Depression Detection in Clinical Interviews with LLM-Empowered Structural Element Graph

Zhuang Chen1  Jiawen Deng2  Jinfeng Zhou1  Jincenzi Wu1  Tiejun Qian3  Minlie Huang1,*

1CoAI Group, DCST, TUPM, IAI, BNRIST, Tsinghua University
2University of Electronic Science and Technology of China
3Wuhan University

zhchen-nlp@mail.tsinghua.edu.cn aihuang@tsinghua.edu.cn

Abstract

Depression is a widespread mental health disorder affecting millions globally. Clinical interviews are the gold standard for assessing depression, but they heavily rely on scarce professional clinicians, highlighting the need for automated detection systems. However, existing methods only capture part of the relevant elements in clinical interviews, unable to incorporate all depressive cues. Moreover, the scarcity of participant data, due to privacy concerns and collection challenges, intrinsically constrains interview modeling. To address these limitations, in this paper, we propose a structural element graph (SEGA), which transforms the clinical interview into an expertise-inspired directed acyclic graph for comprehensive modeling. Additionally, we further empower SEGA by devising novel principle-guided data augmentation with large language models (LLMs) to supplement high-quality synthetic data and enable graph contrastive learning. Extensive evaluations on two real-world clinical datasets, in both English and Chinese, show that SEGA significantly outperforms baseline methods and powerful LLMs like GPT-3.5 and GPT-4.

1 Introduction

Depression is a pervasive mental health disorder that affects a significant portion of the global population. According to the World Health Organization (WHO)1, over 300 million people suffer from depression, casting a profound shadow on individuals, families, and society. Furthermore, in many communities, limited awareness and stigmatization surrounding mental health can lead to under-diagnosis and under-treatment, which underscores the imperative for effective depression screening.

To date, clinical interviews remain the standard method for assessing depression severity (He et al., 2016). However, analyzing clinical interviews has become relatively feasible, by either an “Wizard-of-Oz” virtual interviewer (Kelley, 1983) or a fully autonomous scripted agent (DeVault et al., 2014). However, analyzing clinical interviews for depression assessment remains a domain reserved for highly specialized clinicians.

As an alternative, automatic assessment of clinical interviews has emerged as a promising assistance for depression detection (Valstar et al., 2016). Clinical interviews mainly encapsulate four types of elements: the interview questions2, and the answer transcripts/videos, enabling super-

1https://www.who.int/news-room/fact-sheets/detail/depression

2Denote the text of questions. The video and audio of the interviewer are usually tool-synthesized without information.

Figure 1: A clinical interview for assessing depression. (2022). In Figure 1, we present an example interview from DAIC-WOZ (Gratch et al., 2014), a widely-used corpus for clinical depression detection. As can be seen, clinical interviews typically take the form of semi-structured multi-round question-answering dialogues. The interviewer follows an outline while flexibly posing questions relevant to personal experience and mental states, and the participant answers are recorded to assess the state of depression. Currently, conducting clinical interviews has become relatively feasible, by either an “Wizard-of-Oz” virtual interviewer (Kelley, 1983) or a fully autonomous scripted agent (DeVault et al., 2014). However, analyzing clinical interviews for depression assessment remains a domain reserved for highly specialized clinicians.

As an alternative, automatic assessment of clinical interviews has emerged as a promising assistance for depression detection (Valstar et al., 2016). Clinical interviews mainly encapsulate four types of elements: the interview questions and the answer transcripts/videos, enabling super-
Method | Elements in Clinical Interviews
--- | ---
Burdisso et al. (2023) | T V Q
Fan et al. (2022) | T A V Q
Shen et al. (2022) | T A V Q
Yoon et al. (2022) | T A V Q
Fang et al. (2022) | T A V Q
Niu et al. (2021) | T A V Q

SEGA (Ours) | T A V Q

Table 1: A glance at latest detection methods in modeling the transcript T, audio A, video V, and question Q.

To address the above limitations, we propose an LLM-empowered Structural Element GrAph (SEGA) for comprehensive interview modeling under limited resources. 1) To make full use of all four types of elements, we transform the clinical interview into SEGA, a directed acyclic graph. We first gather expertise about processing interview elements in depression assessment, then align SEGA’s structure with this human experience to facilitate representation learning. Specifically, within each round, we define the information flow between elements with noise suppression, allowing for element interaction while avoiding interference. Across multiple rounds, we specify simple yet efficient information propagation and aggregation approaches, so as to preserve the complete element features for predicting depressive status. 2) To alleviate the negative impact of data scarcity on inter-view modeling, we introduce a novel LLM-based data augmentation approach guided by principles of integrity, authenticity, respectfulness, consistency, and informality. Along with the graph contrastive learning with real and synthetic data, we further advance the performance of SEGA under limited resources. We conduct experiments on two real-world clinical interview datasets in English and Chinese, respectively. The results demonstrate that SEGA significantly surpasses existing baseline methods and powerful LMs like GPT-4 for depression detection in clinical interviews.

2 Related Work

2.1 Depression Detection

Depression is a debilitating mental condition that affects a person’s thoughts, feelings, and behavior. The automatic analysis of clinical interviews has been recognized as a promising assistant for psychologists and psychiatrists to improve the time efficiency and diagnostic consistency of depression detection (Valstar et al., 2016; Zou et al., 2022).

Most studies leverage the participant answers for depressive assessment, which can be categorized into three types: 1) Uni-modal methods typically model either transcripts or audios. For transcripts, existing work focuses on aggregating word representations for prediction (Mallol-Ragolta et al., 2019; Burdisso et al., 2023), or further incorporates affective and mental health lexicons (Xezonaki et al., 2020; Villatoro-Tello et al., 2021). For audios, existing studies have investigated the utility of shallow acoustic features like MFCCs (Taguchi et al., 2018; Huang et al., 2022) and deep audio representations from neural networks like CNN (Ma et al., 2016; Zhang et al., 2021; Sardari et al., 2022; Sun et al., 2022). 2) Bi-modal methods typically combine transcripts and audios as the two most informative modalities via feature fusion (Al Hanai et al., 2018; Wu et al., 2022; Guo et al., 2022; Seneviratne and Espy-Wilson, 2022) or prediction ensembling (Niu et al., 2021; Shen et al., 2022). Yoon et al. (2022) model both audio and video modalities with a cross-attention transformer, but the performance is relatively limited without transcripts. 3) Tri-modal methods consider the answer transcripts, audios, and videos via multi-modal fusion conducted at word-level (Rohanian et al., 2019) or utterance-level (Guo et al., 2020; Zhang et al., 2020; Zheng et al., 2020; Fang et al., 2022; Prabhu et al., 2022).
In addition to the three modalities of the answers, a few studies (Niu et al., 2021; Flores et al., 2022) have incorporated the interview questions as supplementary context to better extract salient cues from participant answers. Different from the above methods, we resort to human experience in depression assessment to construct the structural element graph, so as to effectively model all four types of interview elements under limited resources.

2.2 LLMs Application in Healthcare

Recent years have witnessed the rapid development and powerful capability of large language models (LLMs), such as OpenAI’s ChatGPT (Ouyang et al., 2022), Anthropic’s Claude (Bai et al., 2022), Google’s LaMDA (Thoppilan et al., 2022), and Meta’s LLaMA (Touvron et al., 2023). LLMs have shown promising potential in powering the healthcare domain. Wang et al. (2023) and Chen et al. (2023) use LLMs to play doctor and patient roles, respectively, to generate medical dialogue data. Yang et al. (2023), Bisercic et al. (2023), and Jo et al. (2023) leverage LLMs to assist diagnosis via medical report generation, tabular data extraction, and emotional support deployment, respectively. Sarker et al. (2023) and Liyanage et al. (2023) conduct data augmentation via LLMs to enhance medication event identification and wellness dimension classification.

Despite these advances, LLMs remain underexplored in the realm of depression detection in clinical interviews. In this work, we meticulously craft principle-based prompts to direct LLMs in generating high-quality data for depression detection, complemented by responsible quality verification and ethical consideration.

3 Methodology

3.1 Task Definition

In a clinical interview $I$ from the corpus $C$, the interviewer asks the participant a series of $m$ questions (recorded as text) $Q = \{q_1, ..., q_m\}$, typically probing into their feelings, experiences and mental states. The participant’s responses are recorded in three modalities - transcripts $T = \{t_1, ..., t_m\}$ capturing the textual content, audios $A = \{a_1, ..., a_m\}$ capturing tone and prosody, and videos $V = \{v_1, ..., v_m\}$ capturing facial expressions and body movements. Each question $q_i$ and transcript $t_i$ comprises a word sequence, while each audio $a_i$ and video $v_i$ comprises a frame sequence sampled with a certain frequency. The goal is to analyze these elements and predict a label $y \in \{1, 0\}$ denoting whether the participant is depressed or not.

3.2 Structural Element Graph

For depression detection in clinical interviews, we propose the structural element graph (SEGA) based on human experience for comprehensive interview modeling. Figure 2 presents the architecture of SEGA, which comprises three main components: element feature extraction, graph construction with expertise, and graph learning and prediction.

3.2.1 Element Feature Extraction

In clinical interviews, each element serves as a unique indicator of the depressive state. Interview questions may probe for negative emotions, shedding light on underlying feelings of hopelessness or worthlessness. Answer transcripts may reveal a higher frequency of negative words, reflecting a depressed mindset. Answer audios can hint at depression through a flatter affect, slower speech rates, or prolonged pauses. Answer videos capture visual cues like reduced facial expressivity or subdued reactions. Together, these elements form a rich tapestry of information that helps in assessing an individual’s depressive state.

Below we show how to extract element features. Within a question-answer round, the question ($q_i$) transcript ($t_i$) is an utterance involving the text modality. We first map every word in the utterance using a pre-trained word embedder, then compute the mean of word vectors to obtain the utterance representation $q_i / t_i$. For the audio ($a_i$) and video ($v_i$) modalities of the answer utterance, they are already processed into frames at a certain frequency, where the raw features are specified by the used dataset. We then process those frames according to the timestamps in the corresponding transcript. Specifically, we look at the start and end timestamps of the utterance in the transcript, then average the frames spanning the duration of this utterance to derive audio and video representations $a_i / v_i$ for the answer utterance. By this means, we ensure that all element features are strictly aligned at the granularity of a question-answer round.

3.2.2 Graph Construction with Expertise

Based on the element features, we follow specific expertise for depression assessment to derive the graph structure and facilitate interview modeling.
Expertise within Each Round: 1) Element Importance. Studies on multi-modal learning have demonstrated that texts contain the richest semantics (Rohanian et al., 2019; Hazarika et al., 2022). Therefore, for each round, we anchor the transcript $t_i$ as the central node, and treat the audio $a_i$ and video $v_i$ as auxiliary nodes. The question $q_i$ is also recognized as a supplementary node since it completes the contextual information of the answer. 2) Information Flow. To reflect the status of nodes, we build a directed acyclic graph (Shen et al., 2021) where information is passed unidirectionally from the auxiliary nodes to the central node. 3) Noise Suppression. To avoid direct interference from the background noise in audio and video elements (Baltrušaitis et al., 2018), we additionally insert a proxy node $p_i$ within each round. It aims to first distill evidence from the audio and video, transform it into the text vector space, and then let the transcript node selectively absorb useful information. The representation of $p_i$ is initialized with the representation of $t_i$. Besides, each element node has a self-loop edge to preserve its own information.

Expertise across Multiple Rounds: 1) Information Propagation. Since an interview topic (e.g., querying family relationships) may span multiple rounds, we link adjacent central nodes ($t_i$) for capturing temporal dependencies. 2) Information Aggregation. To represent the semantics of the entire clinical interview, we introduce an additional virtual “summary” node ($s$). We then pass information unidirectionally from all central nodes ($t_i$) to the summary node $s$. The representation of $s$ is initialized by averaging all $t_i$ representations.

The structural element graph is a simple yet effective architecture. On the one hand, it fully utilizes the evidence in four types of elements embedded in clinical interviews. On the other hand, it benefits from prior human experience and sets a good starting point for interview modeling.

3.2.3 Graph Learning and Prediction

Based on the constructed structural element graph, we employ Graph Attention Network (GAT) (Velickovic et al., 2018) to capture the interactions among elements and accordingly update representations. For node $i$ and its neighborhood nodes $\mathcal{N}(i)$, we iteratively updates the node representation $h_i \in \mathbb{R}^{d_h}$ by aggregating neighborhood node representations using multi-head attention:

$$r_{ij}^{(k)} = \text{LeakyReLU}(w^T [W^{(k)} h_i || W^{(k)} h_j]), \quad (1)$$

$$o_{ij}^{(k)} = \text{softmax}_j(r_{ij}^{(k)}) = \frac{\exp(r_{ij}^{(k)})}{\sum_{m \in \mathcal{N}_i} \exp(r_{im}^{(k)})}, \quad (2)$$

$$h_i^{(k)} = \text{ELU} \left( \sum_{j \in \mathcal{N}_i} o_{ij}^{(k)} W^{(k)} h_j \right), \quad (3)$$

$$h_i = \|_{k=1}^{K} h_i^{(k)}, \quad (4)$$

where $k$ denotes an attention head, $W^{(k)}$ is the weight matrix, $w$ is the weight vector, $o_{ij}^{(k)}$ is the attention weight, $h_i^{(k)}$ is the output feature vector for head $k$, $\|$ denotes vector concatenation. After the computation of $L$ GAT layers, we collect the representation $s$ of the summary node after propagation as the final feature of the interview. We then feed it into a feed-forward layer to predict the depressive status $\hat{y}$ and compute the cross-entropy loss $\mathcal{L}_{ce}$ for detection:

$$\hat{y} = \text{softmax}(W_s(s)), \quad (5)$$

$$\mathcal{L}_{ce} = -y \log(\hat{y}) + (1 - y) \log(1 - \hat{y}). \quad (6)$$

3.3 A Helping Hand from LLMs

While SEGAl intrinsically aligns with the expertise of processing interviews, its learning capability is still constrained by the limited clinical resources. Due to collection difficulties and privacy concerns, the training set size of commonly-used public clinical depression detection corpora is only about 100. Nevertheless, large language models (LLMs) possess an extremely strong human-like
text generation capability, offering immense potential for generating synthetic data to alleviate data scarcity. Hereby, we propose to design a principle-based prompt to carefully control the LLM for data augmentation. The goal is to simulate diverse participant responses by rephrasing transcripts with the same meanings but different vocabularies.

3.3.1 Principle-based Augmentation

While LLMs are capable of generating fluent and coherent text, their outputs can sometimes deviate from the original meaning or contain inappropriate content if not properly constrained. This presents challenges when utilizing them for data augmentation in sensitive domains like mental healthcare. Therefore, to ensure LLMs rephrase transcripts faithfully without altering semantics or tones, we guide the synthesizing process with explicit principles. The five principles serve as “guard rails” to steer the model toward making edits that introduce diversity while preserving integrity, authenticity, respectfulness, consistency, and informality.

We list the principles as follows. 1) Integrity of Content: Retain the original meaning and sentiment. Do not change the substance. 2) Conversational Authenticity: Use natural, casual language. Avoid overly formal styles. 3) Respectful Communication: Maintain a respectful tone. Do not make inappropriate/offensive alterations. 4) Consistency in Length: Keep similar length to the original. Avoid excessive shortening/lengthening. 5) Tolerance for Informality: Tolerate some irregularities (omissions, repetitions, filler words) given the conversational context.

We rephrase transcripts in a sliding window with a size of three, applying LLMs with context awareness. In other words, when rephrasing each answer transcript, we use the last three rounds of question-answers as input. Besides, we also provide $N$ hand-crafted demonstrations to facilitate task understanding. As shown in Figure 3, the final input prompt is composed of task instruction, rephrasing principles, and manual demonstrations. After prompting, we collect the outputs of LLMs as the augmented data. For the accompanying audios and videos, due to the untouchability of raw recordings and the lack of LLM-level synthesis tools, we adopt a simple random frame-swapping method.

After synthesizing augmentation data via rephrasing, we manually inspect samples for quality verification. We verify that: 1) The rephrased sentences express the same semantic content as the originals, without introducing factual inconsistencies or changes in meaning. 2) The tone and sentiment remain consistent between the original and augmented versions. 3) No offensive, unethical, or otherwise inappropriate language is generated. 4) Personal details and identifiers are not altered or exposed. Ultimately, steered by the guiding principles, only very few (less than 1%) of the augmented data is filtered out and regenerated.

3.3.2 Empowerment of SEGA

After obtaining the synthetic data $\tilde{C}$, we mix it with the original corpus $C$ to obtain the augmented training data. Furthermore, owing to the improved size and diversity with synthetic data, we introduce the self-supervised contrastive learning with InfoNCE (Oord et al., 2018) loss to encourage distinct representations for depressed and control participants:

$$
L_{cl} = -\sum_{c}^{\text{all}} \log \frac{\exp(s_c^+/\tau)}{\exp(s_c^+/\tau) + \sum_{j=1}^{J} \exp(s_{cj}^+/\tau)}
$$

where $\tau$ denotes a temperature parameter, $s_c^+$ is a certain positive sample with the same label, $s_{cj}$ denotes several negative samples with the opposite

Figure 3: Guided by principles, LLMs synthesize augmentation data by rephrasing the answer utterances.
By updating the final training loss as \( L = L_{ce} + L_{cl} \), we obtain the empowered model SEGA++.

4 Experiment

4.1 Experimental Setup

Datasets

We conduct experiments on two available corpora for clinical depression detection: DAIC-WOZ (Gratch et al., 2014) and EATD (Shen et al., 2022). DAIC-WOZ is a widely-used English dataset that collects interviews from 189 participants, recorded in transcripts, audios, and videos. Each participant is accompanied by a PHQ-8 score (Kroenke et al., 2009), and those with scores \( \geq 10 \) are labeled as depressed ([D]), while the others are labeled as control ([C]). DAIC-WOZ has a fixed data split, and we follow previous studies by using the development set for evaluation since the labels of the test set are unavailable. EATD is a newly released Chinese dataset that consists of interviews conducted with 162 student volunteers, recorded in transcripts and audios. Each participant is accompanied by an SDS score (Zung, 1965), and those with scores \( \geq 53 \) are labeled as depressed. EATD is split into training and development sets via 3-fold cross-validation with train:dev = 2:1. The detailed statistics of datasets are shown in Table 2. The details of interview recordings and pre-processing methods can be found in Appendix A.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Size</th>
<th>Category</th>
<th>Round</th>
<th>Token</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAIC-WOZ</td>
<td>train</td>
<td>[D] 30 [C] 77</td>
<td>((\bar{x}=57))</td>
<td>((\bar{x}=1,394))</td>
<td>((\bar{x}=15m04s))</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>((\bar{x}=57))</td>
<td>((\bar{x}=1,394))</td>
<td>((\bar{x}=15m04s))</td>
</tr>
<tr>
<td></td>
<td>dev</td>
<td>[D] 12 [C] 23</td>
<td>((\bar{x}=55))</td>
<td>((\bar{x}=1,531))</td>
<td>((\bar{x}=17m09s))</td>
</tr>
<tr>
<td>EATD</td>
<td>train</td>
<td>[D] 23/20/24 [C] 85/88/84</td>
<td>((\bar{x}=3))</td>
<td>((\bar{x}=181))</td>
<td>((\bar{x}=47s))</td>
</tr>
<tr>
<td></td>
<td>dev</td>
<td>[D] 7/10/6 [C] 47/44/48</td>
<td>((\bar{x}=3))</td>
<td>((\bar{x}=177))</td>
<td>((\bar{x}=49s))</td>
</tr>
</tbody>
</table>

Table 2: Detailed statistics of DAIC-WOZ and EATD.

Settings

We first discuss feature pre-processing. For DAIC-WOZ, we use the GloVe.840B.300d embeddings (Pennington et al., 2014) to vectorize interview questions and answer transcripts. The videos and audios have already been processed by COVAREP (Degottex et al., 2014) and OpenFace (Baltrušaitis et al., 2016), respectively. For EATD, we use the pre-trained Chinese BERT (Cui et al., 2020) as the word embedder. EATD only provides unprocessed raw audios, so we extract the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) (Eyben et al., 2016) via OpenS-MILE (Eyben et al., 2010) as acoustic features.

For the implementation of SEGA, the dimension of all hidden states is \( d_h=256 \). The layer number of GAT is \( L=2 \). The number of attention head \( K=8 \). We use gpt-3.5-turbo-0613 for principle-guided data augmentation. In the prompt, we include \( N=2 \) manually written samples as the demonstration. We collect three rephrased results for each transcript, but only use the first one since no significant performance difference is observed, resulting in 2x training samples. We perform contrastive learning within a single batch with \( \tau=0.05 \). For DAIC-WOZ/EATD, we train the models for 300/100 epochs using Adam optimizer (Kingma and Ba, 2015) with a learning rate 1e-4/2e-5 and batch size 8. We run all experiments on a single Tesla V100 32G GPU in Ubuntu 16.04. SEGA contains 2.6M trainable parameters and requires 0.33 GPU hours for training. The hyperparameter ranges can be found in Appendix B.

Following previous studies (Burdisso et al., 2023), we use depressed, control (i.e., non-depressed), and macro F1-scores as the metrics. We select the checkpoint for evaluation based on macro F1-scores. The final results for comparison are the average scores of 3 runs with random seeds (for DAIC-WOZ) or folds (for EATD)\(^3\).

4.2 Compared Methods

After reviewing the literature on depression detection, we find that the pre-processing of clinical interviews in different methods varies considerably, making it inaccurate to compare their reported results directly. For instance, Niu et al. (2021) employ manual cleaning of transcripts, while Shen et al. (2022) group every ten participant answers as one sample. Additionally, there are hardly any reported results on the recently released EATD dataset. Therefore, we select the four latest top-performing methods involving participant transcripts, audios, videos, and interview questions used in SEGA. Since most models are not open-sourced, we carefully replicate them following the original papers. Our implementation may serve as a basis for fair comparison in future research.

\( \omega \)-GCN (Burdisso et al., 2023) is a uni-modal model that leverages answer transcripts. EATD-Fusion (Shen et al., 2022) is a bi-modal model that leverages answer transcripts and audios. MFM-Att (Fang et al., 2022) is a tri-modal model that lever-

\(^3\)Code is available at https://github.com/zhchen18/SEGA.
Table 3: Depressed, Control, and Macro F1-scores. The best scores are in bold, and the best baseline scores are underlined. Results with † are significantly better than baselines (p < 0.05) based on a one-tailed unpaired t-test.

<table>
<thead>
<tr>
<th>Method</th>
<th>Element</th>
<th>DAIC-WOZ</th>
<th>EATD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T</td>
<td>A</td>
<td>V</td>
</tr>
<tr>
<td>ω-GCN</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>EATD-Fusion</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>MFM-Att</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>HCAG</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>GPT-3.5 (Zero Shot)</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>GPT-3.5 (Few Shot)</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>GPT-4 (Zero Shot)</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>GPT-4 (Few Shot)</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>SEGA</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>SEGA++</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
</tbody>
</table>

Table 3: Depressed, Control, and Macro F1-scores. The best scores are in bold, and the best baseline scores are underlined. Results with † are significantly better than baselines (p < 0.05) based on a one-tailed unpaired t-test.

4.3 Main Results

We present the performance on DAIC-WOZ and EATD datasets in Table 3. Obviously, SEGA achieves state-of-the-art performance, outperforming the best baseline by 1.12% and 2.36% in terms of macro F1-scores on two datasets, respectively. After LLM-empowered data augmentation, we further compare the performance of GPT-3.5 and GPT-4 by prompting them to determine whether the participant is depressed according to the clinical interview. We consider both zero-shot and few-shot settings, where the latter includes a control and a depressed training sample as two demonstrations. The implementation details of baseline methods can be found in Appendix C.

We then inspect the performance of the GPT family. Despite the strong general capabilities, GPTs’ performance still lags behind existing baselines. We hereby have the following observations: 1) GPT-4 generally outperforms GPT-3.5, benefiting from more parameters and pretraining data. 2) Few-shot models significantly exceed zero-shot models, where the additional depressed and control samples serve as anchors to aid judgment. 3) Performance on English DAIC-WOZ is markedly higher than Chinese EATD, indicating the language preference inside LLMs. 4) LLMs are more accurate at identifying the control class, showing the data encountered during LLM pretraining is also imbalanced, limiting discrimination of the depressed class. Overall, the primary reason for the inferior performance is that LLMs cannot explicitly model the specific pragmatic structure of clinical interviews, and they are inherently less adept at integrating multi-modal evidence such as useful MFCCs and eye movements.

Our key findings can be summarized into two aspects. One is that effective structure modeling, rather than simply introducing more elements, is more important. SEGA’s structural element graph has shown strength in capturing clinical interviews. The other is that, directly applying LLMs to specialized domains like mental health still yields unsatisfactory performance. We effectively absorb applicable knowledge from LLMs with principle-guided augmentation, which empowers SEGA++ to make further performance gains.
5 Analysis

5.1 Ablation Study

We conduct extensive ablation studies to investigate the contribution of different components and present the results on DAIC-WOZ.

In Figure 4, we present models with variations on the graph structure. We first include variants “Transcript”, “Video”, and “Audio” representing variants using only the respective single modality, where the results emphasize the need for integrating different elements for interview modeling. We then examine the graph construction. “FC Graph” uses fully connected subgraphs for each round’s four elements; “Bi Edge” alters unidirectional edges to bidirectional without modifying the structure; “w/o Proxy” involves removing proxy nodes, allowing auxiliary nodes to directly connect with central nodes. The evident performance drop in these structural variants underscores the effectiveness of SEGA based on human experience.

In Figure 5, we present models with different modalities. For bi-modal models, combining audios and videos as “A+V” leads to performance gains compared to these two single modalities. However, adding audios or videos to transcripts, forming “T+A” or “T+V”, results in similar or even decreased performance. This suggests that audios and videos contain relatively limited information, thus potentially complementing each other. In contrast, transcripts contain abundant information, making it challenging for audio or video alone to effectively supplement valuable information. Nevertheless, when all these modalities are combined as “T+V+A”, a slight improvement over the single transcript modality can be achieved. Moreover, modeling interview questions (i.e., SEGA) can further enhance the performance.

In Figure 6, we compare variants for empowerment. The absence of contrastive learning (“w/o CL”) results in a performance decline. Removing principle-based augmentation (“w/o DA”) leads to a performance drop to the level of the original SEGA, highlighting the effectiveness of synthetic data. “Free Gen” leverages LLMs to role-play depressed/control participants and answer questions for augmentation, resulting in a substantial performance decrease, indicating its impracticality. “EDA” (Wei and Zou, 2019) involves easy data augmentation like deletion, insertion, and swapping, which brings insignificant improvement.

5.2 Case Study

To have a close look, we further select a correctly predicted depressed sample from the development set of DAIC-WOZ, and highlight a question-answer round heavily weighted by the summary node per SEGA’s average graph attention scores. The answer transcript and audio’s Mel spectrograms are presented in Figure 7. The video is omitted due to the unavailability of original recordings.

Figure 7: The answer transcript and audio’s Mel spectrograms of a depressed participant.
In the transcript, we mark negative emotion words in blue, expressing the participant’s distress and regret in semantics. Moreover, from the green highlights and Mel spectrograms, we observe slower speech and more pauses, which existing research identifies as typical depressive symptoms (Sobin and Sackeim, 1997). This demonstrates SEGA’s strength in integrating indicative cues across modalities via the structural element graph to make accurate detection.

5.3 Implications on Mental Healthcare

Mental healthcare currently faces challenges such as insufficient specialized professionals, uneven distribution of resources, and stigma surrounding treatment seeking. By proposing SEGA, we absolutely do not intend to replace mental health experts like psychologists and psychiatrists. Rather, our goal is for SEGA to act as a qualified assistant.

For practical implementation, clinical depression assessment can be formulated as a two-step pipeline. The first step is conducting automated interviews, for which existing methods like Wizard-of-Oz or fully automated agents are viable options. The second step is the automatic detection by analyzing the interview recordings to aid in judging the participant’s depressive state, like SEGA. By combining these two steps, we can alleviate the burden on professionals, compensate for imbalanced medical resources, expedite widespread depression screening, and ultimately promote public health and well-being for people worldwide.

6 Conclusion

In this paper, we propose a novel structural element graph SEGA for depression detection in clinical interviews. SEGA comprehensively captures all key elements - interview questions, answer transcripts, audios, and videos - within a simple yet effective directed acyclic graph. To further address data scarcity, we design the empowered SEGA++ via principle-guided augmentation with LLMs and graph contrastive learning. Experiments on two real-world datasets demonstrate the state-of-the-art performance of our method, surpassing the latest baseline methods and zero/few-shot GPT-3.5/4.

Acknowledgments

This work was supported by the National Science Foundation for Distinguished Young Scholars (No. 62125604) and the NSFC Key Project (No. 61936010). This work was also supported by Tsinghua Precision Medicine Foundation and Tsinghua University - Beijing Tsingshang Architectural Decoration Engineering Co., Ltd. Joint Institute for Smart Scene Innovation Design.

Limitations

We discuss the limitations of our work as follows.

1) Task Setting. In this work, we focus on depression detection in clinical interviews, a serious “laboratory setting” typically conducted by mental health experts like psychologists and psychiatrists, representing a standard approach in depression screening. Contrasting this is the “wild setting”, which refers to informal early screenings for daily contents, such as analyzing social media posts (e.g., on Reddit). Due to significant differences between these settings in data structures, modalities, and judgment criteria, intuitively, our method designed for the “laboratory setting” might not effectively transfer to the “wild setting”.

2) Modality Specificity. Due to our reliance on publicly available datasets, our methodology is limited to three modalities: text (interview questions and participant responses), audio (MFCC, formants), and video (facial actions, eye gazes, etc.). We do not include other medically relevant information such as electroencephalograms (EEGs), near-infrared signals, or magnetic resonance imaging (MRI) signals.

3) Corpus Size. Considering the challenges in data collection and privacy considerations, the datasets available for depression detection in clinical interviews usually have limited samples. In experiments, we report the average results across multiple runs and conduct significance testing to ensure that any performance improvements are statistically valid.

4) Data Release. Due to licensing constraints of the DAIC-WoZ and EATD datasets, we are unable to directly release raw interview recordings. Instead, under appropriate licensing or registration, we can provide pre-processed vector features to those who have already applied to obtain the raw data. Access to the original materials is possible for anyone who completes an application form and receives approval from the dataset authors.

Ethical Considerations

We here elaborate on the potential ethical issues.
1) Data Privacy and Consent. We apply and use publicly available datasets, DAIC-WoZ and EATD, for depression detection in clinical interviews. According to the original dataset papers, both datasets have received approval from Institutional Review Boards. All participants are informed that their interviews are for academic research. All personal details like names, ages, and professions are either removed or anonymized, eliminating any risk of personal information exposure. The DAIC-WoZ dataset, involving a video modality, does not provide original videos but de-identified vector features of facial actions and eye gaze, making it impossible to reconstruct the participants’ appearances. Therefore, the information of participants is comprehensively and rigorously protected, with no privacy breaches.

2) Participant Demographics. The DAIC-WoZ dataset, collected by researchers from the Institute for Creative Technologies at the University of Southern California, encompasses participants including the U.S. armed forces veterans and the general public from the Greater Los Angeles metropolitan area. The EATD dataset, collected by researchers from the School of Software Engineering at Tongji University, primarily includes teachers and students from universities in Shanghai. The focus on specific groups in both datasets is a result of objective conditions and research goals, rather than any intention of bias or unfairness towards race, nationality, age, or gender.

3) Role of AI in Diagnosis. Our method aims to serve as an intelligent assistant for mental health experts such as psychologists and psychiatrists, not to replace them. Directly using our method to calculate depression coefficients may cause or perpetuate algorithmic bias, potentially leading to inaccurate diagnoses. Therefore, the model predictions should only be used as a reference, while the final diagnosis must be cautiously determined by professionals. Our method is based solely on completed interview recordings for preliminary screening and does not offer medical advice or intervention.

4) Dataset Access and Use. For the DAIC-WoZ dataset, we meticulously review the DAIC-WoZ End-User License Agreement and submit the necessary official application forms, securing consent from the original authors. For the EATD dataset, we download the data from the author’s official GitHub repository. We strictly use the two datasets exclusively for the research purposes of this work. No unauthorized dissemination or sharing of the data is conducted.

5) LLM Data Augmentation. In the use of large language models for data augmentation, we apply principle-guided rephrasing instead of free generation to control the model behavior. This approach ensures that the model performs semantic-preserving rewrites and does not introduce imaginative, additional information. We rigorously verify all synthetic data to ensure no introduction of any unsafe or harmful content.

References


Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. COVAREP - A


8192


A Details of Dataset Pre-Processing

We here discuss the raw information provided by the dataset and the detailed preprocessing methods.

A.1 DAIC-WOZ

The Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) provides multi-modal data, including transcripts, audios, and videos of clinical interviews.

- Transcript: DAIC-WOZ provides complete textual transcripts of each interview containing interview questions and participant answers. We process transcripts with GloVe.840B.300d word embeddings.
- Audio: DAIC-WOZ provides the raw audio of each interview along with acoustic features extracted using COVAREP, including F0, VUV, NAQ, QQ, H1H2, PSP, MDQ, peakSlope, Rd, Rd_conf, MCEP_0-24, HMPDM_0-24, HMPDD_0-12, and the first 5 formants.
- Video: DAIC-WOZ provides the visual features extracted from each interview using OpenFace, instead of the original video recordings. These include 2D/3D facial landmark points, facial action units, eye gazes, Felzenswalt’s HoG features, and head poses.

A.2 EATD

The Emotional Audio-Textual Depression Corpus (EATD) provides multi-modal data, including transcripts and audio, without video.

- Transcript: EATD provides only the transcribed text of participant answers. We supplemented these with 26 additional interview questions based on the content of the answers. We use pre-trained Chinese BERT to obtain word embeddings for the text.
- Audio: EATD provides audio recordings of each participant’s answers. Since no specific audio processing tool is specified, we extract the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) acoustic features using OpenSmile, which contains features related to frequency-related parameters like formant, energy-related parameters like loudness, and spectral parameters like MFCCs.

B Details of Hyperparameter Ranges

We present the hyperparameter ranges in Table 4. We select all hyperparameters via manual tuning.

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Range</th>
<th>Selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAT &amp; Bi-LSTM hidden state</td>
<td>(64, 128, 256, 512)</td>
<td>256</td>
</tr>
<tr>
<td>number of GAT layers $L$</td>
<td>(1,2,3,4)</td>
<td>2</td>
</tr>
<tr>
<td>number of attention heads $K$</td>
<td>(1, 2, 4, 8, 16)</td>
<td>8</td>
</tr>
<tr>
<td>number of demonstrations $N$</td>
<td>(1, 2, 3)</td>
<td>2</td>
</tr>
<tr>
<td>temperature in contrastive loss $\tau$</td>
<td>(0.01, 0.05, 0.1, 0.5, 1.0)</td>
<td>0.05</td>
</tr>
<tr>
<td>batch size</td>
<td>(4, 8, 16, 32)</td>
<td>8</td>
</tr>
<tr>
<td>learning rate</td>
<td>(1e-5, 2e-5, 5e-5, 1e-4, 2e-4, 5e-5)</td>
<td>1e-4/2e-5</td>
</tr>
</tbody>
</table>

C Details of Baseline Implementation

We here illustrate the implementation of baseline methods.

- $\omega$-GCN is a uni-modal model that leverages answer transcripts for depression detection. It constructs a graph between words and interview documents, and uses TF-IDF, PMI, and PageRank to obtain the edge weights for graph convolution.
- EATD-Fusion is a bi-modal model that leverages answer transcripts and audios for depression detection. It uses a gating mechanism for integrating predictions from two modalities.
- MFM-Att is a tri-modal model that leverages answer transcripts, audios, and videos for depression detection. It first learns features of single modalities, then uses an attention fusion network to obtain fused multi-modal representation. MFM-Att for EATD only uses transcripts and audios since there is no video modality in EATD.
- HCAG is a model that captures the interaction between interview questions and participant answers for depression detection. It concatenates the question and answer in each round as a vertex, and constructs edges between adjacent QA rounds in a
Below is a transcript of an interview between an interviewer and a participant. Based on the content of the interview, determine whether the participant is depressed or not. Only answer 'Yes' or 'No' without any explanations.

{ INPUT CLINICAL INTERVIEW }

Below is a transcript of an interview between an interviewer and a participant. Based on the content of the interview, determine whether the participant is depressed or not. Only answer 'Yes' or 'No' without any explanations.

Example 1
Input: {DEMONSTRATION CLINICAL INTERVIEW 1 }
Output: LABEL OF DEMONSTRATION 1

Example 2
Input: {DEMONSTRATION CLINICAL INTERVIEW 2 }
Output: LABEL OF DEMONSTRATION 2

Target
Input: { INPUT CLINICAL INTERVIEW }
Output:

Table 5: The zero-shot and few-shot prompts for evaluating GPT-3.5 and GPT-4 for depression detection in clinical interviews.

QA graph. GAT is used to update vertex features and make predictions.

- **GPT-3.5** and **GPT-4** are included as the competitors in experiments. Given the length of the clinical interview, for GPT-3.5, we employ the gpt-3.5-turbo-16k-0613 API (the 4k context window of the default gpt-3.5-turbo-0613 is insufficient in length). As for GPT-4, we utilize the standard gpt-4-0613 API with an 8k context window. The zero-shot and few-shot prompts used for evaluating the GPT family in depression detection are shown in Table 5.

D Use of AI Assistants

We use ChatGPT to polish some of the content.