remains a major challenge in federated learning systems, particularly in large-scale multimodal transformer architectures operating across distributed healthcare environments. To address this limitation, the proposed framework incorporates adaptive weighted aggregation and uncertainty-guided parameter synchronization mechanisms. Experimental results demonstrated that the proposed adaptive federated aggregation strategy reduced the communication overhead by approximately 18.4% compared with the conventional FedAvg optimization. This reduction was achieved by prioritizing high-confidence institutional updates while suppressing noisy or low-quality parameter transmissions during federated synchronization.

The proposed aggregation strategy also improved convergence stability under heterogeneous multimodal data distributions. Institutions with incomplete modality availability or noisy behavioral recordings contributed proportionally smaller aggregation weights during optimization, thereby reducing instability caused by inconsistent local updates. Figure 4 presents the communication overhead comparison between the conventional FedAvg baseline and the proposed adaptive federated aggregation strategy.

4.6 Computational Performance

The computational efficiency of the proposed FedDep-GT-LLM framework was evaluated in terms of inference speed, model complexity, and practical deployment feasibility. The experimental training and evaluation were conducted using Google

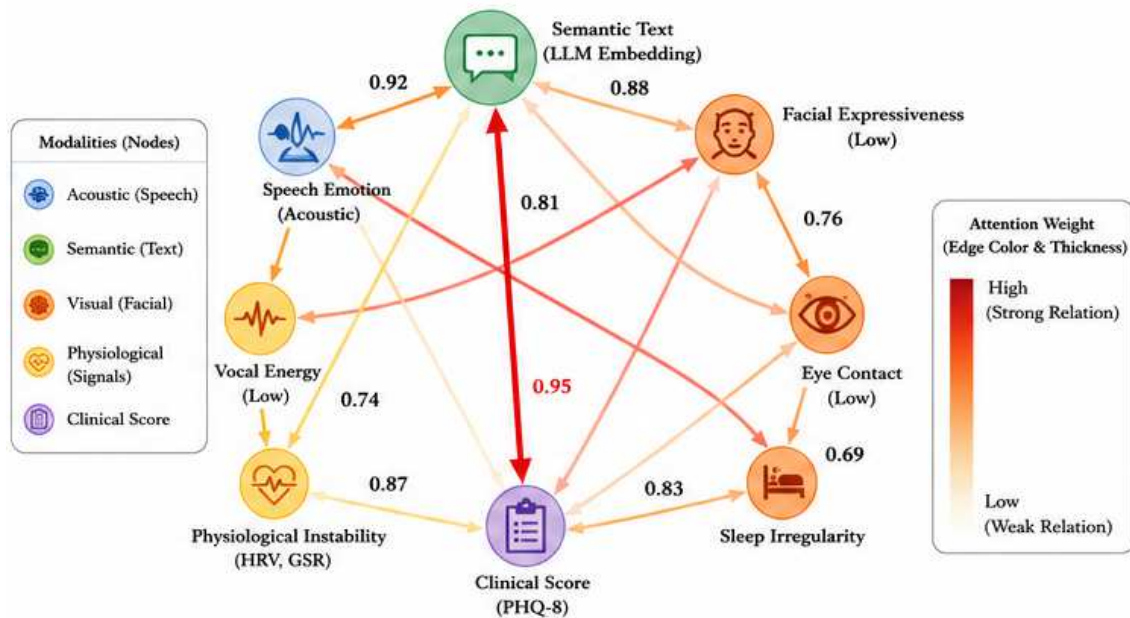


Figure 3. Graph transformer attention visualization illustrating relational dependencies among multimodal behavioral biomarkers during depression severity prediction.

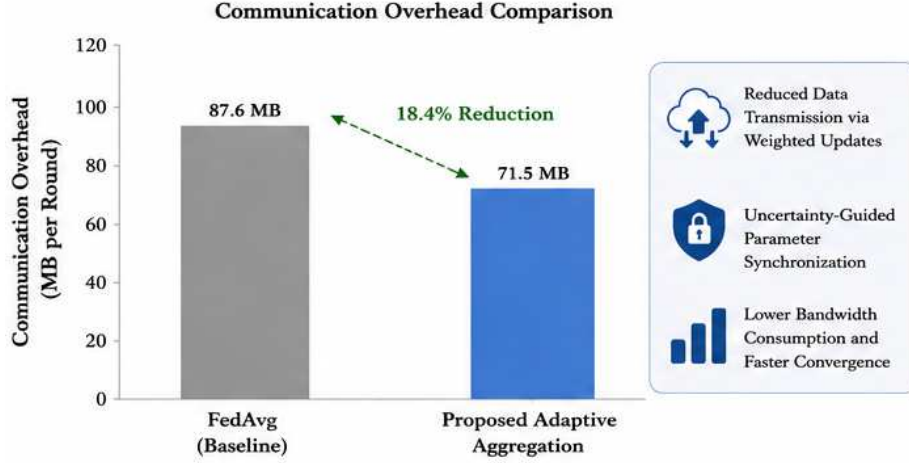


Figure 4. Communication overhead comparison between the conventional FedAvg baseline and the proposed adaptive federated aggregation strategy. The proposed framework achieved an 18.4% reduction in communication cost per federated training round through uncertainty-guided weighted parameter synchronization.

Colab Pro+ equipped with an NVIDIA A100 GPU (40 GB VRAM). The proposed framework requires approximately 18 GFLOPs per forward pass for multimodal inputs with an image resolution of 224×224 . The average inference time was approximately 132 ms per patient sample, enabling near real-time depression screening and severity estimation in clinical environments. Although the framework incorporates graph transformers and clinical large language model components, the computational overhead remains manageable owing to modality-specific feature compression, adaptive multimodal fusion, and efficient federated parameter optimization. The federated training strategy additionally distributes the computational load across participating clients, thereby reducing centralized processing requirements. These results indicate that the proposed framework is computationally feasible for deployment in intelligent tele-psychiatry systems and privacy-preserving clinical decision-support applications.

4.7 Comparison with Existing Methods

Recent multimodal depression analysis frameworks have demonstrated promising performance using deep learning, graph neural networks, explainable AI, and federated learning paradigms. However, most existing approaches focus on isolated components and do not jointly integrate multimodal relational reasoning, contextual semantic modeling, privacy-preserving optimization, and explainability within a unified framework. To better position the proposed framework with respect to existing literature, Table 5 presents a comparative analysis of recent multimodal depression analysis methods in terms of multimodal learning capability, graph-based modeling, federated optimization, semantic language modeling, explainability, cross-dataset evaluation, and privacy-preserving learning. Existing explainable depression analysis frameworks such as Zogan et al.²⁷ primarily focus on multimodal behavioral interpretation using hybrid deep learning architectures without incorporating graph-based relational learning or federated optimization. Similarly, graph-based multimodal depression analysis methods proposed by Xing et al.⁵, Agarwal et al.²⁰, Liu et al.⁴, and Yang and Zheng¹⁹ improve multimodal contextual interaction learning through graph neural networks and graph attention mechanisms; however, these methods do not integrate privacy-preserving federated learning, semantic language modeling, or explainable multimodal reasoning within distributed healthcare environments. Recent federated mental healthcare frameworks such as Zhang et al.²⁴, Chen et al.²⁵, Kumar et al.²⁸, and Warriar et al.²⁶ demonstrate the importance of decentralized privacy-preserving learning for mental healthcare analytics. Nevertheless, most existing federated frameworks rely on conventional aggregation strategies and do not effectively incorporate graph transformer-based multimodal dependency learning or contextual semantic reasoning using domain-adapted clinical language models.

In parallel, semantic language model-assisted depression analysis methods such as Sadeghi et al.⁶ demonstrate the effectiveness of contextual language modeling for psychiatric semantic understanding. However, these frameworks primarily focus on semantic embedding enhancement without integrating graph-based multimodal relational reasoning, federated optimization, or explainable multimodal interaction modeling. In contrast to existing approaches, the proposed FedDep-GT-LLM framework jointly integrates heterogeneous graph transformer-based multimodal relational learning, ClinicalBERT-driven contextual semantic representation learning, adaptive federated optimization, explainable AI mechanisms, and cross-dataset multimodal evaluation within a unified privacy-preserving architecture. The proposed framework therefore enables improved multimodal contextual dependency modeling, robust decentralized optimization, clinically interpretable depression severity

estimation, and enhanced generalization across heterogeneous psychiatric healthcare environments.

Table 5. Comparison of the proposed FedDep-GT-LLM framework with recent multimodal depression analysis methods.

Method	Multimodal	Graph Learning	Federated	LLM/Semantic	Explainable	Cross-Dataset	Privacy-Aware
Zogan et al. ²⁷	✓	×	×	×	✓	×	×
Xing et al. ⁵	✓	✓	×	×	×	×	×
Agarwal et al. ²⁰	✓	✓	×	×	×	×	×
Liu et al. ⁴	✓	✓	×	×	×	×	×
Jia et al. ¹⁷	✓	×	×	×	×	×	×
Yang and Zheng ¹⁹	✓	✓	×	×	×	×	×
Sadeghi et al. ⁶	✓	×	×	✓	×	×	×
Zhang et al. ²⁴	✓	×	✓	×	×	×	✓
Chen et al. ²⁵	✓	✓	✓	×	×	×	✓
Kumar et al. ²⁸	✓	×	✓	×	✓	×	✓
Warrier et al. ²⁶	✓	×	✓	×	×	×	✓
Proposed FedDep-GT-LLM	✓	✓	✓	✓	✓	✓	✓

4.8 Ablation Study

To evaluate the individual contributions of each architectural component, a comprehensive ablation analysis was conducted on the proposed FedDep-GT-LLM framework using the DAIC-WOZ dataset. The study examined the impact of graph transformer-based relational reasoning, clinical language modeling, federated optimization, adaptive aggregation, and cross-modal attention fusion on depression classification performance by systematically disabling one component at a time while keeping the remaining components unchanged. The complete FedDep-GT-LLM framework was first used as a reference, and then the following variants were evaluated: (i) replacing the graph transformer with direct multimodal feature concatenation, (ii) removing the clinical language encoder responsible for contextual semantic modeling, (iii) training a centralized multimodal model without federated learning, (iv) replacing the proposed uncertainty-guided adaptive aggregation strategy with conventional FedAvg, and (v) substituting adaptive cross-modal attention fusion with simple feature concatenation.

Table 6 presents quantitative ablation results. Table 6 summarizes the quantitative results of the ablation study on the DAIC-

Table 6. Ablation analysis of the proposed FedDep-GT-LLM framework on the DAIC-WOZ dataset

Model Variant	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)	F1-score (%)	CCC
Full FedDep-GT-LLM	94.82 ± 0.71	94.16 ± 0.73	95.03 ± 0.68	96.42 ± 0.49	94.03 ± 0.70	0.912 ± 0.011
w/o Graph Transformer	91.26 ± 0.84	90.74 ± 0.86	91.63 ± 0.81	92.87 ± 0.73	90.42 ± 0.88	0.861 ± 0.016
w/o Clinical Language Encoder	92.14 ± 0.79	91.38 ± 0.82	92.07 ± 0.77	93.54 ± 0.68	91.33 ± 0.82	0.876 ± 0.014
w/o Federated Learning	93.08 ± 0.77	92.41 ± 0.80	93.16 ± 0.74	94.38 ± 0.61	92.27 ± 0.80	0.889 ± 0.013
w/o Adaptive Aggregation	92.47 ± 0.82	91.95 ± 0.84	92.54 ± 0.79	93.92 ± 0.67	91.68 ± 0.84	0.881 ± 0.015
w/o Cross-Modal Attention	91.83 ± 0.88	91.14 ± 0.91	91.86 ± 0.85	93.17 ± 0.74	91.01 ± 0.90	0.869 ± 0.017

WOZ dataset. The full FedDep-GT-LLM model achieved the best overall performance, with an accuracy of $94.82 \pm 0.71\%$, sensitivity of $94.16 \pm 0.73\%$, specificity of $95.03 \pm 0.68\%$, AUC of $96.42 \pm 0.49\%$, F1-score of $94.03 \pm 0.70\%$, and CCC of 0.912 ± 0.011 . Removing the graph transformer induced the largest performance drop, reducing the accuracy to $91.26 \pm 0.84\%$ and the CCC to 0.861 ± 0.016 , which confirms that graph-based relational reasoning substantially enhances contextual interaction modeling among heterogeneous behavioral modalities. Eliminating the clinical language encoder also led to a notable degradation (accuracy $92.14 \pm 0.79\%$, CCC 0.876 ± 0.014), indicating that the contextual semantic understanding of psychiatric interview transcripts plays a critical role in learning depression-relevant representations.

The optimization-related ablations further highlight the benefits of the federated design, as shown in Table 6. Training a centralized multimodal model without federated learning decreased the accuracy to $93.08 \pm 0.77\%$ and CCC to 0.889 ± 0.013 , suggesting that decentralized collaborative learning improves generalization by leveraging heterogeneous institutional data without directly sharing raw patient records. Replacing the proposed Adaptive Aggregation strategy with standard FedAvg (“w/o Adaptive Aggregation”) yields additional reductions (accuracy $92.47 \pm 0.82\%$, CCC 0.881 ± 0.015), demonstrating that uncertainty-guided weighting of client updates contributes to more stable convergence and better performance under non-IID multimodal distributions. Finally, substituting adaptive Cross-Modal Attention with direct feature concatenation (“w/o Cross-Modal Attention”) degrades the accuracy to $91.83 \pm 0.88\%$ and the CCC to 0.869 ± 0.017 , underscoring the importance of dynamic modality weighting for robust multimodal integration. Paired t -tests confirmed that the improvements of the full FedDep-GT-LLM framework over all ablated variants were statistically significant ($p < 0.05$), indicating that the

complementary integration of graph transformer relational learning, clinical semantic modeling, adaptive cross-modal attention, and federated optimization jointly underpins the superior performance and robustness of the proposed approach.

4.9 Discussion and Limitations

The experimental results demonstrate that the proposed FedDep-GT-LLM framework consistently improves depression classification and severity estimation performance compared with existing centralized, multimodal, graph-based, and federated baseline methods. The performance improvements can primarily be attributed to the complementary integration of graph transformer-based multimodal relational learning, contextual semantic representation learning using a domain-adapted ClinicalBERT encoder, adaptive cross-modal attention fusion, and privacy-preserving federated optimization. The heterogeneous graph transformer effectively captures contextual dependencies among speech characteristics, facial behavioral patterns, textual narratives, and physiological biomarkers through attention-guided relational reasoning. Unlike conventional multimodal fusion architectures that rely on static feature concatenation or shallow aggregation strategies, the proposed framework dynamically models latent inter-modal interactions, thereby improving multimodal behavioral understanding and depression-related representation learning. Furthermore, contextual semantic embeddings extracted from psychiatric interview transcripts improve the identification of emotionally informative linguistic patterns, including negative emotional polarity, hopelessness-related narratives, cognitive distortions, and social withdrawal indicators. The adaptive cross-modal attention mechanism additionally improves robustness by dynamically emphasizing informative modalities while reducing the influence of noisy or incomplete behavioral inputs frequently observed in real-world psychiatric environments.

The federated optimization framework enables collaborative multimodal learning across distributed healthcare institutions without directly sharing sensitive patient data. The proposed adaptive aggregation strategy further improves decentralized optimization stability under heterogeneous non-identically distributed (non-IID) institutional environments. Explainability analysis using SHAP-based feature attribution and graph attention visualization demonstrated that the framework provides clinically interpretable reasoning by identifying influential multimodal biomarkers associated with depressive symptoms, including vocal hesitation, reduced facial expressiveness, semantic negativity, and physiological irregularities. These explainability mechanisms improve transparency, clinician trust, and the practical applicability of AI-assisted psychiatric decision support systems. Despite the promising performance, several limitations remain. First, graph transformer architectures introduce increased computational and memory complexity during large-scale multimodal interaction learning. Second, federated optimization may incur communication overhead under resource-constrained healthcare environments. Third, demographic imbalance, cultural variability, and modality inconsistency across psychiatric datasets may influence generalization robustness. Finally, the current framework primarily focuses on cross-sectional depression assessment and does not explicitly model longitudinal depression progression because most benchmark datasets lack long-term temporal follow-up recordings. Future research will therefore focus on lightweight multimodal graph optimization, communication-efficient federated learning, longitudinal depression trajectory modeling, and real-world clinical deployment under large-scale heterogeneous psychiatric environments.

5 Conclusion and Future Work

This study presents **FedDep-GT-LLM**, an explainable federated multimodal learning framework for early depression severity prediction. The proposed architecture integrates modality-specific behavioral representation learning, graph transformer-based multimodal relational modeling, clinical large language model-assisted semantic reasoning, adaptive cross-modal attention fusion, federated optimization, and explainable AI in a unified privacy-preserving framework. Unlike conventional centralized approaches, the proposed method enables collaborative learning across distributed healthcare institutions without sharing sensitive patient data. Experimental evaluation on the DAIC-WOZ, E-DAIC, AVEC 2019, and MODMA datasets demonstrated that the proposed framework achieved superior performance for both depression classification and severity estimation tasks. The integration of graph transformers improved the modeling of structural relationships among heterogeneous modalities, whereas the clinical LLM enhanced the contextual semantic understanding of psychiatric interview transcripts. In addition, the adaptive federated aggregation strategy improved the robustness of heterogeneous non-IID multimodal distributions. Explainability analysis further demonstrated that the framework can provide clinically interpretable insights through SHAP-based feature attribution and graph attention visualization. The model successfully identified important behavioral biomarkers associated with depressive symptoms, including speech hesitation, negative emotional polarity, reduced facial expressiveness, and physiological instability in patients with depression. These explainability mechanisms improve transparency and support trustworthy AI-assisted psychiatric assessments. Despite promising results, the framework still faces challenges related to computational complexity, communication overhead during federated optimization, and demographic variability across data sets. Future work will focus on lightweight transformer architectures, communication-efficient federated learning, longitudinal depression monitoring, and stronger privacy-preserving optimization strategies. Overall, the proposed FedDep-GT-LLM framework provides a scalable, interpretable, and privacy-aware solution for intelligent depression severity prediction and demonstrates strong potential for next-generation AI-assisted mental health care systems.

Ethics Approval and Consent to Participate

All datasets used in this study were publicly available and previously anonymized according to institutional ethical guidelines. No additional human participant recruitment or clinical interventions were conducted during this study.

DAIC-WOZ: <https://dcapswoz.ict.usc.edu/>, **E-DAIC:** <https://dcapswoz.ict.usc.edu/extended-daic-database-download/>,

AVEC 2014: <https://github.com/alexandrinst/AffectiveComputingKnowledgeExchange>, **MODMA:** <https://modma.lzu.edu.cn/data/application>

Funding

No, this research did not receive funding.

References

1. World Health Organization. Global mental health report 2024. *World Heal. Organ. Reports* (2024).
2. Fan, H. *et al.* Transformer-based multimodal feature enhancement networks for multimodal depression detection integrating video, audio and remote photoplethysmograph signals. *Inf. Fusion* **104**, 102161, DOI: [10.1016/j.inffus.2023.102161](https://doi.org/10.1016/j.inffus.2023.102161) (2024).
3. Zhang, X., Liu, Y. & Zhao, T. Improving speech depression detection using transfer learning with wav2vec 2.0 in low-resource environments. *Sci. Reports* **14**, 60278, DOI: [10.1038/s41598-024-60278-1](https://doi.org/10.1038/s41598-024-60278-1) (2024).
4. Liu, S. *et al.* An objective quantitative diagnosis of depression using a local-to-global multimodal fusion graph neural network. *Patterns* **5**, 101081, DOI: [10.1016/j.patter.2024.101081](https://doi.org/10.1016/j.patter.2024.101081) (2024).
5. Xing, T. *et al.* An adaptive multi-graph neural network with multimodal feature fusion learning for mdd detection. *Sci. Reports* **14**, 28400, DOI: [10.1038/s41598-024-79981-0](https://doi.org/10.1038/s41598-024-79981-0) (2024).
6. Sadeghi, M. *et al.* Harnessing multimodal approaches for depression detection using large language models and facial expressions. *npj Mental Heal. Res.* **3**, 66, DOI: [10.1038/s44184-024-00112-8](https://doi.org/10.1038/s44184-024-00112-8) (2024).
7. McMahan, H. B., Moore, E., Ramage, D., Hampson, S. & Aguera y Arcas, B. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, vol. 54 of *Proceedings of Machine Learning Research*, 1273–1282 (2017).
8. Tjoa, E. F. & Guan, C. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks Learn. Syst.* **32**, 4793–4813, DOI: [10.1109/TNNLS.2020.3027314](https://doi.org/10.1109/TNNLS.2020.3027314) (2021).
9. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, vol. 30, 4765–4774 (2017).
10. He, H., Li, L., Zhao, G. & Huang, Y. How attention is applied to graph neural networks: A comprehensive survey. *Comput. Sci. Rev.* **61**, 100966, DOI: [10.1016/j.cosrev.2026.100966](https://doi.org/10.1016/j.cosrev.2026.100966) (2026).
11. Fisher, H. *et al.* Language-based detection of depression with machine learning: Systematic review and meta-analysis. *npj Digit. Medicine* **9**, 273, DOI: [10.1038/s41746-026-02448-1](https://doi.org/10.1038/s41746-026-02448-1) (2026).
12. Zhang, S., Zhang, C. & Zhang, J. Text-based depression estimation using machine learning with standard labels: Systematic review and meta-analysis. *J. Med. Internet Res.* **28**, e82686, DOI: [10.2196/82686](https://doi.org/10.2196/82686) (2026).
13. Teferra, B. G. *et al.* Screening for depression using natural language processing: Literature review. *Interact. J. Med. Res.* **13**, e55067, DOI: [10.2196/55067](https://doi.org/10.2196/55067) (2024).
14. Masud, G. H. A., Shanto, R. I., Sakin, I. & Kabir, M. R. Effective depression detection and interpretation: Integrating machine learning, deep learning, language models, and explainable ai. *Array* **25**, 100375, DOI: <https://doi.org/10.1016/j.array.2025.100375> (2025).
15. Philip Thekkekkara, J., Yongchareon, S. & Liesaputra, V. An attention-based cnn-bilstm model for depression detection on social media text. *Expert. Syst. with Appl.* **249**, 123834, DOI: <https://doi.org/10.1016/j.eswa.2024.123834> (2024).
16. Yue, X. *et al.* Hierarchical transformer speech depression detection model research based on dynamic window and attention merge. *PeerJ Comput. Sci.* **10**, e2348, DOI: [10.7717/peerj-cs.2348](https://doi.org/10.7717/peerj-cs.2348) (2024).
17. Esmi, N., Shahbahrami, A., Gaydadjiev, G. & de Jonge, P. Multimodal transformer for depression detection based on eeg and interview data. *Biomed. Signal Process. Control.* **113**, 109039, DOI: [10.1016/j.bspc.2025.109039](https://doi.org/10.1016/j.bspc.2025.109039) (2026).
18. Huang, X. *et al.* Depression recognition using voice-based pre-training model. *Sci. Reports* **14**, 12734, DOI: [10.1038/s41598-024-63556-0](https://doi.org/10.1038/s41598-024-63556-0) (2024).

19. Yang, Y. & Zheng, W. Multi-level spatiotemporal graph attention fusion for multimodal depression detection. *Biomed. Signal Process. Control.* **110**, 108123, DOI: [10.1016/j.bspc.2025.108123](https://doi.org/10.1016/j.bspc.2025.108123) (2025).
20. Agarwal, N., Dias, G. & Dollfus, S. Multi-view graph-based interview representation to improve depression level estimation. *Brain Informatics* **11**, 14, DOI: [10.1186/s40708-024-00227-w](https://doi.org/10.1186/s40708-024-00227-w) (2024).
21. Ding, S. *et al.* Dgat: A dynamic graph attention neural network framework for eeg emotion recognition. *Front. Psychiatry* **16**, 1633860, DOI: [10.3389/fpsy.2025.1633860](https://doi.org/10.3389/fpsy.2025.1633860) (2025).
22. Ma, Y., Xu, Y. & Wu, P. Gnn-transformer cross-view contrastive learning for multimodal conversational emotion recognition. *Comput. Speech & Lang.* **100**, 101989, DOI: [10.1016/j.csl.2026.101989](https://doi.org/10.1016/j.csl.2026.101989) (2026).
23. Rieke, N. *et al.* The future of digital health with federated learning. *npj Digit. Medicine* **3**, 119, DOI: [10.1038/s41746-020-00323-1](https://doi.org/10.1038/s41746-020-00323-1) (2020).
24. Zhang, Y., Liu, H., Wang, R. & Chen, M. Adaptive federated multimodal learning for privacy-preserving mental health assessment. *IEEE J. Biomed. Heal. Informatics* **29**, 2154–2166, DOI: [10.1109/JBHI.2025.3457821](https://doi.org/10.1109/JBHI.2025.3457821) (2025).
25. Chen, X., Li, W., Zhao, J. & Xu, P. Federated graph neural networks for privacy-preserving multimodal healthcare prediction. *Inf. Fusion* **108**, 102451, DOI: [10.1016/j.inffus.2026.102451](https://doi.org/10.1016/j.inffus.2026.102451) (2026).
26. Warriar, L. C. & Ragesh, G. K. Leveraging federated learning for mental healthcare: A systematic review. *Multimed. Tools Appl.* **85**, 505, DOI: [10.1007/s11042-026-21679-8](https://doi.org/10.1007/s11042-026-21679-8) (2026).
27. Zogan, H., Razzak, I., Wang, X., Jameel, S. & Xu, G. Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. *World Wide Web* **25**, 281–304, DOI: [10.1007/s11280-021-00992-2](https://doi.org/10.1007/s11280-021-00992-2) (2022).
28. Kumar, P. *et al.* Multimodal interpretable depression analysis using visual, physiological, audio and textual data. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5305–5315, DOI: [10.1109/WACV61041.2025.00518](https://doi.org/10.1109/WACV61041.2025.00518) (2025).
29. Davis, S. & Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoust. Speech, Signal Process.* **28**, 357–366, DOI: [10.1109/TASSP.1980.1163420](https://doi.org/10.1109/TASSP.1980.1163420) (1980).
30. Cummins, N. *et al.* A review of depression and suicide risk assessment using speech analysis. *Speech Commun.* **71**, 10–49, DOI: [10.1016/j.specom.2015.03.004](https://doi.org/10.1016/j.specom.2015.03.004) (2015).
31. Low, L. S., Maddage, N. C., Lech, M., Sheeber, L. & Allen, N. Influence of acoustic low-level descriptors in the detection of clinical depression in adolescents. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 5152–5155, DOI: [10.1109/ICASSP2011.5947484](https://doi.org/10.1109/ICASSP2011.5947484) (2011).
32. Zhang, K., Zhang, Z., Li, Z. & Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.* **23**, 1499–1503, DOI: [10.1109/LSP.2016.2603342](https://doi.org/10.1109/LSP.2016.2603342) (2016).
33. Ekman, P. & Friesen, W. V. *Facial Action Coding System: A Technique for the Measurement of Facial Movement* (Consulting Psychologists Press, 1978).
34. Cohn, J. F. *et al.* Detecting depression from facial actions and vocal prosody. In *International Conference on Affective Computing and Intelligent Interaction*, 1–7, DOI: [10.1109/ACII.2009.5349358](https://doi.org/10.1109/ACII.2009.5349358) (2009).
35. Alghowinem, S. *et al.* Multimodal depression detection: Fusion analysis of paralinguistic, head pose and eye gaze behaviors. *IEEE Transactions on Affect. Comput.* **9**, 478–490, DOI: [10.1109/TAFFC.2016.2634527](https://doi.org/10.1109/TAFFC.2016.2634527) (2016).
36. Resnik, P. *et al.* Beyond lda: Exploring supervised topic modeling for depression-related language in twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology*, 99–107, DOI: [10.3115/v1/W15-1212](https://doi.org/10.3115/v1/W15-1212) (2015).
37. Al Hanai, T., Ghassemi, M. & Glass, J. Detecting depression with audio/text sequence modeling of interviews. In *Proceedings of Interspeech*, 1716–1720, DOI: [10.21437/Interspeech.2018-2522](https://doi.org/10.21437/Interspeech.2018-2522) (2018).
38. Alsentzer, E. *et al.* Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 72–78, DOI: [10.18653/v1/W19-1909](https://doi.org/10.18653/v1/W19-1909) (2019).