

Explainable Multimodal Depression Recognition in Clinical Interviews via PHQ-Aligned Symptom Summarization

Wenjie Zheng, Qiming Xie, Jianfei Yu*, Yang Wang, Lei Cao, Fei Wang*, Shijin Wang, Rui Xia*, and Chengqing Zong, *Fellow, IEEE*

Abstract—Recent advances in multimodal depression recognition for clinical interviews (MDRC) have demonstrated the potential of AI systems by integrating textual, acoustic, and facial cues. However, existing methods pay limited attention to interpretability, thereby constraining reproducibility and clinician review. To address this, we introduce Explain-MDRC, an explainable MDRC framework that mirrors clinical workflows by generating structured symptom summaries from text and integrating them with nonverbal cues for recognition. Specifically, we construct Explain-DAIC, a dataset based on DAIC-WOZ and enriched with PHQ-8-aligned summary annotations, providing a foundation for developing models with built-in interpretability. We further propose PhqCML, a model that combines PHQ-8-aligned symptom summarization with PHQ-aware contrastive learning and summary-informed multimodal fusion. Automated metrics and expert evaluations show that Explain-MDRC improves recognition performance and provides more interpretable, clinician-readable intermediate evidence, suggesting a promising direction for transparent AI-assisted depression recognition research.

Index Terms—Explainable Depression Recognition, Structured Symptom Summarization, PHQ-aware Contrastive Learning

I. INTRODUCTION

Multimodal affective computing aims to automatically recognize and interpret human emotional states by integrating signals from multiple modalities, such as text, audio, and facial expressions [1]–[3]. These techniques have demonstrated significant potential in healthcare applications, enabling AI systems to assist in clinical assessment through the analysis of patients’ behavioral and emotional cues [4], [5]. Among mental health conditions, depression affects approximately 332 million people worldwide, yet a substantial proportion of cases remain unrecognized or undiagnosed (World Health Organization, 2024)¹. To address this gap, multimodal depression recognition for clinical interviews (MDRC) has been proposed to assist in the evaluation of depression severity, leveraging

interviewer-participant interactions enriched with multimodal data, including textual, acoustic, and facial information. Given its potential to improve clinical decision-making and enable earlier intervention, MDRC has garnered growing attention in recent years [6], [7].

Most existing MDRC studies formulate the task as a classification problem, employing strategies such as signal aggregation [8], [9], temporal modeling [10], [11], intra-modality representation learning [12], [13], and guidance from clinical instruments such as the Patient Health Questionnaire (PHQ) [14]–[17]. These studies have substantially advanced the state of the art in MDRC by improving predictive accuracy through well-designed frameworks. **However**, current MDRC research remains primarily focused on recognition performance, often overlooking the need to generate human-interpretable evidence that aligns with clinical reasoning processes. This gap is particularly concerning, as recent studies emphasize that interpretability is essential for developing clinically transparent, human-centered, and clinician-reviewable AI systems [18]–[21]. Moreover, enhancing interpretability may also strengthen recognition performance by encouraging models to ground predictions in clinically meaningful and explainable cues.

The necessity of such interpretability becomes clear when considering the standard clinical workflow for depression. In mental health practice, clinicians typically conduct structured or semi-structured interviews to assess a patient’s mental health status. Throughout the interview, they often evaluate symptom domains outlined in standardized depression scales and explore potential underlying causes. Importantly, clinicians synthesize this information into structured yet interpretative symptom summaries, informing clinical assessment impressions. Moreover, nonverbal cues, such as acoustic and facial information, are observed and integrated with these verbal summaries to support a more holistic assessment. The inability of current MDRC systems to generate such clinically meaningful summaries represents a gap between recognition capability and clinician-reviewable evidence. Addressing this requires moving beyond classification tasks to develop models capable of producing structured, evidence-backed symptom summaries that mirror clinicians’ reasoning processes.

To this end, we introduce a framework named Explainable Multimodal Depression Recognition in Clinical Interviews (**Explain-MDRC**) that explicitly simulates core aspects of clinical workflows. As illustrated in Fig. 1, given a clinical

Wenjie Zheng, Qiming Xie, and Jianfei Yu are with the School of Artificial Intelligence, Nanjing University of Science & Technology, Nanjing, China. Yang Wang and Fei Wang are with the Department of Psychiatry, the Affiliated Brain Hospital of Nanjing Medical University, Nanjing, China. Lei Cao is with the Faculty of Psychology, Beijing Normal University, Beijing, China. Shijin Wang is with the iFLYTEK AI Research (Central China), Hefei, China. Rui Xia is with the School of Intelligence Science and Technology, Nanjing University, China. Chengqing Zong is with the Institute of Automation, Chinese Academy of Sciences, Beijing, China.

*Corresponding authors: Jianfei Yu (e-mail: jfyu@njnu.edu.cn), Fei Wang (e-mail: fei.wang@yale.edu), and Rui Xia (e-mail: rxia@nju.edu.cn).

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

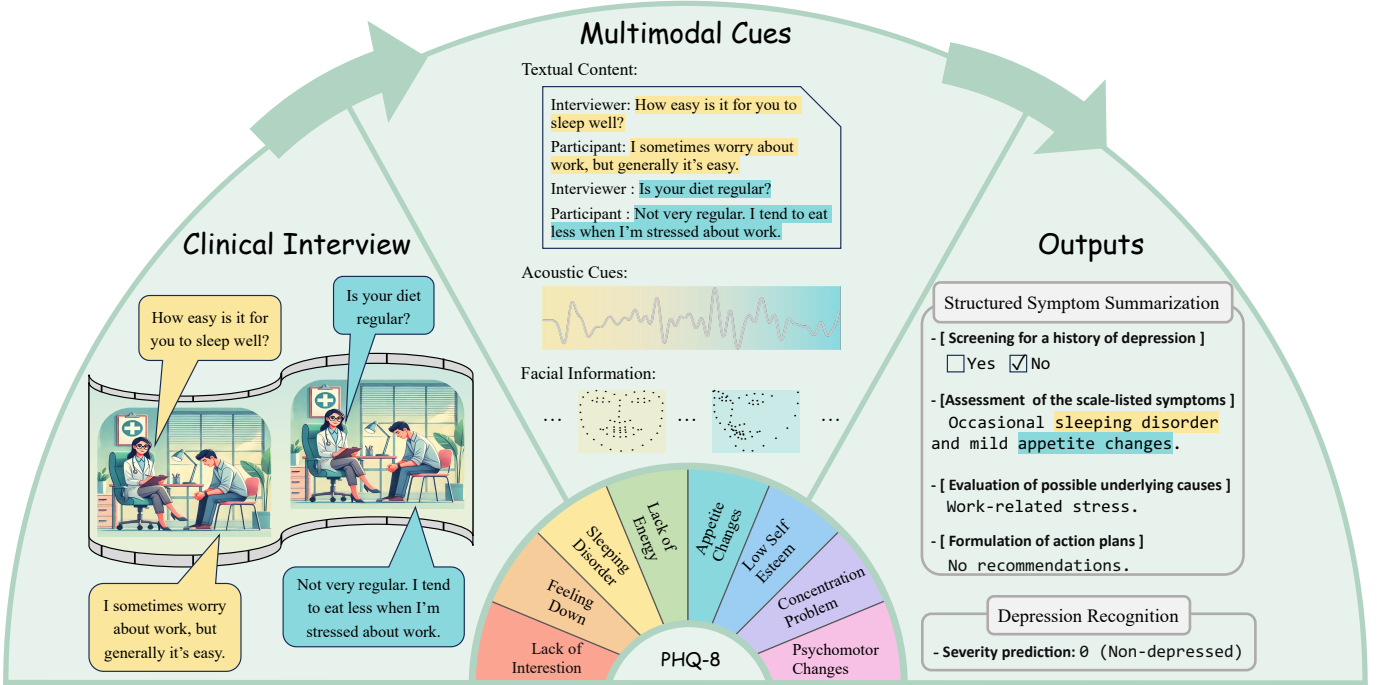


Fig. 1. The overview of our explainable MDRC framework, which leverages the PHQ-8 as a central pivot to provide structured and interpretable evidence for AI-assisted depression assessment. Highlighted dialogue text indicates supporting evidence for individual symptom components (e.g., yellow span = sleeping disorder).

interview scenario between an interviewer and a participant, Explain-MDRC first extracts multimodal cues, including textual content, acoustic cues, and facial information. It then performs two output tasks: (1) generating a structured symptom summary aligned with PHQ-8 [15], and (2) predicting the depression recognition label by integrating the generated summary with nonverbal cues. Specifically, the structured summarization encompasses four key components aligned with standard psychiatric evaluation procedures [22]: ① screening for a history of depression, ② assessment of the scale-listed symptoms, ③ evaluation of possible underlying causes, and ④ formulation of action plans, with supporting dialogue evidence anchoring the first three components.

To operationalize the Explain-MDRC framework, we first construct **Explain-DAIC** by extending the widely used MDRC dataset DAIC-WOZ [23], [24] with manually annotated PHQ-8-aligned symptom summaries. Unlike conventional depression-recognition datasets that mainly provide interview recordings and session-level labels, Explain-DAIC explicitly links participant utterances to clinically meaningful symptom dimensions, while retaining the original textual, acoustic, and visual signals for multimodal recognition. This design supports both clinician-readable symptom explanation generation and the evaluation of whether such explanations improve downstream depression recognition.

Building on this dataset, we propose **PhqCML** (PHQ-aware Contrastive Multimodal Learning), a trainable two-stage method that makes the symptom summary an explicit intermediate reasoning layer rather than a post-hoc explanation. In the first stage, PhqCML instruction-tunes an LLM to generate structured symptom summaries, and uses PHQ-

aware contrastive learning to align each summary’s hidden representation with the participant’s eight-dimensional PHQ-8 profile during training. In the second stage, PhqCML fuses the resulting PHQ-aware summary representation with acoustic and visual features for depression recognition.

Experiments on the Explain-DAIC dataset support the effectiveness of the proposed method. Specifically, PhqCML outperforms the previous state-of-the-art by 7.10 absolute percentage points in Macro-F1, underscoring that incorporating structured symptom summaries not only enhances recognition performance but also improves model interpretability. In a clinical expert evaluation, PhqCML outperformed GPT-5.5 on completeness, correctness, and conciseness, with all three comparisons reaching $p < 0.001$ under case-level one-sample Wilcoxon signed-rank tests, indicating improved clinical readability of the generated symptom summaries.

The main contributions are summarized as follows: ① We formulate Explain-MDRC, an explainable MDRC framework that couples PHQ-8-aligned symptom summarization with multimodal depression recognition, making clinician-readable symptom evidence an explicit intermediate output. ② We construct Explain-DAIC, a DAIC-WOZ-derived dataset enriched with manually annotated PHQ-8-aligned symptom summaries, enabling joint study of symptom explanation generation and interview-based depression recognition. ③ We propose PhqCML, which combines PHQ-aware contrastive learning and summary-informed multimodal fusion, achieving stronger recognition performance while providing structured intermediate evidence in clinical interviews.

Collectively, this work moves MDRC from black-box classification toward clinician-aligned symptom reasoning for AI-

TABLE I

COMPARISON OF THE PROPOSED EXPLAIN-DAIC DATASET WITH MAINSTREAM MULTIMODAL DEPRESSION RECOGNITION DATASETS FOR CLINICAL INTERVIEWS. THE TURNS IN PARENTHESES DENOTE THE TOTAL NUMBER OF QUESTION-RESPONSE UNITS IN THE DATASET. “SYM. SUM.” REFERS TO THE SYMPTOM SUMMARIZATION TASK AND “DEP. REC.” CORRESPONDS TO THE DEPRESSION RECOGNITION TASK.

Dataset	Modality	# Subjects	# Dialogues (Turns)	Duration	Language	Sym. Sum.	Dep. Rec.
CMDC (2022) [25]	text, audio, video	78	78 (1,848)	46.0h	Mandarin	✗	✓
EATD-Corpus (2022) [26]	text, audio	162	162 (486)	2.3h	Mandarin	✗	✓
DAIC-WOZ (2016) [24]	text, audio, video	189	189 (21,484)	47.3h	English	✗	✓
Explain-DAIC (Ours)	text, audio, video	189	189 (21,484)	47.3h	English	✓	✓

assisted depression recognition research.

II. RELATED WORK

A. Depression Recognition Datasets

Depression recognition aims to assess depressive severity from linguistic, acoustic, visual, or multimodal behavioral cues. Early text-oriented resources provide symptom-level supervision from social media or self-reported mental-health descriptions, including PHQ-9-aligned tweets [27], PRIMATE [28], PsySym [29], and BDI-Sen [30]. These datasets annotate symptoms at tweet, sentence, or question-response levels, enabling symptom-aware text modeling and fine-grained linguistic analysis. However, because they lack synchronized acoustic and visual behavior, they are not designed for multimodal clinical interview assessment. As research moved toward more ecologically valid scenarios, multimodal depression recognition datasets emerged across social media, in-the-wild videos, controlled tasks, and clinical interviews. Social-media datasets include multiTwitterDep [31], multiRedditDep [32], and the meme-level RESTORE dataset with fine-grained symptom labels [33]; in-the-wild resources collect video-level samples from public platforms [34]–[36]; and task-oriented corpora such as AVEC 2013/2014 [37], [38] and MODMA [39] record participants under controlled protocols. These resources broaden the sensing context from written self-expression to speech, facial behavior, and task responses, but their labels are usually user-, video-, or session-level. Clinical-interview datasets, including CMDC [25], EATD-Corpus [26], and DAIC-WOZ, capture richer doctor-patient or interviewer-participant interactions.

B. Depression Recognition Methods

Existing depression recognition methods improve prediction by introducing clinically informed supervision and richer behavioral representations. Early PHQ-aware studies use screening interviews, questionnaire responses, or symptom spans as intermediate signals for depression prediction. Rinaldi et al. [40] model interview prompts with latent categories, Nguyen et al. [16] leverage questionnaire-response prediction to improve generalization, and Zhang et al. [41] learn symptom-related spans with PHQ-aware contrastive objectives. Other work broadens the modeling scope by learning shared and disorder-specific cues across mental disorders [42] or by capturing local topic patterns and global temporal context for severity prediction [43].

With the rapid advancement of LLMs, several studies have begun integrating LLMs into mental-health assessment pipelines. Yang et al. [44] and Xu et al. [45] build domain-specialized mental-health LLMs based on foundation models such as LLaMA and Alpaca [46], while Yang et al. [47] and Hu et al. [48] explore interactive agent-based psychometric assessment. These studies improve task coverage and interaction ability, yet they mainly target broad mental-health prediction, screening, or adaptive questioning. Recent interview-oriented methods further use LLMs to extract depression indicators or estimate rating-scale scores from clinical conversations [49]–[51], suggesting that language models can recover clinically meaningful cues from long interviews. Multimodal systems such as IDRL [52], PsyGAT [53], and dynamic summary generation [54] also improve individual-aware, psychologically grounded, or summary-guided reasoning. Despite these advances, most MDRC systems still focus on final depression labels or task-specific explanations, rather than building PHQ-8-aligned, evidence-grounded symptom summaries as an explicit intermediate clinical representation.

C. Medical Dialogue Summarization

In the general medical field [55], Krishna et al. [56] employ a T5-based model to generate structured clinical summaries from physician-patient conversations. Michalopoulos et al. [57] introduce MedicalSum, a framework for summarizing medical conversations by integrating medical domain knowledge from the Unified Medical Language System. With the rise of LLMs, Van et al. [58] evaluate eight adapted LLMs across four clinical summarization tasks and demonstrated that, in many cases, these models can generate summaries comparable to or surpassing those produced by medical experts. Ghosh et al. [59] introduce the MMQS multimodal dataset and propose a CLIP- and LLM-based framework that integrates visual cues with textual queries to generate more accurate and medically informed question summaries. Saab et al. [60] introduce MedGemini, whose strong multimodal and long-context reasoning capabilities enable it to exceed human experts in clinical text summarization. Yuan et al. [61] propose LCDS, a logic-controlled system that reduces hallucinations in discharge summary generation through source-summary mapping, rule-based constraints, and expert review. In the mental health domain, Yao et al. [62] introduce D⁴, a Chinese depression-diagnosis-oriented dialogue dataset grounded in ICD-11 and DSM-5, supporting multiple clinical tasks and improving both empathy and diagnostic accuracy in consultation dialogue systems. Aragón et al. [63] develop a method that extracts

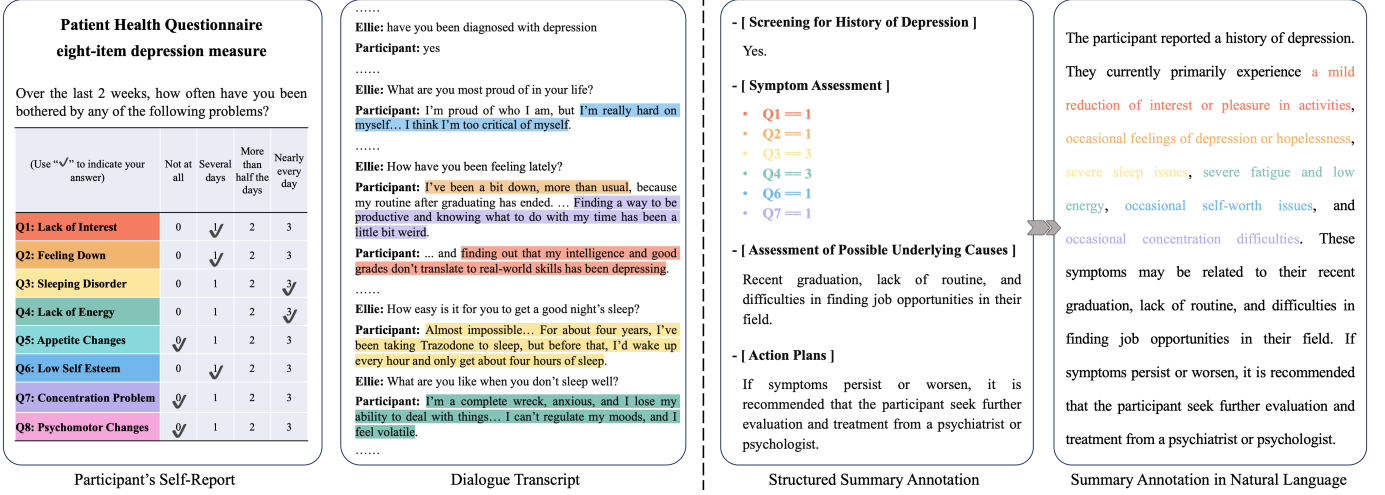


Fig. 2. Annotation process for a participant's structured symptom summary. For each symptom reported in the PHQ-8, annotators identify corresponding dialogue segments and assess possible underlying causes. Additional symptoms emerging during the conversation (such as Concentration Problem in this example) are also included in the summary.

BDI-related evidence from social media posts and uses an LLM to estimate answers to the 21 BDI-II items, achieving competitive performance with expert assessments.

III. DATASET

A. Data Source

Since no existing MDRC dataset provides the structured symptom summaries required to develop the Explain-MDRC framework, we construct a new dataset by extending DAIC-WOZ, a widely used MDRC dataset of clinical interviews with veterans in which a clinician-controlled virtual agent (Ellie) conducts semi-structured interviews. Each participant provided informed consent and completed the PHQ-8 prior to the session.

B. Motivation for Explain-DAIC Annotation

In clinical interview settings, semi-structured interviews do not strictly follow the PHQ-8; clinicians interpret spontaneously disclosed information and integrate nonverbal cues to inform clinical assessment impressions. **However**, DAIC-WOZ lacks structured annotations of participants' specific depressive symptoms and underlying causes, which limits its use for interpretable modeling. To address this gap, we use the PHQ-8 item structure as an annotation scaffold for organizing dialogue-grounded symptom evidence, while retaining the original DAIC-WOZ binary depression labels as the classification target. This yields Explain-DAIC, a dataset of PHQ-8-aligned symptom summaries supporting more transparent AI-assisted depression recognition research without redefining the original DAIC-WOZ recognition task.

C. Symptom Summarization Annotation

We employ three clinically trained mental health professionals who were supervised by a licensed psychiatrist to annotate structured symptom summaries. To support the development of an explainable framework, annotators are instructed

to construct symptom summaries based on the interviewer-participant dialogue transcripts from the DAIC-WOZ dataset.

Annotation Guidelines. Following discussions with the licensed psychiatrist, we define four required annotation components based on dialogue content. First, annotators are required to determine whether the participant has a history of depression. Second, annotators identify the participant's depressive symptoms, using the original DAIC-WOZ PHQ-8 item scores as a reference for the symptom dimensions to be examined, rather than as direct text to be copied into the summary. For each symptom indicated by the PHQ-8, annotators must locate corresponding explanatory evidence spans from the dialogue whenever possible; item-level symptom descriptions may also be adjusted when the dialogue clearly contradicts the self-reported score (e.g., a participant exhibits *occasional concentration difficulties* during the interview not reflected in the self-report). Third, annotators extract supporting evidence from the dialogue to infer potential underlying causes of the identified symptoms. Finally, based on the assembled dialogue-grounded symptom evidence, annotators provide conservative screening-oriented recommendations. Together, these four components constitute the final structured symptom summary.

Summarization Format. Each symptom summary consists of four components: ① Screening for a history of depression. If Ellie asks the participant whether they have been diagnosed with depression and the participant responds “Yes,” the first sentence begins with: “*The participant reported a history of depression.*” ② Assessment of the scale-listed symptoms. Conversely, if the participant does not have a history of depression, the summary starts directly with the second sentence: “*The participant primarily experiences [symptoms].*” ③ Evaluation of possible underlying causes. The third sentence provides an explanation for the onset or exacerbation of these symptoms, such as: “*These symptoms may be related to their [underlying causes].*” ④ Formulation of action plans. This component provides a conservative screening-oriented recommendation for clinician-readable reporting, for example: “*If symptoms*

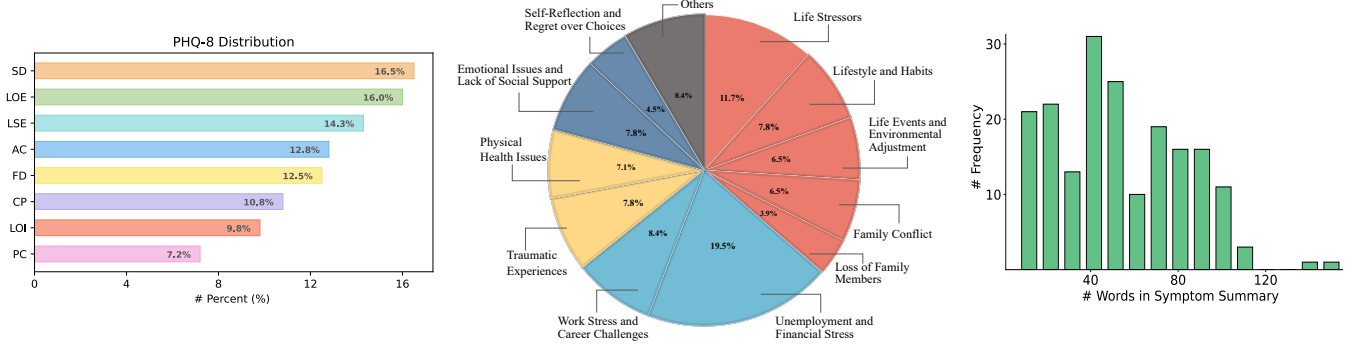


Fig. 3. Analysis of the annotated symptom summaries in Explain-DAIC. **Left:** PHQ-8 symptom mention distribution (SD: Sleeping Disorder, LOE: Lack of Energy, LSE: Low Self-Esteem, AC: Appetite Changes, FD: Feeling Down, CP: Concentration Problem, LOI: Lack of Interest, PC: Psychomotor Changes). **Middle:** Frequently identified causes. **Right:** Symptom summary length distribution.

persist or worsen, it is recommended that the participant seek further professional evaluation from a psychiatrist or psychologist.”

Annotation Procedure. As illustrated in Fig. 2 (a), each annotator receives the participant’s self-reported PHQ-8 questionnaire results and the corresponding semi-structured interview transcript for constructing the reference annotations. Annotators independently perform structured annotation following the established guidelines. The structured annotations are then converted into coherent natural language descriptions, which serve as explanatory evidence for downstream depression recognition models.

Annotation Consistency Verification. For certain symptoms, such as *lack of energy*, the severity to which they are expressed by the participant in dialogue can be ambiguous, and the underlying causes may be interpreted differently due to annotators’ subjective judgments. To account for this variability, three trained annotators independently annotate symptom summaries, and we adopt Fleiss’ κ [64] to measure the inter-annotator agreement. Two annotations are considered consistent if their Intersection-over-Union (IoU) score exceeds 0.5, following the partial-match convention widely used in NLP span annotation [65].

We report κ in two stages. The pre-consensus first-pass $\kappa = 0.71$ reflects raw agreement before any re-annotation, which corresponds to substantial agreement under the Landis–Koch scale [66]. Samples below the consistency threshold are re-annotated through structured disagreement resolution until the required level is achieved; in total, 19.1% (9/47) of dialogues undergo this adjudication. The post-consensus κ rises to **0.84**, indicating near-perfect agreement. We additionally report per-component κ across the four summary fields: history of depression ($\kappa_h = 0.91$), symptom items ($\kappa_s = 0.82$), contextual stressors ($\kappa_c = 0.79$), and action plan ($\kappa_a = 0.93$). Agreement is highest on structured, factual fields (history, action plan) and lowest on contextual stressors, where subjective judgment about what constitutes a stressor introduces unavoidable variability, consistent with the ambiguity illustrated by the *lack of energy* example above.

D. Dataset Analysis

Fig. 3 analyzes the annotated symptom summaries in Explain-DAIC. The **left panel** shows the distribution of PHQ-8 symptom mentions across participants, with SD (Sleeping Disorder) and LOE (Lack of Energy) appearing most frequently. This pattern is consistent with the clinical observation that sleep disturbance and fatigue are common symptoms in depression-related interviews, including cases with mild or subthreshold severity. The **middle panel** summarizes the frequently identified contextual causes: Unemployment and Financial Stress (19.5%), Life Stressors (11.7%), and Work Stress and Career Challenges (8.4%) are the top three categories, collectively accounting for 39.6%. These results suggest that economic, life-event, and occupational stressors constitute the dominant contextual factors in Explain-DAIC. The **right panel** shows that most symptom summaries contain fewer than 80 words, reflecting the dataset’s predominance of minimal-to-mild depressive symptoms while still leaving room for longer summaries when participants report more complex symptom profiles or underlying causes.

E. Relationship to Original DAIC-WOZ Labels

Explain-DAIC augments DAIC-WOZ with PHQ-guided, dialogue-grounded symptom evidence profiles, but does not redefine the original depression-recognition task. We retain the original DAIC-WOZ binary labels for downstream recognition to preserve comparability with prior MDRC studies. The PHQ-8 item scores are used as annotation references and training-time supervision for PHQ-aware representation learning; at inference time, the PHQ-8 item vector is not used as an input feature. Thus, the task is framed as interview-based prediction of a standardized screening-derived label, rather than independent clinical diagnosis.

The refined symptom-evidence totals remain highly aligned with the original self-reported PHQ-8 totals (Pearson $r = 0.985$). If these refined item-level scores were summed and thresholded at the conventional PHQ-8 cutoff, only 6 of 189 participants would cross the binary boundary. We therefore keep the original DAIC-WOZ labels unchanged and treat these cases as construct-discordant examples between self-reported screening scores and interview-derived evidence.

Stage 1: PHQ-aware Symptom Summarization

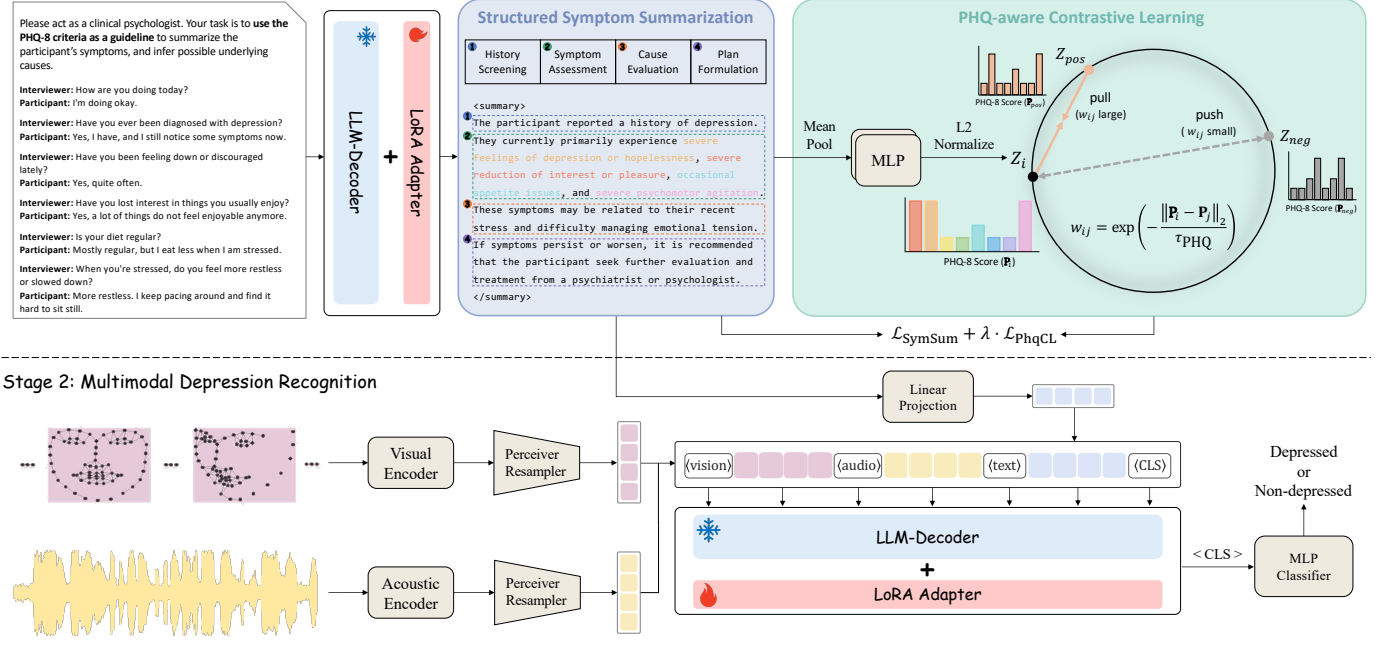


Fig. 4. The overview of our proposed PHQ-aware contrastive multimodal learning (PhqCML) method.

IV. METHODOLOGY

A. Formulation of the Proposed Task

Each clinical interview sample contains an interviewer-participant dialogue and the participant's nonverbal acoustic and visual cues. Let $u_i = (u_i^t, u_i^a, u_i^v)$ denote the multimodal input of the i -th interview, where u_i^t is the dialogue transcript, u_i^a is the audio recording, and u_i^v denotes the participant's visual information. Conventional MDRC aims to predict the binary depression label y_i directly from these multimodal cues:

$$\hat{y}_i^{\text{DR}} = f_{\text{MDRC}}(u_i^t, u_i^a, u_i^v), \quad (1)$$

where $\hat{y}_i^{\text{DR}} \in \{\text{depressed}, \text{non-depressed}\}$. Following the standard DAIC-WOZ protocol, the binary label is defined by the original self-reported PHQ-8 cutoff. Accordingly, Explain-MDRC is framed as interview-based prediction of a standardized screening label, coupled with PHQ-8-aligned symptom summarization, rather than independent clinical diagnosis.

Explain-MDRC extends this formulation by introducing a clinician-readable intermediate output. Let $\mathcal{D} = \{(u_i, s_i, y_i)\}_{i=1}^N$ denote the Explain-DAIC corpus, where s_i is the PHQ-8-aligned structured symptom summary annotated from the transcript. Given an interview, Explain-MDRC first generates a PHQ-8-aligned symptom summary:

$$\hat{s}_i = g_{\text{sum}}(u_i^t), \quad (2)$$

and then predicts the depression label by combining the generated summary with nonverbal multimodal cues:

$$\hat{y}_i^{\text{DR}} = f_{\text{Explain-MDRC}}(\hat{s}_i, u_i^a, u_i^v), \quad (3)$$

where $\hat{s}_i$ serves as explicit symptom-level evidence associated with the final prediction. Thus, the proposed task couples

PHQ-8-aligned symptom summarization with multimodal depression recognition, making the explanatory evidence part of the prediction pipeline rather than a post-hoc interpretation.

B. Framework Overview

As shown in Fig. 4, PhqCML adopts a unified two-stage architecture. The first stage trains an LLM to produce a PHQ-8-aligned symptom summary, with PHQ-aware contrastive learning (PhqCL) using each participant's eight-dimensional PHQ score vector as training-time clinical supervision on the summary representation, so that summaries from participants with similar PHQ-8 profiles are pulled together and those with dissimilar profiles pushed apart in the LLM's hidden space. In the second stage, the resulting PHQ-aware summary is fused with acoustic and visual features in an LLM-based architecture to produce the final depression prediction.

C. Stage 1: PHQ-aware Symptom Summarization

In clinical practice, depression assessment is anchored to a small set of standardized symptom dimensions, most commonly the eight items of the PHQ-8 scale. A useful symptom summary should read as clear clinical text and faithfully reflect a participant's profile across these eight dimensions. We designed the first stage of Explain-MDRC around this dual requirement.

A natural starting point is to fine-tune an LLM to generate the structured summary using the standard negative log-likelihood (NLL) objective. This objective shapes what the summary says at the token level, but it does not constrain where the resulting summary lies in the model's internal representation space. In practice, this means that two participants with very different PHQ-8 profiles can produce summaries

whose internal representations are almost indistinguishable. Such collapse weakens the subsequent multimodal depression recognition stage, where the summary representation is fused with acoustic and visual cues, and it also blurs the clinical signal that the summary is intended to carry.

To mitigate this weakly structured representation space, we treat each summary as a point in an eight-dimensional clinical space defined by the PHQ-8 scale during training. We then add a geometric training objective: it encourages summary representations from participants with similar PHQ-8 profiles to be close, while separating representations from clinically dissimilar profiles through the contrastive denominator. The first-stage model is trained with two complementary signals: a standard NLL loss that supervises summary generation at the surface level, and a PHQ-aware contrastive loss, denoted $\mathcal{L}_{\text{PhqCL}}$, that supervises the clinical geometry of the underlying representation.

1) **Symptom Summary Generation:** We employ an LLM as the generation backbone and instruction-tune it to produce structured symptom summaries that follow the four-component clinical format introduced in Section III-C. The generated summary is delimited by the markers `<summary>` and `</summary>`, which also serve as anchors for representation extraction in the contrastive objective described below.

Formally, given a participant’s interview dialogue $D_i = \{d_1, \dots, d_n\}$ and an instruction prompt I as input, the LLM generates the symptom summary s_i autoregressively:

$$s_i = \text{LLM}_\theta(I \oplus D_i), \quad (4)$$

where $\oplus$ denotes text concatenation and θ denotes the trainable parameters. The generation objective is the standard token-level NLL loss:

$$\mathcal{L}_{\text{SymSum}} = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^{|s_i|} \log P_\theta(s_{i,j} | s_{i,<j}, I \oplus D_i), \quad (5)$$

where $s_{i,j}$ denotes the j -th token of the target summary and $s_{i,<j}$ denotes all preceding tokens.

2) **PHQ-aware Contrastive Learning:** The NLL objective shapes the model’s output distribution at the token level, but does not constrain how participants’ summarization representations are organized in the LLM’s hidden space. We therefore introduce a contrastive objective that operates directly on summarization representations, supervised during training by participants’ eight-dimensional PHQ-8 score vectors $\mathbf{p}_i \in \{0, 1, 2, 3\}^8$. A representation space that respects this clinical proximity provides more clinically informative features for the subsequent multimodal recognition stage.

Summarization representation. For each training example i , we extract the final-layer LLM hidden states corresponding to the generated symptom-summarization span, apply mean pooling, and map the pooled vector into a contrastive embedding space using a two-layer MLP projection head g_ϕ followed by ℓ_2 normalization:

$$\mathbf{z}_i = \frac{g_\phi(\text{MeanPool}(\mathbf{h}_i^{(\text{sum})}))}{\|g_\phi(\text{MeanPool}(\mathbf{h}_i^{(\text{sum})}))\|_2} \in \mathbb{S}^{d-1}, \quad (6)$$

where $\mathbf{h}_i^{(\text{sum})}$ denotes the final-layer hidden states over the summarization span, and $\mathbb{S}^{d-1}$ is the unit hypersphere in $\mathbb{R}^d$.

Contrastive learning with continuous PHQ supervision (PhqCL). These PHQ-8 profiles provide a structured clinical coordinate system for measuring participant similarity. We therefore encourage representations to reflect graded proximity in the PHQ-8 space, so that participants with similar symptom profiles are pulled closer while clinically dissimilar profiles are separated through the contrastive denominator. For each anchor i and candidate j , we define a clinical similarity weight

$$w_{ij} = \exp\left(-\frac{\|\mathbf{p}_i - \mathbf{p}_j\|_2}{\tau_{\text{PHQ}}}\right) \in (0, 1], \quad (7)$$

where τ_{PHQ} controls the decay rate of clinical similarity. This formulation replaces hard positive-pair assignment with graded PHQ-8 similarity, making PhqCL a soft clinical extension of supervised contrastive learning [67].

The PHQ-aware contrastive loss is then a soft-weighted InfoNCE:

$$\mathcal{L}_{\text{PhqCL}} = -\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \frac{\sum_{j \in \mathcal{N}(i)} w_{ij} \log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_j / \tau)}{\sum_{k \in \mathcal{N}(i)} \exp(\mathbf{z}_i \cdot \mathbf{z}_k / \tau)}}{\sum_{j \in \mathcal{N}(i)} w_{ij}}, \quad (8)$$

where $\mathcal{B}$ is the mini-batch, $\mathcal{N}(i)$ excludes the anchor, and τ is the contrastive temperature. Candidates with higher PHQ-8 similarity receive larger positive weights, while clinically distant candidates contribute mainly through the contrastive denominator.

3) **Joint optimization:** We jointly optimize symptom summarization generation and PHQ-aware representation alignment with the combined loss:

$$\mathcal{L} = \mathcal{L}_{\text{SymSum}} + \lambda \cdot \mathcal{L}_{\text{PhqCL}}, \quad (9)$$

where λ is a trade-off hyperparameter.

D. Stage 2: Multimodal Depression Recognition

After obtaining the PHQ-aware symptom summaries, we perform multimodal depression recognition by fusing textual, acoustic, and visual representations within a unified LLM architecture. We first extract unimodal embeddings $\{\mathbf{X}^t, \mathbf{X}^a, \mathbf{X}^v\}$ from the three modalities:

- **Textual:** We extract symptom-aware textual embeddings from the LLM hidden states corresponding to the generated symptom summary s_i . We use the final transformer layer hidden states $\mathbf{X}^t \in \mathbb{R}^{l_t \times d_t}$ as the textual representation, where l_t denotes the summary token length and d_t is the LLM hidden dimension. These embeddings are extracted offline from the first-stage LLM and projected into the second-stage LLM’s embedding space via a linear projection layer.
- **Acoustic:** We adopt Wav2Vec2², fine-tuned on Common Voice 6.1, to obtain frame-level acoustic embeddings $\mathbf{X}^a \in \mathbb{R}^{l_a \times d_a}$, where l_a denotes the number of audio frames.
- **Visual:** Given a 3D facial landmark sequence $\mathcal{F} = (f_1, \dots, f_n)$, we employ PointMLP [68] to extract visual embeddings $\mathbf{X}^v \in \mathbb{R}^{l_v \times d_v}$, where l_v denotes the length of the facial sequence.

²<https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english>

Since the acoustic and visual sequences are much longer than the textual summary, directly feeding them into the LLM would be computationally prohibitive. We therefore employ a Perceiver Resampler-based modality encoder [69] for each non-textual modality, which compresses variable-length feature sequences into a fixed set of N_m latent tokens via cross-attention between learnable queries and the input features, and projects them to the LLM hidden dimension:

$$\mathbf{H}^m = \text{PerceiverResampler}(\mathbf{X}^m) \in \mathbb{R}^{N_m \times d_h}, \quad m \in \{a, v\}, \quad (10)$$

where d_h denotes the LLM hidden size.

We then construct a unified multimodal input sequence by concatenating the compressed modality tokens with special delimiter tokens and the projected textual embeddings:

$$\mathbf{Z} = [\langle \text{vision} \rangle \mathbf{H}^v \langle \text{audio} \rangle \mathbf{H}^a \langle \text{text} \rangle \mathbf{H}^t \langle \text{CLS} \rangle], \quad (11)$$

where $\mathbf{H}^t$ denotes the projected textual hidden states, and the special tokens $\langle \text{audio} \rangle$, $\langle \text{vision} \rangle$ serve as modality boundary markers. The entire sequence is fed as input embeddings into an LLM, which performs cross-modal fusion through its self-attention layers:

$$\mathbf{O} = \text{LLM}_\phi(\mathbf{Z}), \quad (12)$$

Finally, the hidden state at the $\langle \text{CLS} \rangle$ position is extracted and passed through a classification head to obtain the predicted depression recognition label:

$$\hat{y}_i^{\text{DR}} = \text{Softmax}(\text{MLP}(\mathbf{O}_{[\text{CLS}]})), \quad (13)$$

where $\mathbf{O}_{[\text{CLS}]}$ denotes the LLM's last-layer hidden state at the $\langle \text{CLS} \rangle$ token position. The model is optimized with the standard cross-entropy loss:

$$\mathcal{L}_{\text{DR}} = -\frac{1}{N} \sum_{i=1}^N \sum_{k=1}^C y_{i,k}^{\text{DR}} \log(\hat{y}_{i,k}^{\text{DR}}), \quad (14)$$

where C denotes the number of recognition-label classes.

V. EXPERIMENTS

A. Evaluation Metrics

We evaluate the framework on symptom summarization and depression recognition. **Symptom summarization.** To evaluate the quality of generated PHQ-aware symptom summaries, we adopt standard automatic metrics widely used in text summarization [58]. Specifically, we report ROUGE-1, ROUGE-2, and ROUGE-L to measure n-gram overlap and structural similarity, BLEU to assess surface-level precision, and BERTScore to capture semantic consistency between generated summaries and references. **Depression recognition.** We follow previous studies [34], [70] and evaluate the model on binary classification (depressed vs. non-depressed) using precision, recall, and Macro-F1. Moreover, to ensure consistency with previous studies [71], [72] for experiments on the development set, we report the F1 scores for the depressed class and the control class instead of precision and recall.

B. Baseline Systems

General-purpose LLM API baselines. We include five strong general-purpose LLM API baselines: *GPT-5.5*, *Claude-Opus-4.8*, *Gemini-3.5-Flash*, *GLM-5.2*, and *Qwen3.7-Max*. Each API baseline is evaluated under both zero-shot and four-shot settings using the interview transcript as input. These models provide a direct reference for whether prompt-based general-purpose reasoning is sufficient for Explain-MDRC without task-specific training.

General multimodal and MDRC baselines. *MuT* [73] is a transformer-based fusion model that uses cross-modal attention to capture interactions among unaligned modalities. *MISA* [74] encourages the shared space to capture cross-modal commonality while preserving private modality cues. These two methods provide strong general-purpose multimodal fusion references. *D-vlog* [34] models depressive signals from naturalistic vlog-style behavior, making it a representative depression-oriented multimodal baseline. *EmoGRU* [26] emphasizes emotional dynamics in audio-text sequences through recurrent modeling. *HiQuE* [75] introduces hierarchical question embeddings to better exploit the structure of clinical interviews. *DMPF* [70] focuses on speech-based depression detection by decoupling complementary feature perspectives before fusion. *MTRB* [76] treats doctor-patient interviews as multi-instance inputs, reducing the reliance on a single global interview representation.

LLM training and ablation baselines. We implement LLM-based training baselines with *Llama-3.2-3B-Instruct* and *Qwen3-4B-Instruct-2507*. *DepRec* directly maps raw multimodal inputs to the depression label, testing whether an instruction-tuned LLM can perform end-to-end recognition without explicit symptom supervision. *MuTsk* keeps the same two-stage symptom summarization and depression recognition pipeline as *PhqCML*, but removes PHQ-aware contrastive learning; it therefore isolates the contribution of the proposed representation objective.

Development-split references. Because prior DAIC-WOZ studies commonly report development-set results, we also evaluate on the development split of Explain-DAIC for direct comparison with earlier work. We report strong general-purpose LLM API references, including *Qwen3-Omni-30B*, *DeepSeek-v3.2*, and *GPT-5.5*. We also include specialized MDRC systems. *ω -GCN* [77] builds a node-weighted graph convolutional network over transcript units to exploit conversational structure. *Wav2VecDep* [78] transfers wav2vec2.0 speech representations to depression detection. *AudLadLLM* [79] augments LLM reasoning with acoustic landmarks, linking speech cues with textual context. *SEGA* [72] constructs structural-element graphs for LLM-assisted depression analysis. *AgentMental* [48] formulates assessment as an interactive multi-agent process. *PsyGAT* [53] uses psychologically grounded graph attention to model clinically relevant cues.

C. Experimental Settings

Symptom summarization setup. We select *Llama-3.2-3B-Instruct* and *Qwen3-4B-Instruct-2507* as the LLM backbones,

TABLE II
COMPARISON RESULTS ON THE EXPLAIN-DAIC DATASET. “–” INDICATES THAT THE CORRESPONDING METHOD DOES NOT PRODUCE SYMPTOM SUMMARIES. THE BEST RESULTS ARE MARKED IN BOLD.

Methods	Symptom Summarization					Depression Recognition		
	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERTScore	Precision	Recall	Macro-F1
GPT-5.5 (zero-shot)	38.34	18.43	28.50	15.85	73.89	52.38	78.57	70.41
Claude-Opus-4.8 (zero-shot)	29.63	12.79	20.73	8.67	68.92	66.67	42.86	68.34
Gemini-3.5-Flash (zero-shot)	43.65	24.48	35.86	19.45	76.92	60.00	42.86	66.43
GLM-5.2 (zero-shot)	33.51	15.44	23.99	11.54	70.65	66.67	42.86	68.34
Qwen3.7-Max (zero-shot)	43.20	23.11	34.75	21.52	75.83	70.00	50.00	72.02
GPT-5.5 (four-shot)	42.68	21.36	33.07	19.66	76.28	54.55	85.71	72.99
Claude-Opus-4.8 (four-shot)	43.02	22.93	33.13	18.24	75.44	66.67	42.86	68.34
Gemini-3.5-Flash (four-shot)	49.21	30.47	42.03	22.85	78.90	66.67	42.86	68.34
GLM-5.2 (four-shot)	43.69	24.52	34.44	20.76	75.66	62.50	35.71	64.39
Qwen3.7-Max (four-shot)	49.84	30.20	42.10	23.78	78.39	63.64	50.00	70.03
MuT (ACL’ 2019) [73]	–	–	–	–	–	73.00	74.00	74.00
MISA (ACM MM’ 2020) [74]	–	–	–	–	–	74.00	77.00	74.00
D-vlog (AAAI’ 2020) [34]	–	–	–	–	–	73.00	72.00	73.00
EmoGRU (ICASSP’ 2022) [26]	–	–	–	–	–	75.00	78.00	75.00
HiQuE (CIKM’ 2024) [75]	–	–	–	–	–	78.00	80.00	79.00
DMPF (TAFFC’ 2025) [70]	–	–	–	–	–	79.92	81.30	80.30
MTRB (Sci. Rep. 2025) [76]	–	–	–	–	–	89.00	86.00	88.00
DepRec-Llama3.2	–	–	–	–	–	83.33	89.39	84.07
DepRec-Qwen3	–	–	–	–	–	83.93	86.77	84.87
MuTsk-Llama3.2	54.50	35.97	48.73	26.46	80.40	86.84	92.42	88.33
MuTsk-Qwen3	55.40	35.59	48.95	24.36	81.25	91.18	95.45	92.78
PhqCML-Llama3.2 (Ours)	56.00	37.66	50.50	27.20	81.33	91.18	95.45	92.78
PhqCML-Qwen3 (Ours)	56.79	38.64	50.91	27.16	81.80	93.75	96.97	95.10

with the maximum input length set to 8,192 tokens and the maximum generation length set to 4,096 tokens. We employ LoRA with rank $r = 16$ and scaling factor $\alpha = 32$. In PhqCL, the projection head g_ϕ is a two-layer MLP with GELU activation and dropout 0.1, mapping the LLM hidden dimension to a 128-dimensional contrastive embedding. We use a FIFO memory bank of size $M = 256$ to store recent (z_j, p_j) pairs, which is updated without gradients. The batch size is 2, the learning rate is $5e-5$, and the number of training epochs is 7. **Depression recognition setup.** We reuse the same LLM backbones with LoRA. For the Perceiver Resampler, both audio and vision use 32 latent queries. The maximum sequence length for audio and vision inputs is set to the mean plus three standard deviations of the training set. The batch size is 4, the learning rate is $5e-5$, and the number of training epochs is 5. Both stages are optimized with AdamW. **Reproducibility.** All experiments are conducted using PyTorch 2.6.0 with CUDA 11.8 on an NVIDIA A800 GPU. We run each trainable experiment with three random seeds and report the mean performance. The code, configurations, prompts, and processed annotation files will be released upon publication, subject to DAIC-WOZ data-use and redistribution constraints.

D. Main Results

Table II reports the performance of different methods on the proposed Explain-DAIC dataset. Under our evaluation protocol, PhqCML achieves the strongest results across the reported metrics. The results support three main observations.

General-purpose LLM baselines remain far behind task-specific PHQ-aware learning. The strongest API baseline reaches 72.99% Macro-F1 with *GPT-5.5* under four-shot prompting, which remains far below *PhqCML*. This result suggests that prompting general-purpose large models alone is insufficient for reliable Explain-MDRC.

PhqCML improves recognition while exposing intermediate evidence. Beyond the API comparison, *PhqCML-Qwen3* reaches 95.10% Macro-F1, outperforming the strongest prior MDRC baseline under the compared protocol (*MTRB*) by 7.10 percentage points while producing PHQ-8-aligned symptom summaries alongside its predictions. Compared with single-stage LLM baselines that directly map raw dialogue to labels, *PhqCML* also improves Macro-F1 by 8.71 points over *DepRec-Llama3.2* and 10.23 points over *DepRec-Qwen3*. These results suggest that explicit symptom summarization provides useful intermediate supervision rather than merely serving as an explanation layer.

PHQ-aware contrastive learning consistently benefits both stages. Compared with *MuTsk*, which uses the same two-stage pipeline without the PHQ-aware contrastive learning loss, *PhqCML* improves both symptom summarization and depression recognition across Llama3.2 and Qwen3 backbones. On summarization, the largest gains appear on ROUGE-2 (+1.69 for Llama3.2 and +3.05 for Qwen3), indicating better coverage of clinically meaningful symptom phrases. The same benefit carries into recognition, where *PhqCML* improves Macro-F1 by 4.45 and 2.32 points over the corresponding *MuTsk* variants. Because the pipeline is otherwise unchanged, these gains isolate the effect of PHQ-aware contrastive learning: aligning summary representations with each participant’s PHQ-8 profile makes the intermediate summaries more useful for downstream multimodal fusion.

E. In-depth Analysis of PHQ-aware Contrastive Learning (PhqCL)

To understand what *PhqCL* contributes to *PhqCML* beyond the aggregate gains reported above, we conduct representation probing, attention-shift analysis, comparison with supervised contrastive learning, and hyperparameter sensitivity analysis.

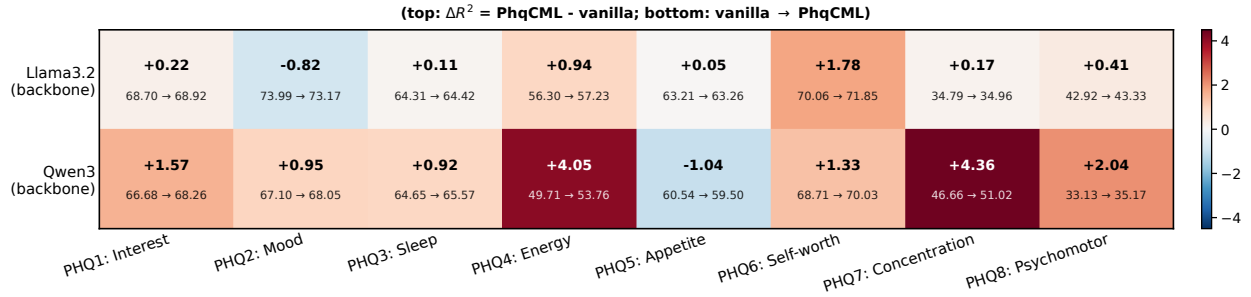


Fig. 5. Per-item PHQ-8 linear probing of symptom-summary embeddings from *PhqCML* and its vanilla counterpart, *MulTsk*. For each backbone, we train eight ridge regressors on frozen mean-pooled summary embeddings to predict individual PHQ-8 item scores on the test set. Each cell shows the change in test R^2 , $\Delta R^2 = R^2_{\text{PhqCML}} - R^2_{\text{MulTsk}}$, on the top, and the absolute $\text{MulTsk} \rightarrow \text{PhqCML}$ R^2 values on the bottom. Positive values and warmer colors indicate that PHQ-aware contrastive learning makes the representations more linearly predictive of the corresponding symptom item.

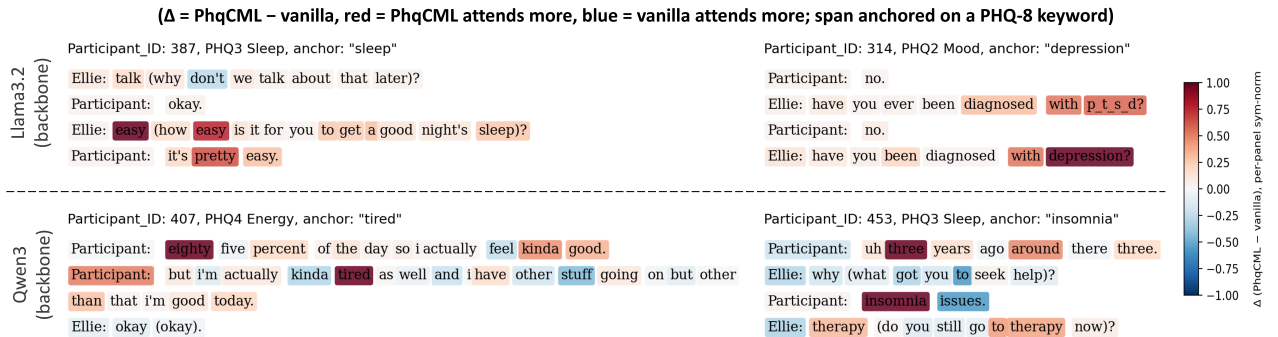


Fig. 6. Token-level attention shift of *PhqCML* versus vanilla (*MulTsk*) around PHQ-8 symptom keywords, averaged over the last four decoder layers, all heads, and all summary query positions.

PhqCML improves per-item PHQ-8 probing R^2 over vanilla. We first test whether PhqCL makes summary embeddings encode item-level symptom information. We fit eight ridge regressors from frozen summary embeddings to PHQ-8 item scores. As shown in Fig. 5, *PhqCL* improves probing R^2 on seven of eight items for both backbones, with mean gains of +0.36 and +1.77 percentage points for Llama3.2 and Qwen3. The largest improvements appear on clinically salient items such as Concentration and Energy for Qwen3, suggesting that PHQ-aware supervision makes summary representations more discriminative at the symptom level.

PhqCL redirects attention towards PHQ-8 symptom keywords. We then inspect whether the representation objective changes which transcript tokens the model attends to during summarization. We compare token-level attention patterns between *PhqCML* and the vanilla model. Fig. 6 shows that positive attention shifts concentrate around PHQ-8 symptom anchors and nearby clinical context, while back-channel and filler tokens receive lower relative attention. This supports the view that *PhqCL* improves summary quality by reweighting transcript evidence toward symptom-relevant cues.

Comparison with supervised contrastive learning (SupCL). To test whether continuous PHQ-8 supervision is more informative than binary class supervision, Fig. 7 compares *PhqCL* with *SupCL*, which uses binary depression labels to form hard contrastive pairs. *PhqCL* matches or exceeds *SupCL* on all metrics, with clearer advantages on summary fidelity (e.g., Qwen3 ROUGE-2 +3.05 vs. +1.13)

and Llama3.2 recognition (Macro-F1 +4.45 vs. +2.19). This indicates that continuous PHQ-8 profiles provide finer clinical supervision than coarse class labels.

Hyperparameter analysis. Finally, we evaluate the sensitivity of PhqCL to its two key hyperparameters. Fig. 8 sweeps the loss weight λ and PHQ-distance temperature τ_{PHQ} . Both backbones prefer $\tau_{\text{PHQ}} = 1$, with different optimal loss weights (0.07 for Llama3.2 and 0.15 for Qwen3). Performance is stable around the selected settings, whereas overly large λ over-regularizes generation and extreme temperatures weaken PHQ-aware similarity modeling.

F. Modality Ablation Study

TABLE III
MODALITY ABLATION OF PHQCML.

Methods	Depression Recognition		
	Precision	Recall	Macro-F1
PhqCML-Llama3.2	91.18	95.45	92.78
– w/o Audio	88.89	93.94	90.52
– w/o Vision	88.89	93.94	90.52
– w/o Audio, Vision	86.84	92.42	88.33
PhqCML-Qwen3	93.75	96.97	95.10
– w/o Audio	91.18	95.45	92.78
– w/o Vision	91.18	95.45	92.78
– w/o Audio, Vision	88.89	93.94	90.52

To evaluate the contribution of each input modality, we conduct ablation studies on both *PhqCML-Llama3.2* and

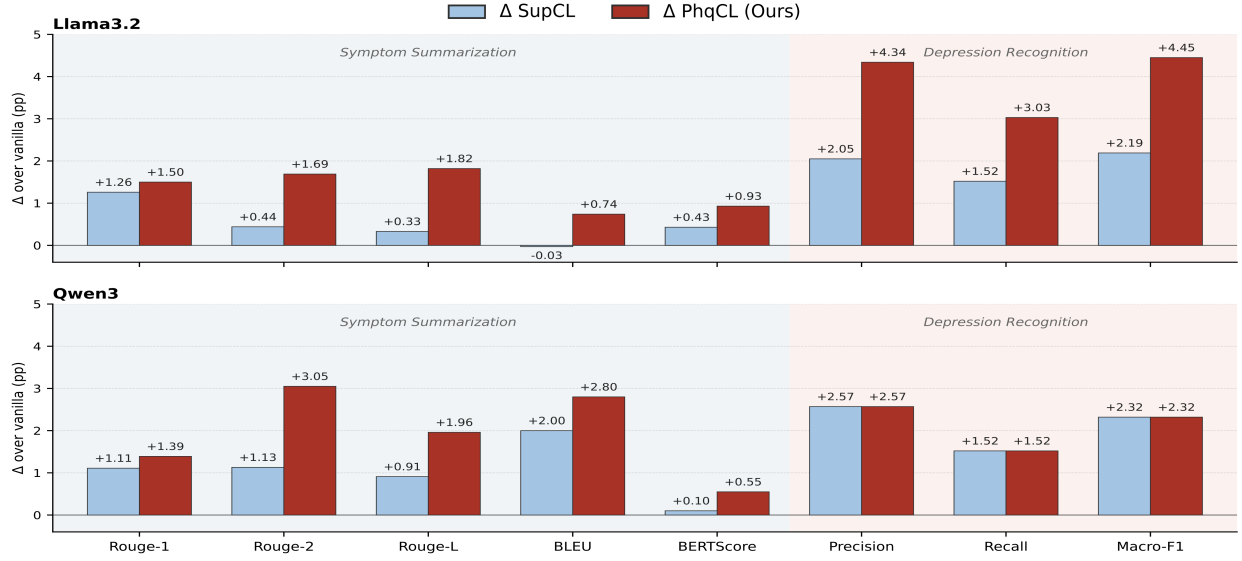


Fig. 7. Comparison of *PhqCL* (ours) against supervised contrastive learning (*SupCL*). Bars report the absolute gain (Δ , in percentage points) over the vanilla.

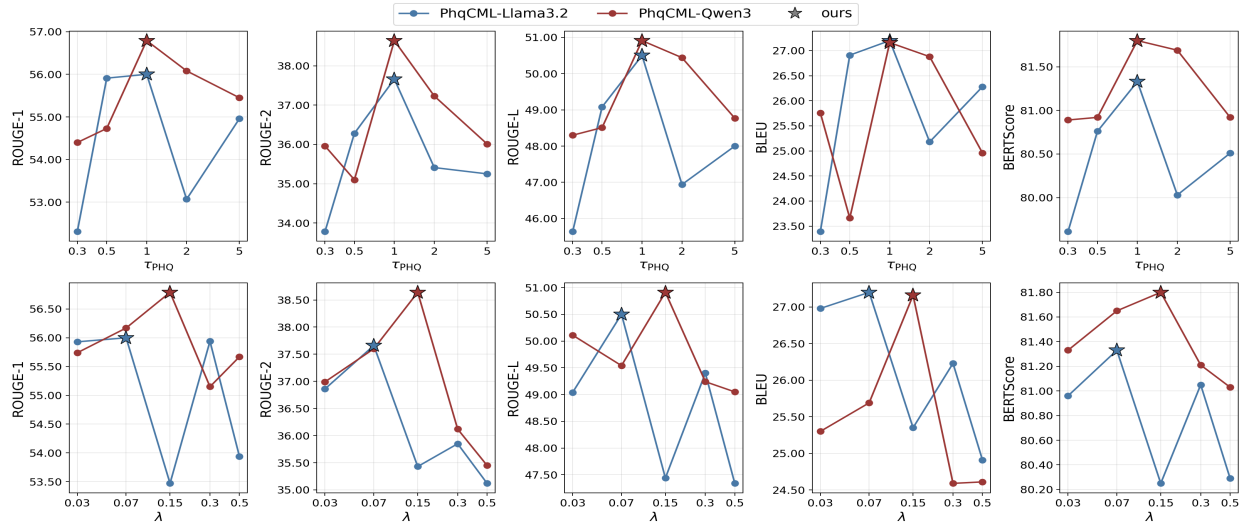


Fig. 8. Hyperparameter analysis of *PhqCML*. Stars (*) mark each backbone’s chosen configuration, lying at the peak of its corresponding curve.

PhqCML-Qwen3 by progressively removing acoustic and visual modalities (Table III). Removing audio or vision each causes a Macro-F1 drop of 2.32 points for Qwen3 and 2.26 points for Llama3.2; removing both modalities leads to drops of 4.58 and 4.45 points, respectively. These consistent degradations suggest that nonverbal modalities provide complementary cues (prosody, vocal quality, and facial expression) beyond the textual stream for depression recognition.

G. Performance on the DAIC-WOZ Development Split

Since Explain-DAIC is derived from DAIC-WOZ, where prior studies report on the development split, we evaluate *PhqCML* on this split for direct comparison (Table IV). General-purpose LLMs lag behind specialized MDRC models: even *GPT-5.5* reaches only 76.31 Macro-F1, 13 points below *PsyGAT* (89.99), suggesting that prompting alone is insufficient for robust depression recognition. Both *PhqCML*

TABLE IV
COMPARISON AGAINST BASELINES ON THE EXPLAIN-DAIC DEVELOPMENT SET. “DEPRE” / “CONTR” DENOTE F1 SCORES FOR THE POSITIVE (DEPRESSED) AND NEGATIVE (CONTROL) CLASSES.

Methods	Depression Recognition		
	Depre	Contr	Macro-F1
Qwen3-Omni-30B (four-shot)	65.00	74.07	69.54
DeepSeek-v3.2 (four-shot)	68.75	70.59	69.67
GPT-5.5 (four-shot)	68.75	83.75	76.31
ω -GCN (INTER_SPEECH’23) [77]	78.26	89.36	83.81
Wav2VecDep (Sci.Rep.’24) [78]	—	—	79.00
AudLadLLM (EMNLP’24) [79]	—	—	83.30
SEGA (NAACL’24) [72]	84.62	90.91	87.76
AgentMental (AAAI’26) [48]	80.00	92.00	86.00
PsyGAT (arXiv’26) [53]	84.96	93.02	89.99
PhqCML-Llama3.2 (Ours)	92.31	95.45	93.88
PhqCML-Qwen3 (Ours)	92.31	95.45	93.88

TABLE V
CROSS-DATASET DEPRESSION PREDICTION ON CMDC BY INFERRING
PHQ-STYLE SYMPTOM SCORES FROM DIALOGUE.

Methods	Depression Recognition		
	Precision	Recall	Macro-F1
IIFDD (Inf.Fus.'24) [80]	95.00	93.00	93.00
VT-MSD (Sci.Rep.'25) [81]	97.52	96.64	97.08
DepressInstruct (Inf.Fus.'26) [82]	—	—	98.18
PhqCML-SymSum-Llama3.2	90.23	89.42	89.80
PhqCML-SymSum-Llama3.2*	91.35	91.35	91.35
PhqCML-SymSum-Qwen3	92.48	93.27	92.86
PhqCML-SymSum-Qwen3*	94.83	97.12	95.79

variants achieve the strongest reported results: *PhqCML-Llama3.2* and *PhqCML-Qwen3* each reach 93.88 Macro-F1, surpassing *PsyGAT* by 3.89 points. Beyond this quantitative gain, *PhqCML* generates a structured PHQ-8-aligned symptom summary alongside each prediction, providing clinician-readable intermediate evidence that prior classification-only methods lack.

H. Cross-dataset Depression Prediction via Dialogue-based PHQ-style Assessment

To examine whether the dialogue-based symptom-assessment ability of *PhqCML* transfers beyond Explain-DAIC, we conduct an external cross-dataset experiment on CMDC [25], a Mandarin PHQ-9-based semi-structured clinical-interview corpus for binary depression recognition. Unlike conventional depression-recognition models that directly map an interview to a binary label, *PhqCML-SymSum* first performs PHQ-style symptom assessment from the dialogue transcript by generating a structured symptom summary and estimating item-level symptom scores. The depression label can then be inferred from the estimated scale profile. This design allows us to test whether the model has learned transferable clinical-scale-guided reasoning rather than only dataset-specific label cues.

CMDC differs from Explain-DAIC in language, interview protocol, and the acquisition and feature format of acoustic and visual modalities. As a result, the second-stage multimodal depression-recognition head trained on Explain-DAIC cannot be directly applied to CMDC. We therefore focus this experiment on the first stage of *PhqCML*, namely *PhqCML-SymSum*, and evaluate whether its dialogue-based PHQ assessment ability can transfer to CMDC transcripts.

Specifically, we translate the Mandarin CMDC transcripts into English and apply the Explain-DAIC-trained *PhqCML-SymSum* without any CMDC-specific fine-tuning. Given only the interview transcript, the model generates a structured symptom summary and estimates PHQ-style item scores. Cases with an estimated total score greater than 9 are classified as depressed, following the commonly used PHQ cutoff for moderate depressive symptoms [14]. Because Explain-DAIC is annotated with PHQ-8 whereas CMDC follows PHQ-9, we further construct *PhqCML-SymSum** by adding a prompt-level instruction for the PHQ-9 suicidality item. This variant adapts

the output format to the PHQ-9 setting but does not retrain the model on CMDC.

Table V compares our method with recent strong CMDC baselines. Without CMDC-specific fine-tuning, *PhqCML-SymSum-Qwen3* achieves 92.86 Macro-F1, which is comparable to *IIFDD* with 93.00 Macro-F1. After incorporating the suicidality item, *PhqCML-SymSum-Qwen3** reaches 95.79 Macro-F1 and 97.12 Recall, narrowing the gap to CMDC-specialized multimodal systems such as *VT-MSD*. These results suggest that *PhqCML-SymSum* can transfer its dialogue-based PHQ-style assessment ability across datasets, languages, and screening scales. Rather than directly fitting a CMDC-specific classifier, the model infers depression labels through an intermediate symptom-score profile, providing a more clinically structured prediction pathway.

I. Clinical Expert Evaluation

Automatic metrics may not fully capture whether generated summaries are clinically readable and evidence-grounded. We therefore recruited three licensed psychiatrists for a blinded paired preference evaluation on the full Explain-DAIC test set. Following prior work [58], the experts compared anonymized summaries generated by *PhqCML-Qwen3* and *GPT-5.5* along three clinically oriented dimensions: completeness, correctness, and conciseness. Each paired comparison was accompanied by the corresponding interview transcript, allowing experts to assess whether the generated symptom descriptions were supported by the source dialogue. **Completeness** measures whether the summary captures the major symptom domains and contextual evidence expressed in the interview, rather than only repeating the most explicit statements. **Correctness** focuses on factual and clinical faithfulness, including whether symptom descriptions are grounded in the participant’s dialogue and avoid over-interpreting ambiguous cues. **Conciseness** evaluates whether the summary suppresses redundant, off-topic, or weakly supported details while preserving information needed for assessment.

For each test case, the two system outputs were presented as paired summaries without revealing model identities, and each psychiatrist independently assigned a preference score from -2 to 2 . The psychiatrists were instructed to base their judgments only on the source dialogue and the anonymized summaries, focusing on whether the generated symptom descriptions were supported by the interview evidence and clinically useful for review. Positive scores indicate that *PhqCML-Qwen3* is preferred, negative scores indicate that *GPT-5.5* is preferred, and zero denotes no clear preference. For statistical testing, we first averaged the three psychiatrists’ preference scores for each test case and each evaluation dimension, and then applied a one-sample Wilcoxon signed-rank test to the resulting case-level mean preference scores against zero. Results are reported as mean $\pm$ SEM. This design complements automatic overlap metrics by asking whether the generated explanations satisfy practical clinical requirements: a useful summary should cover the relevant PHQ-8 symptom evidence, avoid unsupported symptom attributions, and remain concise enough for efficient review.

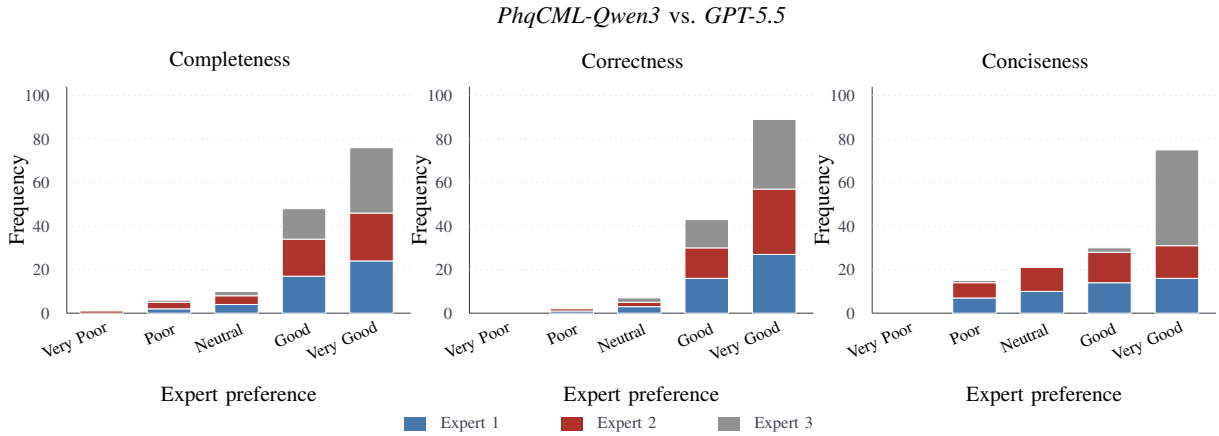


Fig. 9. Per-expert frequency distributions of preference ratings for *PhqCML-Qwen3* against *GPT-5.5* across completeness, correctness, and conciseness. Stacked bars show the frequency of ratings assigned by each expert. Preference scores range from -2 (Very Poor; strongly favoring *GPT-5.5*) to 2 (Very Good; strongly favoring *PhqCML-Qwen3*).

TABLE VI
EXPERT PREFERENCE SCORES FOR PHQCML-QWEN3 AGAINST GPT-5.5
(MEAN $\pm$ SEM). ** DENOTES $p < 0.001$ USING A ONE-SAMPLE
WILCOXON SIGNED-RANK TEST ON CASE-LEVEL MEAN PREFERENCE
SCORES.

Models	Completeness	Correctness	Conciseness
<i>PhqCML</i> vs. <i>GPT-5.5</i>	1.35 ± 0.14 **	1.54 ± 0.08 **	1.17 ± 0.12 **

Under this clinically oriented protocol, the results in Table VI show consistently positive expert preferences for *PhqCML* over *GPT-5.5* in completeness, correctness, and conciseness ($p < 0.001$). The largest margin appears in correctness, the dimension that most directly tests whether symptom descriptions are grounded in interview evidence and avoid unsupported clinical inference. This indicates that the expert preference is not merely driven by longer or more fluent summaries, but by explanations that more faithfully connect patient-reported evidence with PHQ-8 symptom interpretations. These findings support the central motivation of Explain-MDRC: explicit PHQ-aware training improves the clinical readability and evidence grounding of the intermediate explanation, while also improving downstream recognition scores.

Fig. 9 further examines the distribution of expert preferences for each psychiatrist, complementing the aggregate case-level test with a per-expert view. Across the three psychiatrists, positive preferences for *PhqCML* account for 88.0% of judgments on completeness, 93.6% on correctness, and 74.5% on conciseness. The distributions are concentrated in the “Good” and “Very Good” categories, especially for correctness, where very few cases favor *GPT-5.5*. This pattern suggests that *PhqCML* more consistently avoids hallucinated or clinically unsupported symptom descriptions, which is important for clinician-reviewable mental-health assessment tools. Conciseness shows a slightly broader distribution than the other two dimensions, indicating that some generated summaries may still include nonessential details. Nevertheless, the overall preference remains positive, suggesting that the structured

PHQ-aware formulation improves focus without sacrificing symptom coverage.

VI. LIMITATIONS

Despite these promising results, several limitations remain. **First**, Explain-MDRC does not yet incorporate external clinical knowledge sources, such as psychiatric knowledge graphs or ontologies, which may improve the credibility and domain alignment of symptom interpretation. **Second**, the current framework follows a static learning paradigm; future work should explore continuous learning or human-in-the-loop adaptation so clinicians can verify outputs and provide corrective feedback, consistent with human-centered XAI principles that emphasize target-user understanding and calibrated trust [21]. **Third**, real-world deployment requires broader safety evaluation, including bias, unintended output, spurious contextual shortcuts, and misuse monitoring in high-stakes clinical contexts [83]. Accordingly, the current framework should be interpreted as a screening-oriented research prototype rather than a deployable diagnostic system; prospective clinical validation with clinician-adjudicated diagnostic labels is required before real-world use. **Finally**, although this work anchors explainability around PHQ-8, extending the framework to multiple validated scales, other mental disorders, and multilingual or culturally diverse settings would improve generalizability and equitable access [84].

VII. CONCLUSION

This work presents Explain-MDRC, an explainable framework that reorients multimodal depression recognition in clinical interviews from direct label prediction toward PHQ-8-aligned symptom reasoning. By constructing Explain-DAIC and introducing structured symptom summaries as an intermediate representation, Explain-MDRC explicitly links interview evidence to clinically meaningful symptom dimensions before integrating acoustic and visual cues for final recognition. *PhqCML* instantiates this framework with PHQ-aware contrastive learning, which shapes summary representations according to item-level clinical profiles and provides clinically

structured evidence for multimodal recognition. Experiments and clinical expert evaluations show that this design improves recognition performance while providing structured, interpretable symptom-level evidence. These results support clinically structured intermediate reasoning as a practical direction for transparent AI-assisted depression recognition.

REFERENCES

- [1] Z. Lian, L. Sun, Y. Ren, H. Gu, H. Sun, L. Chen, B. Liu, and J. Tao, "Merbench: A unified evaluation benchmark for multimodal emotion recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.
- [2] X. Wang and C. Zong, "Learning emotion category representation to detect emotion relations across languages," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [3] J. Li, X. Wang, and Z. Zeng, "Tracing intricate cues in dialogue: Joint graph structure and sentiment dynamics for multimodal emotion recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [4] J. N. Acosta, G. J. Falcone, P. Rajpurkar, and E. J. Topol, "Multimodal biomedical ai," *Nature Medicine*, 2022.
- [5] T. Tu, M. Schaekermann, A. Palepu, K. Saab, J. Freyberg, R. Tanno, A. Wang, B. Li, M. Amin, Y. Cheng *et al.*, "Towards conversational diagnostic artificial intelligence," *Nature*, pp. 1–9, 2025.
- [6] L. Hansen, R. Rocca, A. Simonsen, L. Olsen, A. Parola, V. Bliksted, N. Ladegaard, D. Bang, K. Tylén, E. Weed *et al.*, "Speech-and text-based classification of neuropsychiatric conditions in a multidagnostic setting," *Nature Mental Health*, 2023.
- [7] J. J. Liu, B. Borsari, Y. Li, S. X. Liu, Y. Gao, X. Xin, S. Lou, M. Jensen, D. Garrido-Martín, T. L. Verplaetse *et al.*, "Digital phenotyping from wearables using ai characterizes psychiatric disorders and identifies genetic associations," *Cell*, 2024.
- [8] Y. Gong and C. Poellabauer, "Topic modeling based multi-modal depression detection," in *Proceedings of the 7th annual workshop on Audio/Visual emotion challenge*, 2017, pp. 69–76.
- [9] A. Ray, S. Kumar, R. Reddy, P. Mukherjee, and R. Garg, "Multi-level attention network using text, audio and video for depression prediction," in *Proceedings of the 9th international on audio/visual emotion challenge and workshop*, 2019, pp. 81–88.
- [10] K. Mao, W. Zhang, D. B. Wang, A. Li, R. Jiao, Y. Zhu, B. Wu, T. Zheng, L. Qian, W. Lyu *et al.*, "Prediction of depression severity based on the prosodic and semantic features with bidirectional lstm and time distributed cnn," *IEEE transactions on affective computing*, vol. 14, no. 3, pp. 2251–2265, 2022.
- [11] Z. Zhou, Y. Guo, S. Hao, and R. Hong, "Hierarchical multifeature fusion via audio-response-level modeling for depression detection," *IEEE transactions on computational social systems*, vol. 10, no. 5, pp. 2797–2805, 2022.
- [12] W. Wu, C. Zhang, and P. C. Woodland, "Self-supervised representations in speech-based depression detection," in *ICASSP. IEEE*, 2023, pp. 1–5.
- [13] X. Zhang, H. Liu, K. Xu, Q. Zhang, D. Liu, B. Ahmed, and J. Epps, "When llms meets acoustic landmarks: An efficient approach to integrate speech into large language models for depression detection," in *Proc. of EMNLP*, 2024.
- [14] K. Kroenke, R. L. Spitzer, and J. B. Williams, "The phq-9: validity of a brief depression severity measure," *Journal of general internal medicine*, vol. 16, no. 9, pp. 606–613, 2001.
- [15] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. Williams, J. T. Berry, and A. H. Mokdad, "The phq-8 as a measure of current depression in the general population," *Journal of affective disorders*, 2009.
- [16] T. Nguyen, A. Yates, A. Zirikly, B. Desmet, and A. Cohan, "Improving the generalizability of depression detection by leveraging clinical questionnaires," in *ACL. ACL*, 2022, pp. 8446–8459.
- [17] B. Yang, P. Wang, M. Cao, X. Zhu, S. Wang, R. Ni, and C. Yang, "Uncertainty-aware label contrastive distribution learning for automatic depression detection," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 2, pp. 2979–2989, 2023.
- [18] S. Kundu, "Ai in medicine must be explainable," *Nature Medicine*, 2021.
- [19] D. W. Joyce, A. Kormilitzin, K. A. Smith, and A. Cipriani, "Explainable artificial intelligence for mental health through transparency and interpretability for understandability," *npj Digital Medicine*, 2023.
- [20] T. Savage, A. Nayak, R. Gallo, E. Rangan, and J. H. Chen, "Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine," *npj Digital Medicine*, 2024.
- [21] Y. Rong, T. Leemann, T.-T. Nguyen, L. Fiedler, P. Qian, V. Unhelkar, T. Seidel, G. Kasneci, and E. Kasneci, "Towards human-centered explainable ai: A survey of user studies for model explanations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 4, pp. 2104–2122, 2024.
- [22] J. J. Silverman, M. Galanter, M. Jackson-Triche, D. G. Jacobs, J. W. Lomax II, M. B. Riba, L. D. Tong, K. E. Watkins *et al.*, "The American Psychiatric Association practice guidelines for the psychiatric evaluation of adults," *American Journal of Psychiatry*, vol. 172, no. 8, pp. 798–802, 2015.
- [23] J. Gratch, R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella *et al.*, "The distress analysis interview corpus of human and computer interviews," in *Proc. of LREC*, 2014.
- [24] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proc. of the 6th international workshop on audio/visual emotion challenge*, 2016.
- [25] B. Zou, J. Han, Y. Wang, R. Liu, S. Zhao, L. Feng, X. Lyu, and H. Ma, "Semi-structural interview-based chinese multimodal depression corpus towards automatic preliminary screening of depressive disorders," *IEEE Transactions on Affective Computing*, 2022.
- [26] Y. Shen, H. Yang, and L. Lin, "Automatic depression detection: An emotional audio-textual corpus and a gru/bilstm-based model," in *Proc. of ICASSP*, 2022.
- [27] S. Yadav, J. Chauhan, J. P. Sain, K. Thirunarayan, A. Sheth, and J. Schumm, "Identifying depressive symptoms from tweets: Figurative language enabled multitask learning framework," in *COLING*, 2020, pp. 696–709.
- [28] S. Gupta, A. Agarwal, M. Gaur, K. Roy, V. Narayanan, P. Kumaraguru, and A. Sheth, "Learning to automate follow-up question generation using process knowledge for depression triage on reddit posts," in *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, 2022, pp. 137–147.
- [29] Z. Zhang, S. Chen, M. Wu, and K. Zhu, "Symptom identification for interpretable detection of multiple mental disorders on social media," in *EMNLP*, 2022, pp. 9970–9985.
- [30] A. Pérez, J. Parapar, Á. Barreiro, and S. Lopez-Larrosa, "Bdi-sen: A sentence dataset for clinical symptoms of depression," in *SIGIR*, 2023, pp. 2996–3006.
- [31] T. Gui, L. Zhu, Q. Zhang, M. Peng, X. Zhou, K. Ding, and Z. Chen, "Cooperative multimodal approach to depression detection in twitter," in *Proc. of AAAI*, 2019.
- [32] A.-S. Uban, B. Chulvi, and P. Rosso, "Explainability of depression detection on social media: From deep learning models to psychological interpretations and multimodality," in *Early Detection of Mental Health Disorders by Social Media Monitoring: The First Five Years of the eRisk Project*, 2022.
- [33] S. Yadav, C. Caragea, C. Zhao, N. Kumari, M. Solberg, and T. Sharma, "Towards identifying fine-grained depression symptoms from memes," in *Proc. of ACL*, 2023.
- [34] J. Yoon, C. Kang, S. Kim, and J. Han, "D-vlog: Multimodal vlog dataset for depression detection," in *Proc. of AAAI*, 2022.
- [35] M. A. L. Sawadogo, F. Pala, G. Singh, I. Selmi, P. Puteaux, and A. Othmani, "Ptds in the wild: a video database for studying post-traumatic stress disorder recognition in unconstrained environments," *Multimedia Tools and Applications*, 2024.
- [36] L. He, K. Chen, J. Zhao, Y. Wang, E. Pei, H. Chen, J. Jiang, S. Zhang, J. Zhang, Z. Wang *et al.*, "Lmvd: A large-scale multimodal vlog dataset for depression detection in the wild," *Information Fusion*, p. 103632, 2025.
- [37] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "Avec 2013: the continuous audio/visual emotion and depression recognition challenge," in *Proc. of the 3rd ACM international workshop on Audio/visual emotion challenge*, 2013.
- [38] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "Avec 2014: 3d dimensional affect and depression recognition challenge," in *Proc. of the 4th international workshop on audio/visual emotion challenge*, 2014.
- [39] H. Cai, Z. Yuan, Y. Gao, S. Sun, N. Li, F. Tian, H. Xiao, J. Li, Z. Yang, X. Li *et al.*, "A multi-modal open dataset for mental-disorder analysis," *Scientific Data*, 2022.

- [40] A. Rinaldi, J. E. F. Tree, and S. Chaturvedi, "Predicting depression in screening interviews from latent categorization of interview prompts," in *ACL*, 2020, pp. 7–18.
- [41] T. Zhang, K. Yang, H. Alhuzali, B. Liu, and S. Ananiadou, "Phq-aware depressive symptoms identification with similarity contrastive learning on social media," *Information Processing & Management*, vol. 60, no. 5, p. 103417, 2023.
- [42] S. Chen, Z. Zhang, M. Wu, and K. Zhu, "Detection of multiple mental disorders from social media with two-stream psychiatric experts," in *EMNLP*, 2023, pp. 9071–9084.
- [43] G. Huang, J. Li, H. Lu, M. Guo, and S. Chen, "Rethinking inconsistent context and imbalanced regression in depression severity prediction," *IEEE Transactions on Affective Computing*, vol. 15, no. 4, pp. 2154–2168, 2024.
- [44] K. Yang, T. Zhang, Z. Kuang, Q. Xie, J. Huang, and S. Ananiadou, "Mentallama: interpretable mental health analysis on social media with large language models," in *WWW*, 2024, pp. 4489–4500.
- [45] X. Xu, B. Yao, Y. Dong, S. Gabriel, H. Yu, J. Hendler, M. Ghassemi, A. K. Dey, and D. Wang, "Mental-llm: Leveraging large language models for mental health prediction via online text data," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 1, pp. 1–32, 2024.
- [46] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto, "Stanford alpaca: An instruction-following llama model," 2023.
- [47] Q. Yang, Z. Wang, H. Chen, S. Wang, Y. Pu, X. Gao, W. Huang, S. Song, and G. Huang, "Psychogat: A novel psychological measurement paradigm through interactive fiction games with llm agents," in *ACL*, 2024, pp. 14470–14505.
- [48] J. Hu, A. Wang, Q. Xie, Z. Li, H. Ma, and D. Guo, "Agentmental: An interactive multi-agent framework for explainable and adaptive mental health assessment," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 37, 2026, pp. 31050–31058.
- [49] M. Sadeghi, R. Richer, B. Egger, L. Schindler-Gmelch, L. H. Rupp, F. Rahimi, M. Berking, and B. M. Eskofier, "Harnessing multimodal approaches for depression detection using large language models and facial expressions," *npj Mental Health Research*, 2024.
- [50] Y. Li, S. Shao, M. Milling, and B. W. Schuller, "Large language models for depression recognition in spoken language integrating psychological knowledge," *Frontiers in Computer Science*, vol. 7, 2025.
- [51] G. Y. Kebe, J. M. Girard, E. Liebenthal, J. Baker, F. De la Torre, and L.-P. Morency, "LLAMADRS: Evaluating open-source LLMs on real clinical interviews—to reason or not to reason?" *arXiv preprint arXiv:2501.03624*, 2025.
- [52] C. Wang, J. Liang, P. Cao, J. Yang, and O. R. Zaiane, "IDRL: An individual-aware multimodal depression-related representation learning framework for depression diagnosis," *arXiv preprint arXiv:2603.11644*, 2026.
- [53] R. R. Vyalla, K. Prasad, A. Anand, E. Cambria, S. Ji, F. S. Alamri, and Z. Wang, "Psychologically-grounded graph modeling for interpretable depression detection," *arXiv preprint arXiv:2604.24126*, 2026.
- [54] S. Teng, J. Liu, H. Sun, Y. Li, S. Chai, R. Hou, T. Tateyama, L. Lin, and Y.-W. Chen, "Dynamic summary generation for interpretable multimodal depression detection," *arXiv preprint arXiv:2604.11334*, 2026.
- [55] K. Zhang, R. Zhou, E. Adhikarla, Z. Yan, Y. Liu, J. Yu, Z. Liu, X. Chen, B. D. Davison, H. Ren *et al.*, "A generalist vision-language foundation model for diverse biomedical tasks," *Nature Medicine*, 2024.
- [56] K. Krishna, S. Khosla, J. P. Bigham, and Z. C. Lipton, "Generating soap notes from doctor-patient conversations using modular summarization techniques," in *Proc. of ACL*, 2021.
- [57] G. Michalopoulos, K. Williams, G. Singh, and T. Lin, "Medicalsum: A guided clinical abstractive summarization model for generating medical reports from patient-doctor conversations," in *Proc. of EMNLP (Findings)*, 2022.
- [58] D. Van Veen, C. Van Uden, L. Blankemeier, J.-B. Delbrouck, A. Aali, C. Bluethgen, A. Pareek, M. Polacin, E. P. Reis, A. Seehofnerová *et al.*, "Adapted large language models can outperform medical experts in clinical text summarization," *Nature Medicine*, 2024.
- [59] A. Ghosh, A. Acharya, R. Jain, S. Saha, A. Chadha, and S. Sinha, "Clipsyntel: clip and llm synergy for multimodal question summarization in healthcare," in *Proc. of AAAI*, 2024.
- [60] K. Saab, T. Tu, W.-H. Weng, R. Tanno, D. Stutz, E. Wulczyn, F. Zhang, T. Strother, C. Park, E. Vedadi *et al.*, "Capabilities of gemini models in medicine," *arXiv preprint arXiv:2404.18416*, 2024.
- [61] C. Yuan, X. Rui, Y. Fan, Y. Fan, B. Zhong, J. Wang, W. Zhang, and T. Ruan, "Lclds: A logic-controlled discharge summary generation system supporting source attribution and expert review," in *ACL*, 2025, pp. 284–294.
- [62] B. Yao, C. Shi, L. Zou, L. Dai, M. Wu, L. Chen, Z. Wang, and K. Yu, "D4: a chinese dialogue dataset for depression-diagnosis-oriented chat," in *Proc. of EMNLP*, 2022.
- [63] M. E. Aragón, J. Parapar, and D. E. Losada, "Delving into the depths: Evaluating depression severity through bdi-biased summaries," in *Proc. of 9th Workshop on Computational Linguistics and Clinical Psychology*, 2024.
- [64] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological bulletin*, 1971.
- [65] Ö. Uzuner, B. R. South, S. Shen, and S. L. DuVall, "2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 552–556, 2011.
- [66] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, 1977.
- [67] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18661–18673, 2020.
- [68] X. Ma, C. Qin, H. You, H. Ran, and Y. Fu, "Rethinking network design and local geometry in point cloud: A simple residual mlp framework," in *Proc. of ICLR*, 2022.
- [69] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in neural information processing systems*, vol. 35, pp. 23716–23736, 2022.
- [70] M. Zhao, H. Gao, L. Zhao, Z. Wang, F. Wang, W. Zheng, J. Li, and C. Liu, "Decoupled multi-perspective fusion for speech depression detection," *IEEE Transactions on Affective Computing*, 2025.
- [71] M. Niu, K. Chen, Q. Chen, and L. Yang, "Hcag: A hierarchical context-aware graph attention model for depression detection," in *Proc. of ICASSP*, 2021.
- [72] Z. Chen, J. Deng, J. Zhou, J. Wu, T. Qian, and M. Huang, "Depression detection in clinical interviews with llm-empowered structural element graph," in *Proc. of NAACL*, 2024.
- [73] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. of ACL*, 2019.
- [74] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proc. of ACM MM*, 2020.
- [75] J. Jung, C. Kang, J. Yoon, S. Kim, and J. Han, "Hique: Hierarchical question embedding network for multimodal depression detection," in *Proc. of CIKM*, 2024.
- [76] X. Zhang, C. Li, W. Chen, J. Zheng, and F. Li, "Optimizing depression detection in clinical doctor-patient interviews using a multi-instance learning framework," *Scientific Reports*, vol. 15, no. 1, p. 6637, 2025.
- [77] S. Burdisso, E. Villatoro-Tello, S. R. Madikeri, and P. Motlíček, "Node-weighted graph convolutional network for depression detection in transcribed clinical interviews," in *Proc. of INTERSPEECH*, 2023.
- [78] X. Zhang, X. Zhang, W. Chen, C. Li, and C. Yu, "Improving speech depression detection using transfer learning with wav2vec 2.0 in low-resource environments," *Scientific Reports*, 2024.
- [79] X. Zhang, H. Liu, K. Xu, Q. Zhang, D. Liu, B. Ahmed, and J. Epps, "When LLMs meets acoustic landmarks: An efficient approach to integrate speech into large language models for depression detection," in *Proc. of EMNLP*, 2024.
- [80] J. Chen, Y. Hu, Q. Lai, W. Wang, J. Chen, H. Liu, G. Srivastava, A. K. Bashir, and X. Hu, "Iifdd: Intra and inter-modal fusion for depression detection with multi-modal information from internet of medical things," *Information Fusion*, vol. 102, p. 102017, 2024.
- [81] Z. Xu, Y. Gao, F. Wang, L. Zhang, L. Zhang, J. Wang, and J. Shu, "Depression detection methods based on multimodal fusion of voice and text," *Scientific Reports*, vol. 15, no. 1, p. 21907, 2025.
- [82] D. Li, L. Ding, Z. Wang, and K. Zhao, "Depressinstruct: Instruction tuning of large speech-language models for depression detection," *Information Fusion*, vol. 130, p. 104077, 2026.
- [83] D. Yang, K. Yang, H. Kuang, Z. Chen, Y. Wang, and L. Zhang, "Towards context-aware emotion recognition debiasing from a causal demystification perspective via de-confounded training," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 10663–10680, 2024.
- [84] M. B. First, H. A. Pincus, J. B. Levine, J. B. Williams, B. Ustun, and R. Peele, "Clinical utility as a criterion for revising psychiatric diagnoses," *American Journal of Psychiatry*, 2004.