

Unveiling the Landscape of Clinical Depression Assessment: From Behavioral Signatures to Psychiatric Reasoning

Zhuang Chen¹, Guanqun Bi², Wen Zhang³, Jiawei Hu⁴,
Aoyun Wang¹, Xiya Xiao⁵, Kun Feng⁶, Minlie Huang²

¹School of Computer Science and Engineering, Central South University

²CoAI Group, DCST, IAI, BNRI, Tsinghua University ³University of International Relations

⁴Central China Normal University ⁵Lingxin AI ⁶Yuquan Hospital, Tsinghua University

zhchen18@foxmail.com aihuang@tsinghua.edu.cn

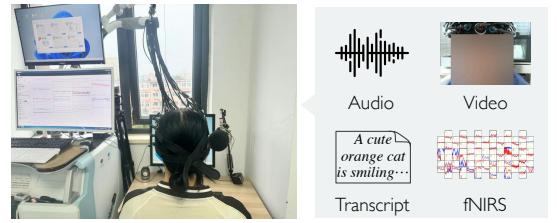
Abstract

Depression is a widespread mental disorder that affects millions worldwide. While automated depression assessment shows promise, most studies rely on limited or non-clinically validated data, and often prioritize complex model design over real-world effectiveness. In this paper, we aim to unveil the landscape of clinical depression assessment. We introduce C-MIND, a *clinical neuropsychiatric multimodal diagnosis* dataset collected over two years from real hospital visits. Each participant completes three structured psychiatric tasks and receives a final diagnosis from expert clinicians, with informative audio, video, transcript, and functional near-infrared spectroscopy (fNIRS) signals recorded. Using C-MIND, we first analyze *behavioral signatures* relevant to diagnosis. We train a range of classical models to quantify how different tasks and modalities contribute to diagnostic performance, and dissect the effectiveness of their combinations. We then explore whether LLMs can perform *psychiatric reasoning* like clinicians and identify their clear limitations in realistic clinical settings. In response, we propose to guide the reasoning process with clinical expertise and consistently improves LLM diagnostic performance by up to 10% in Macro-F1 score. We aim to build an infrastructure for clinical depression assessment from both data and algorithmic perspectives, enabling C-MIND to facilitate grounded and reliable research for mental healthcare.

1 Introduction

Depression is a widespread and serious mental disorder that places a heavy burden on individuals and public health systems worldwide. While automated assessment shows promise for offering objective and scalable support, its real-world clinical utility remains limited due to a lack of clinically grounded data (Cummins et al., 2015; Sarsam et al., 2024). Most widely used datasets rely on self-reported questionnaires rather than diagnoses made

Synchronized Multimodal Recordings



Psychiatric Task Battery



Gold-Standard Clinical Diagnosis

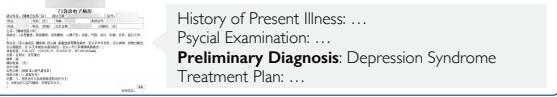


Figure 1: C-MIND integrates multimodal recordings from psychiatric tasks with clinical diagnosis.

by trained clinicians (Gratch et al., 2014; Tadesse et al., 2019). Even the few pioneering studies that include clinical diagnoses often suffer from small sample sizes (< 30 patients) and limited behavioral tasks or modalities (Cai et al., 2022; Zou et al., 2022). These constraints lead many studies to focus on sophisticated model design in controlled settings instead of addressing the full complexity of real clinical data. As a result, a clear picture of what effective automated clinical depression assessment entails has yet to emerge (Sarsam et al., 2024).

In this paper, we aim to unveil this landscape through a three-pronged investigation: 1) establishing a new, clinically grounded data foundation, 2) analyzing the core behavioral signatures, and 3) advancing clinically guided psychiatric reasoning for diagnosis. First, we introduce **C-MIND**: the Clinical Multimodal Neuropsychiatric Diagnosis dataset. Over a two-year period, we build this dataset from a real hospital setting. It comprises 169 participants who each complete three distinct psychiatric tasks, including Interview (Gratch et al.,

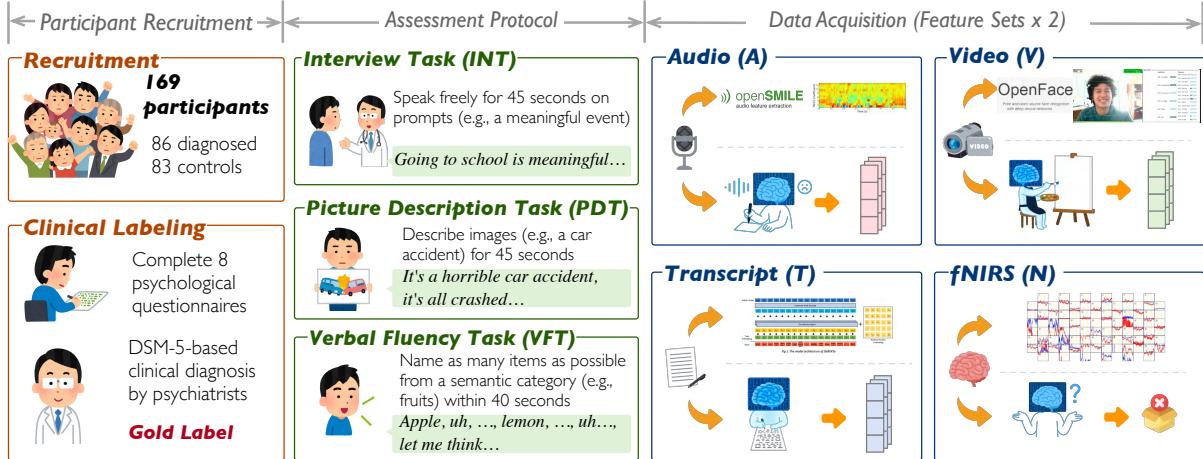


Figure 2: C-MIND collection pipeline, outlining participant recruitment, assessment protocol, and data acquisition.

2014), Picture Description (Ramponi et al., 2010b), and Verbal Fluency (Fossati et al., 2003b). We capture four synchronized modalities (Audio, Video, Transcript, and fNIRS (Cui et al., 2011)) for each session. Crucially, every participant receives a face-to-face diagnostic interview with senior psychiatrists, whose final clinical diagnosis, made according to DSM-5 (American Psychiatric Association, 2013) criteria, serves as the gold-standard ground truth. The dataset is further enriched with detailed medical records and a battery of eight psychometric questionnaires. C-MIND’s scale, clinical grounding, and richness in tasks and modalities far exceed previously available resources.

Leveraging C-MIND, we conduct an in-depth analysis of **behavioral signatures**, defined as observable patterns in speech, facial expression, and neural activity indicative of depressive states. We train a range of modeling backbones to systematically quantify the diagnostic value of different tasks and modalities, revealing that audio and video are the most informative, while the picture description task best elicits depressive markers. Fusing modalities (e.g., Audio+Video) or tasks (e.g., Interview+Picture Description) further enhances performance and robustness, providing clear empirical guidance for designing future assessment systems.

Beyond analyzing predictive signals, we explore whether Large Language Models (LLMs) perform **psychiatric reasoning**. We evaluate seven top-tier text and multimodal LLMs and find clear limitations in their ability to handle real-world clinical data. In response, we propose a novel method that guides the LLM’s reasoning process using structured clinical expertise. This approach significantly boosts diagnostic performance by up to 10% in

Macro-F1 score, demonstrating a promising direction for developing clinically informed computational models.

Our main contributions can be summarized as follows:

- We introduce C-MIND, a clinically validated depression diagnosis dataset with rich tasks and modalities.
- We provide a comprehensive analysis of behavioral signatures, offering clear, data-driven insights into the discriminative power of different tasks and modalities.
- We demonstrate the limitations of LLMs in clinical assessment and propose a novel psychiatric reasoning mechanism that significantly boosts performance.

We believe this work builds a critical infrastructure for the field and provides a blueprint for developing computational systems that are not only effective, but also clinically grounded and trustworthy.

2 C-MIND Collection

We present **C-MIND**, a Clinical Multimodal Neuropsychiatric Diagnosis dataset designed for depression assessment. To ensure ecological validity, data quality, and clinical reliability, we follow a comprehensive collection protocol. Below, we describe participant recruitment, assessment procedures, and data acquisition in detail.

2.1 Participant Recruitment

We recruited participants from December 2022 to April 2025 at the psychiatric department of a university-affiliated hospital. Recruitment was conducted through internal announcements. Volunteers who met the inclusion criteria and were eval-

| Statistic | Depression | Control | Total |
|-----------------------|--------------|-------------|-------------|
| Subject | 86 | 83 | 169 |
| Gender (M/F %) | 36.05/63.95 | 44.58/55.42 | 40.34/59.66 |
| Age (Mean±SD) | 33.49 ±16.47 | 32.47±10.90 | 32.99±13.98 |
| Duration (s) | 171.84 | 125.01 | 152.16 |
| Word Count | 392 | 519 | 445 |

Table 1: Detailed statistics of C-MIND.

uated by a chief psychiatrist together with an associate chief psychiatrist according to DSM-5 (American Psychiatric Association, 2013) were invited to participate after providing written informed consent.

All procedures receive full approval from the university’s Institutional Review Board (IRB), and strict measures are in place to protect participant confidentiality. The final cohort consists of 169 participants, including 86 individuals diagnosed with Major Depressive Disorder (MDD) and 83 healthy controls (HC). Table 1 presents detailed statistics of the C-MIND cohort, including group size, gender distribution, age, average speech duration, and word count. For future public release, we will follow IRB-approved protocols to ensure responsible data sharing, including strict de-identification and a request-based access process.

2.2 Assessment Protocol

The assessment protocol for each participant includes two main parts: a formal face-to-face clinical diagnosis (used to obtain clinically validated labels), and a series of psychiatric tasks (used to collect rich multimodal behavioral signatures). Due to space constraints, we provide detailed experimental materials, guidelines, and procedures in the Technical Appendix.

2.2.1 Clinical Diagnosis

Participants are interviewed by a clinical team comprising a chief and an associate chief psychiatrist, each with over ten years of experience. The team conducts a face-to-face diagnostic interview and makes a high-confidence diagnosis based on DSM-5 criteria. A detailed medical record is maintained for each participant. Participants also complete a battery of eight psychometric questionnaires, including: *HAMD* (Hamilton, 1960), *HAMA* (Hamilton, 1959), *SDS* (Zung, 1965), *SAS* (Zung, 1971), *PSQI* (Buysse et al., 1989), *16PF* (Cattell et al., 1970), *SCL-90* (Derogatis, 1977), and *HCL-32* (Angst et al., 2005). These

instruments assess depressive symptoms, anxiety, sleep quality, and personality traits. In this study, we use only the clinical diagnosis as the ground-truth label; questionnaire data are reserved for future work.

2.2.2 Psychiatric Tasks

All tasks take place in a quiet, controlled laboratory with ambient noise below 60dB. Participants sit in front of a monitor displaying instructions and are asked to remain seated, minimize movement, and maintain a fixed distance from the microphone (approx. 20 cm). We design three structured tasks that elicit cognitive and emotional markers of depression:

- **Interview Task (INT):** Participants speak for 45 seconds in response to autobiographical prompts. This task elicits emotional expression and narrative patterns indicative of depression (e.g., negative sentiment, frequent first-person use) (Rinaldi et al., 2020). Prompts include: “something that made you angry,” “a meaningful event,” and “your favorite food.”
- **Picture Description Task (PDT):** Participants describe a given image for 45 seconds. This task captures visual interpretation and emotional valence. Depressed individuals may show negative bias or limited detail (Ramponi et al., 2010a). Images include: “a cat,” “a car accident,” and “a spaceship.”
- **Verbal Fluency Task (VFT):** Participants list items in a semantic category (e.g., “fruits”) within 40 seconds. This evaluates semantic memory and executive function, which are often impaired in depression (Akiyama et al., 2018; Fossati et al., 2003b). Categories include: “four-legged animals,” “fruits,” “cities,” and “vegetables.”

In total, each participant completes ten tests across the three tasks. We record four synchronized modalities during the psychiatric tasks. We use a studio microphone to capture high-quality audio (44.1kHz, 24-bit, WAV). A camera records video (640x480, 30fps), and a functional near-infrared spectroscopy (fNIRS) device measures blood oxygenation changes in the prefrontal cortex. Each of the 10 tasks is recorded separately in audio. Using timestamps, we segment the continuous video and fNIRS recordings to align all modalities. Transcripts are generated using a commercial speech recognition system and are manually proofread by human annotators to ensure accuracy.

| Availability | Dataset | Language | Subj. (MDD/HC) | Tasks | Modalities | Ground Truth Labels |
|--------------------|--------------------------------|----------|----------------|----------------|----------------|--------------------------------------|
| No | Oizys (Lin et al., 2022) | Chinese | 103 (56/47) | READ | A | C.D., HAMD-17 |
| | Guo et al. (2021) | Chinese | 208 (104/104) | INT, READ, PDT | A, V | PHQ-9, BDI |
| | Liu et al. (2021) | Chinese | 50 (25/25) | INT | A | BDI |
| | DEPAC (Tasnim et al., 2022) | English | 552 (134/418) | VFT, PDT, READ | A | PHQ-9, GAD-7 |
| Yes (w/o C.D.) | DAIC-WOZ (Gratch et al., 2014) | English | 142 (42/100) | INT | A, V, T | PHQ-8 |
| | EATD (Shen et al. 2022) | Chinese | 162 (30/132) | INT | A, T | SDS |
| Yes (with C.D.) | MODMA (Cai et al., 2022) | Chinese | 53 (24/29) | INT, READ, PDT | A, EEG | C.D., PHQ-9 |
| | CMDC (Zou et al., 2022) | Chinese | 78 (26/52) | INT | A, V, T | C.D. ⁻ , HAMD-17, PHQ-9 |
| | C-MIND (Ours) | Chinese | 169 (86/83) | INT, PDT, VFT | A, V, T, fNIRS | C.D. ⁺ , 8 Questionnaires |

Table 2: Comparison of depression datasets. C.D.=Clinical Diagnosis, INT=Interview, PDT=Picture Description, VFT=Verbal Fluency, READ=Reading, A=Audio, V=Video, T=Transcript, fNIRS=functional near-infrared spectroscopy. “C.D.⁻” means the control group is not confirmed by clinical diagnosis. “C.D.⁺” means our C-MIND further provides detailed medical records.

2.3 Comparison with Existing Datasets

To situate C-MIND within the current research landscape, we compare it with other depression datasets collected in controlled environments, while excluding those based on social media data (Yoon et al., 2022).

As summarized in Table 2, many existing datasets (e.g., DAIC-WOZ (Gratch et al., 2014)) rely on self-report instruments like PHQ-8 and lack clinical validation. While MODMA (Cai et al., 2022) and CMDC (Zou et al., 2022) include clinical diagnoses, they have smaller sample sizes (53/78 participants with only 24/26 patients, respectively), limited tasks, and fewer modalities.

In contrast, C-MIND goes far beyond existing resources in every aspect: it offers a larger and balanced sample size, more diverse psychiatric tasks (INT, PDT, VFT), richer modalities (Audio, Video, Text, fNIRS), expert-verified clinical diagnoses, and comprehensive psychometric data—making it a uniquely comprehensive and clinically grounded benchmark for depression research.

3 Methodology

As shown in Figure 3, we design a two-part methodological framework to uncover the mechanisms of clinical depression assessment: 1) modeling behavioral signatures across tasks and modalities, and 2) simulating psychiatric reasoning through guided LLMs. Due to space limitations, we present only core formulations here; details of models and prompts can be found in Technical Appendix.

3.1 Behavioral Signature Modeling

We define behavioral signatures as the informational cues derived from different psychiatric tasks

and data modalities. To quantify the diagnostic relevance of different behavioral cues, we follow a structured modeling pipeline. First, we extract feature representations from four modalities (audio, video, text, fNIRS) for each psychiatric task. Then, we train learning models to predict depression status based on feature sets. By evaluating each task-modality combination and their fusions, we aim to uncover which behavioral signatures most robustly reflect clinical diagnoses. The entire process is designed to simulate and analyze how observable behaviors align with psychiatric assessment.

Feature Representations We extract two feature sets from four synchronized modalities: audio, video, transcript, and fNIRS. **1) Classical Feature Set.** We apply standard feature extraction pipelines. Audio signals are encoded using OpenSmile’s eGeMAPS (88 dimensions), while video-based facial behavior is represented using OpenFace (4,963 dimensions), including action units, gaze, and head pose. Textual transcripts are embedded using DeBERTa. For fNIRS, statistical features are computed from 45 optical channels, resulting in a 630-dimensional vector. **2) Foundation Model Feature Set.** We also extract semantic-level embeddings using pretrained foundation models: Qwen2-Audio-7B (Chu et al., 2023) for audio, Qwen2.5-VL-72B for video (Bai et al., 2025), and Qwen3-235B-A22B for transcript (Yang et al., 2025). Each modality is segmented by task, encoded through the model’s final hidden layer, and globally max-pooled over time to yield fixed-length vectors (4096 dimensions for audio and transcript, 8192 for video). fNIRS remains represented by classical statistics due to the absence of public pre-trained encoders.

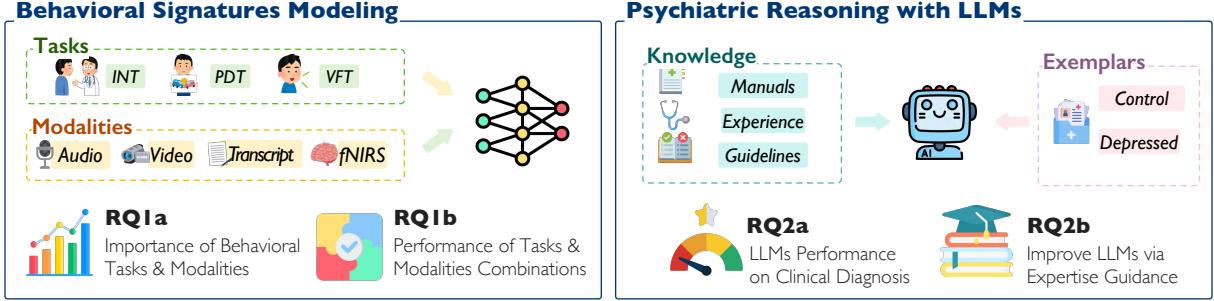


Figure 3: Overview of the two-part research methodology, addressing: 1) the diagnostic value of behavioral signatures (RQ1) and 2) the performance and enhancement of psychiatric reasoning in LLMs (RQ2).

Task & Modality Modeling We denote the task set as $\mathcal{T} = \{\text{INT}, \text{PDT}, \text{VFT}\}$ and the modality set as $\mathcal{M} = \{\text{Audio}, \text{Video}, \text{Transcript}, \text{fNIRS}\}$. For each subject i , let $X_{m,t}^{(i)}$ represent the features from modality m and task t . We train a classifier to map them to the clinical label:

$$f_\theta : X_{m,t}^{(i)} \rightarrow y^{(i)}, \quad y^{(i)} \in \{0, 1\}$$

This allows us to quantify the diagnostic value of each task-modality pair. To further explore whether aggregating behavioral evidence enhances performance, we conduct two types of fusion experiments. In *task fusion*, we fix the modality and concatenate features across all tasks: $X_{\text{fused}}^{(i)} = \text{Concat}(X_{m,t_1}^{(i)}, \dots, X_{m,t_k}^{(i)})$ for every $m \in \mathcal{M}$. This setup allows us to assess whether different task designs provide complementary cognitive and affective signals. In *modality fusion*, we fix the task set and concatenate features across all modalities: $X_{\text{fused}}^{(i)} = \text{Concat}(X_{m_1,t}^{(i)}, \dots, X_{m_k,t}^{(i)})$ for every $t \in \mathcal{T}$, followed by aggregation across tasks. This setting evaluates how multimodal observations enhance detection when applied consistently across structured clinical tasks. In both settings, the fused representation is passed to the same predictive function f_θ for classification.

3.2 Psychiatric Reasoning with LLMs

We examine whether LLMs can emulate clinician-like diagnostic reasoning based solely on transcripts (for text LLMs) or multimodal signals (for MLLM) from three structured psychiatric tasks: interview, picture description, and verbal fluency, spanning a total of ten tests (T1–T10). Each subject’s signals are concatenated into a single prompt, and the LLM is tasked with predicting a binary diagnostic label.

We compare three reasoning strategies. **1) Direct Prediction** asks the model to infer the diagnosis

directly from the input without any explanation. **2) Vanilla Reasoning** encourages the model to engage in free-form, step-by-step reasoning before making a prediction. **3) Psychiatric Reasoning** is our proposed method, which guides the model through a structured reasoning process grounded in clinical expertise. This approach reflects our key insight: effective psychiatric diagnosis depends not only on what is said, but also on how it is expressed under different task demands. To operationalize this, the prompt incorporates task definitions and expert-informed behavioral expectations, helping the LLM attend to symptom-relevant cues in a way that aligns with real clinical reasoning. The essential structure is summarized below.

Interview tasks probe emotional tone and autobiographical specificity... Picture description tasks assess imagination, semantic flow, and emotional valence... Verbal fluency tasks test lexical diversity and cognitive flexibility... Depressed individuals often express negative sentiment, lack detail, and repeat or simplify content... Healthy individuals tend to produce specific, emotionally rich, and well-organized language...

Let $\mathcal{P}^{(i)} = \{T_1^{(i)}, \dots, T_{10}^{(i)}\}$ represent the transcript prompt for subject i . The LLM performs a binary classification $g_\phi(\mathcal{P}^{(i)}) \rightarrow y^{(i)} \in \{0, 1\}$. When clinical guidance \mathcal{K} is included, the model performs $g_\phi(\mathcal{P}^{(i)}, \mathcal{K}) \rightarrow y^{(i)}$, using the embedded knowledge to attend to diagnostically meaningful patterns.

4 Experiments & Analysis

4.1 Experimental Settings

We conduct all experiments on C-MIND. To ensure robust evaluation, we randomly split the dataset into training, validation, and test sets following a 6:2:2 ratio. We report Macro-F1 as the main evaluation metric. Full metrics, including Precision,

| Modality | Feature Set | Interview (INT) | | | | | | Picture Description (PDT) | | | | | | Verbal Fluency (VFT) | | | | | |
|-----------------------|-------------|-----------------|-------|-------|-------|-------|-------|---------------------------|-------|-------|-------|-------|-------|----------------------|-------|-------|-------|-------|-------|
| | | LSTM | CNN | MLP | k-NN | RF | SVM | LSTM | CNN | MLP | k-NN | RF | SVM | LSTM | CNN | MLP | k-NN | RF | SVM |
| Audio (A) | OpenSmile | 72.15 | 84.25 | 82.31 | 85.18 | 91.17 | 91.11 | 69.49 | 88.24 | 81.21 | 94.10 | 88.24 | 85.18 | 70.90 | 84.28 | 79.39 | 76.14 | 88.24 | 82.29 |
| | Qwen-Audio | 72.57 | 78.41 | 58.39 | 55.54 | 69.26 | 76.39 | 74.07 | 71.54 | 67.58 | 69.64 | 72.47 | 79.39 | 76.93 | 82.33 | 61.63 | 78.96 | 76.43 | 76.39 |
| Video (V) | OpenFace | 70.41 | 71.47 | 69.00 | 72.94 | 72.40 | 73.32 | 72.25 | 75.46 | 65.42 | 69.64 | 74.49 | 64.58 | 74.64 | 74.49 | 69.54 | 69.64 | 75.43 | 64.71 |
| | Qwen-VL | 79.97 | 83.31 | 76.92 | 85.28 | 84.28 | 85.28 | 79.91 | 86.27 | 80.32 | 88.24 | 83.29 | 85.28 | 78.73 | 84.25 | 74.26 | 67.39 | 81.28 | 85.28 |
| Transcript (T) | DeBERTa | 60.93 | 64.33 | 51.92 | 46.32 | 58.02 | 52.28 | 66.36 | 62.16 | 48.54 | 55.54 | 52.87 | 51.43 | 57.83 | 56.54 | 56.37 | 62.64 | 56.44 | 55.84 |
| | Qwen | 60.57 | 64.69 | 58.16 | 60.92 | 61.37 | 61.73 | 67.69 | 68.38 | 56.81 | 58.88 | 67.58 | 67.39 | 55.07 | 59.29 | 53.83 | 76.14 | 62.13 | 64.21 |
| fNIRS (N) | Statistics | 62.54 | 71.27 | 59.61 | 43.33 | 73.49 | 61.46 | 65.98 | 69.06 | 54.83 | 42.05 | 73.30 | 45.36 | 62.55 | 64.05 | 65.57 | 50.18 | 75.43 | 46.88 |

Table 3: Performance (Macro-F1) of different models and feature sets, evaluated per task-modality combination. The color of each block corresponds to the average performance of the results within that block (a darker shade indicates a higher average).

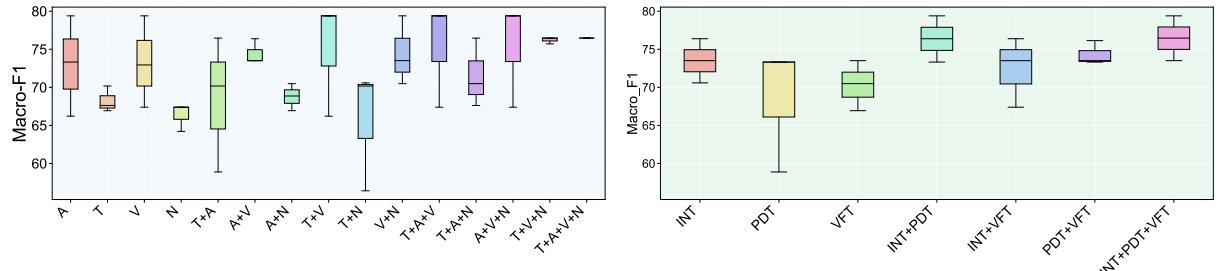


Figure 4: Performance (Macro-F1) of modality fusion (left) and task fusion (right). Combining signals generally improves the Macro-F1 score and reduces performance variance.

Recall, and per-class F1 scores, are available in the Technical Appendix. All results are averaged over five independent runs with different random seeds. Our analysis aims to answer two key research questions (RQs):

- RQ1:** What are the contributions of different behavioral tasks and modalities to depression assessment?
- RQ2:** Can LLMs reason like clinical psychiatrists, and how can knowledge injection improve their performance?

To address RQ1, we benchmark a suite of classical learning backbones, including LSTM, CNN, MLP, k-NN, Random Forest (RF), and SVM. To address RQ2, we evaluate several leading LLMs. The text-based LLMs, including GPT-4o, GPT-o3, DeepSeek-V3, DeepSeek-R1, and Qwen3-235B-A22B-T/NT (thinking/non-thinking mode), use only the transcript as input. In contrast, the multimodal model Qwen2.5-Omni processes a combination of audio, video, and transcript. Due to space limitations, detailed model architectures, parameters, and versions are provided in the Technical Appendix.

4.2 RQ1: The Power of Behavioral Signatures

Tasks and Modalities As shown in Table 3, we evaluate each task and modality using two feature sets. Our analysis reveals that Audio and Video are the most informative modalities, though their effectiveness is deeply intertwined with the psychiatric task being performed. Each task is designed to probe different cognitive and emotional facets, and their diagnostic power comes from how well these probes elicit observable, depression-related behavioral markers.

The Picture Description Task (PDT), for instance, excels in this regard, proving to be the most effective probe in our analysis. This is clinically intuitive as it assesses for emotional and attentional biases. Depressed individuals may exhibit a negative interpretation bias or provide less detailed descriptions, which is reflected not only in word choice (Transcript) but crucially in a flat vocal tone (Audio) and blunted affect (Video). This is evidenced by the top-performing model, which achieved a 94.10% Macro-F1 score using audio features from the PDT. Similarly, the Verbal Fluency Task (VFT) also shows remarkable performance, particularly with audio features. VFT assesses executive functions and semantic memory, which are often impaired in depression. This cog-

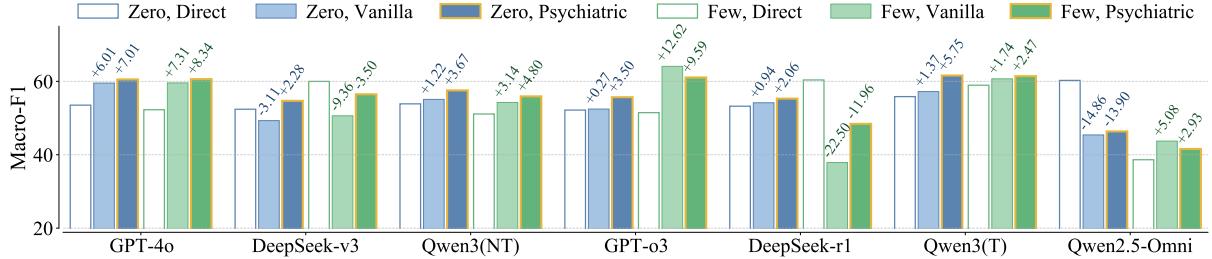


Figure 5: Performance comparison of LLMs using three reasoning strategies in zero/few-shot settings. Numbers above the bars indicate the performance change relative to the corresponding "Direct" baseline.

nitive deficit doesn't just manifest as a lower word count, but more saliently as acoustic patterns like longer pauses, frequent hesitation markers (e.g., "uh", "um"), and reduced prosodic variation. These are precisely the signals captured by audio analysis, explaining its success. The Interview task (INT) remains a robust baseline because its autobiographical prompts are effective at eliciting narratives laden with depressive markers like negative sentiment and overgeneralization, signals that are present across Audio, Video, and Transcript.

Fusion Improves Robustness As illustrated in Figure 4, a clear and consistent finding is that fusing evidence from multiple sources enhances diagnostic performance. Combining modalities (e.g., Audio and Video) or integrating tasks (e.g., INT and PDT) consistently leads to higher Macro-F1 scores and, critically, more stable and reliable predictions by reducing variance. This underscores the value of a holistic assessment strategy, where a richer, multi-faceted view of a participant's behavior provides a more robust foundation for clinical inference than any single signal alone.

4.3 RQ2: Psychiatric Reasoning with LLMs

Reasoning Strategies We evaluate seven leading LLMs under three prompting strategies: *Direct Prediction*, *Vanilla Reasoning*, and our proposed *Psychiatric Reasoning*, across both zero-shot and few-shot conditions (Figure 5). Several consistent patterns emerge.

1) Psychiatric Reasoning consistently improves zero-shot performance. Across most non-thinking models (e.g., GPT-4o, GPT-o3, Qwen3(NT)), the structured psychiatric prompt yields stable gains (e.g., +7.01% for GPT-4o, +3.67% for Qwen3(NT)), outperforming both Direct and Vanilla strategies. 2) Few-shot performance gains vary, and can conflict with structured guidance. For GPT-4o, Vanilla Reasoning under few-shot im-

proves significantly (+7.31%), but the gain from Psychiatric Reasoning (+8.34%) suggests that explicit guidance remains beneficial. In contrast, DeepSeek-v3 and DeepSeek-r1 show degradation under few-shot reasoning, likely due to incompatibility between pretrained reasoning paths and injected prompts. 3) Models with internal reasoning protocols may conflict with external prompts. DeepSeek-r1 exhibits a significant drop when reasoning is added, especially under few-shot settings (-22.5% with Vanilla and -11.96% with Psychiatric prompts), highlighting potential interference from overlaying external logic on built-in reasoning. 4) Multimodal input does not guarantee improved performance. The multimodal model Qwen2.5-Omni consistently underperforms across all prompting settings, achieving just 46.36% with Psychiatric Reasoning (zero-shot), which is worse than most transcript-only models. This suggests that general-purpose multimodal LLMs currently lack the fine-grained capability to utilize clinical non-verbal cues effectively without task-specific tuning.

Notably, even the best transcript-based LLM with Psychiatric Reasoning (GPT-4o, 60.53%) still falls short of the 68.24% achieved by a supervised model trained directly on transcript Qwen features, further emphasizing the performance gap between prompting-based and discriminative approaches in high-stakes diagnosis.

Task Fusion On LLMs To examine whether LLMs benefit from task-level information integration, we aggregate transcripts from multiple psychiatric tasks (Figure 6). Results align with RQ1: combining tasks improves performance. For example, GPT-4o's Macro-F1 improves from 50.48% (INT only) to 55.68% (INT+VFT), and further to 60.53% when Psychiatric Reasoning is applied. The best-performing configuration under transcript-only input is GPT-4o with INT+VFT and Psychiatric prompt. Similar trends are observed with GPT-

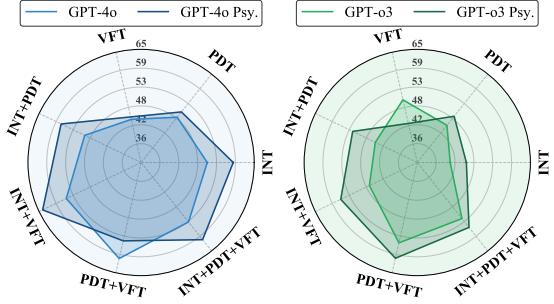


Figure 6: Performance of task combinations on LLMs.

o3. However, tasks like VFT remain underutilized in isolation, likely due to the loss of timing and repetition patterns during transcription. This supports our earlier findings that linguistic transcripts alone cannot fully capture the cognitive and affective richness of certain psychiatric tasks.

4.4 Case Study

The case study presented in Figure 7 serves as a clear illustration that explicitly integrating domain-specific psychiatric knowledge enhance the performance of depression assessment. In our observation, domain knowledge contributes in two crucial ways. First, it guides clinical interpretation of signals. In the case of PDT, rather than overreacting to raw metrics “*low word count*”, reasoning with psychiatric knowledge assesses recognize protective factors like emotional expressiveness, therefore correctly identifying non-pathological cases. Second, knowledge prevents over-weighting isolated negative signals. For example, when encountering “*Good people don’t get good rewards*,” the baseline model treats it as a core depressive marker. Psychiatric reasoning, however, draws on clinical reasoning to distinguish between fleeting complaints and the pervasive negativity typical of depression. By noting the lack of elaboration or supporting cues, it correctly down-weights the phrase.

5 Related Work

Corpus The foundation of depression detection is its data corpora, which have evolved along a hierarchy of evidence, trading scale for clinical validity. Early research leveraged large-scale social media corpora with labels based on user self-disclosure (Shen et al., 2017; Tadesse et al., 2019; Zirikly et al., 2019; Bucur et al., 2025). To improve signal quality, subsequent work introduced datasets collected in controlled settings, where ground truth

| | VFT (Fruit) | PDT (Cat) | INT (Angry Event) |
|-----------------------------------|--|--|---|
| Baseline | Apple, banana, pear, melon, watermelon, orange, lemon, cherry | It’s cute... healing... must have a good owner. | Good people don’t get good rewards. |
| Knowledge + Psychiatric Reasoning | <p>Declining in executive function and motor slowness are typical cognitive impairment features of depression.</p> <p>Focus on content, not just count. Absence of severe repetition or bizarre responses indicates healthy.</p> <p>The performance of healthy individuals in the VFT task fluctuates, and whether it is “significantly” lower than the normal level should be judged based on large-scale baseline data, rather than absolute counts.</p> | <p>Low speech output, reinforcing the diagnosis of “psychomotor retardation.”</p> <p>Ability to express positive emotion and infer positive context are contraindicators for depression.</p> <p>The ability to actively express pleasant emotions and infer others’ positive states reflect normal social cognition, which is typically impaired in severe depression.</p> | <p>Key sign of negative bias. Maps directly to a core depressive symptom.</p> <p>Lack of rumination or additional negative examples makes it a weak signal.</p> <p>It is necessary to differentiate between “state” complaints and “trait” pervasive negative thinking. The key is to observe whether it is accompanied by excessive generalization and rumination (repeated thinking).</p> |

Figure 7: A case contrasting a superficial, baseline interpretation with a nuanced, knowledge-guided assessment for the same healthy participant. Red marks the incorrect assumptions of the baseline, green highlights the correct interpretations guided by psychiatric expertise, and yellow identifies points that are considered but ultimately de-emphasized.

was typically derived from self-report questionnaire scores (Gratch et al., 2014; Valstar et al., 2016; Guo et al., 2021; Tasnim et al., 2022). Representing a move toward the clinical gold standard, the most recent corpora have begun to incorporate formal diagnoses from trained psychiatrists (Cai et al., 2022; Zou et al., 2022; Lin et al., 2022). These pioneering efforts often feature smaller or imbalanced cohorts and are focused on a limited set of behavioral tasks or modalities. Our work therefore introduces a new clinically-validated resource featuring a balanced cohort across diverse tasks and modalities.

Method Paralleling the evolution of datasets, detection methods have shifted from analyzing hand-crafted features within single modalities to learning complex data representations through sophisticated, multimodal architectures. Initial approaches relied on handcrafted features from single modalities, such as text, audio, or video (Fossati et al., 2003a; Cummins et al., 2015; Ma et al., 2016). A consensus has since formed around multimodal fusion models, which integrate these channels using sophisticated attention or transformer architectures to achieve stronger performance (Fan et al., 2019; Wei et al., 2023; Chen et al., 2024; Jia et al., 2025; Wu et al., 2025). The latest frontier involves applying

LLMs to this task. While powerful, research highlights challenges in adapting these general-purpose models for clinical use, noting the need to imbue them with specialized, domain-specific knowledge beyond what is learned from web-scale text (Guo et al., 2024; Wang et al., 2024; Hua et al., 2025; Bi et al., 2025). Our work contributes to this frontier by conducting a comprehensive analysis across different tasks and modalities to clarify their discriminative power, and then proposing a novel psychiatric reasoning mechanism to enhance the clinical awareness of LLMs.

6 Conclusion

We present the clinical multimodal neuropsychiatric diagnosis (C-MIND) dataset, a clinically validated resource collected from real hospital settings, featuring diverse behavioral signals across structured tasks and synchronized modalities. Through systematic analysis, we reveal how specific combinations of tasks and modalities enhance diagnostic stability, providing empirical guidance for system design. We also show that large language models, when guided by structured psychiatric knowledge, can better approximate expert reasoning in complex diagnostic scenarios. By integrating high-quality clinical data with interpretable and knowledge-informed modeling, this work offers a concrete step toward computational systems that are accurate, trustworthy, and deployable in real-world mental healthcare.

References

- T. Akiyama, M. Koeda, Y. Okubo, and M. Kimura. 2018. Hypofunction of left dorsolateral prefrontal cortex in depression during verbal fluency task: A multi-channel near-infrared spectroscopy study. *Journal of Affective Disorders*, 231:83–90.
- American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders*, 5th edition. American Psychiatric Publishing, Arlington, VA.
- J. Angst, R. Adolfsson, F. Benazzi, A. Gamma, E. Han-touche, T. D. Meyer, P. Skeppar, E. Struening, E. Vieta, and J. Scott. 2005. The hcl-32: Towards a self-assessment tool for hypomanic symptoms in outpatients. *Journal of Affective Disorders*, 88(2):217–233.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wen-bin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66. IEEE.
- Guanqun Bi, Zhuang Chen, Zhoufu Liu, Hongkai Wang, Xiya Xiao, Yuqiang Xie, Wen Zhang, Yongkang Huang, Yuxuan Chen, Libiao Peng, and Minlie Huang. 2025. MAGI: Multi-agent guided interview for psychiatric assessment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24898–24921, Vienna, Austria. Association for Computational Linguistics.
- Ana-Maria Bucur, Andreea-Codrina Moldovan, Kru-tika Parvatikar, Marcos Zampieri, Ashiqur R Khuda-Bukhsh, and Liviu P Dinu. 2025. Datasets for depression modeling in social media: An overview. *arXiv preprint arXiv:2503.21513*.
- Daniel J. Buysse, Charles F. Reynolds III, Timothy H. Monk, Susan R. Berman, and David J. Kupfer. 1989. The pittsburgh sleep quality index: A new instrument for psychiatric practice and research. *Psychiatry Research*, 28(2):193–213.
- Hanshu Cai, Zhenqin Yuan, Yiwen Gao, Shuteng Sun, Na Li, Fuze Tian, Han Xiao, Jianxiu Li, Zhengwu Yang, Xiaowei Li, and 1 others. 2022. A multi-modal open dataset for mental-disorder analysis. *Scientific Data*, 9(1):178.
- R. B. Cattell, H. W. Eber, and M. M. Tatsuoka. 1970. *Handbook for the Sixteen Personality Factor Questionnaire (16PF)*. Institute for Personality and Ability Testing, Champaign, IL.
- Zhuang Chen, Jiawen Deng, Jinfeng Zhou, Jincenzi Wu, Tieyun Qian, and Minlie Huang. 2024. Depression detection in clinical interviews with LLM-empowered structural element graph. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8181–8194, Mexico City, Mexico. Association for Computational Linguistics.
- Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.
- Xu Cui, Signe Bray, Daniel M Bryant, Gary H Glover, and Allan L Reiss. 2011. A quantitative comparison of nirs and fmri across multiple cognitive tasks. *NeuroImage*, 54(4):2808–2821.
- Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F Quatieri. 2015. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49.

- L. R. Derogatis. 1977. *SCL-90: Administration, scoring and procedures manual-I for the R(evised) version and other instruments of the Psychopathology Rating Scale Series*. Johns Hopkins University School of Medicine, Baltimore.
- Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462.
- Weiquan Fan, Zhiwei He, Xiaofen Xing, Bolun Cai, and Weirui Lu. 2019. Multi-modality depression detection via multi-scale temporal dilated cnns. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pages 73–80.
- Philippe Fossati, Anne-Marie Ergis, Jean-François Allilaire, and 1 others. 2003a. Qualitative analysis of verbal fluency in depression. *Psychiatry research*, 117(1):17–24.
- Philippe Fossati, Guillaume Le Bastard, Steven Small, Anne-Sophie Rigaud, Jean-Pierre Kahn, Sophie Pilliod, Nematollah Jaafari, Jean-François Allilaire, and Bruno Dubois. 2003b. Verbal fluency and clustering in patients with mood disorders. *Psychiatry Research*, 117(3):187–207.
- Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, and 1 others. 2014. The distress analysis interview corpus of human and computer interviews. In *LREC*, volume 14, pages 3123–3128. Reykjavik.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shitong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Weitong Guo, Hongwu Yang, Zhenyu Liu, Yaping Xu, and Bin Hu. 2021. Deep neural networks for depression recognition based on 2d and 3d facial expressions under emotional stimulus tasks. *Frontiers in neuroscience*, 15:609760.
- Zhijun Guo, Alvina Lai, Johan H Thygesen, Joseph Farrington, Thomas Keen, Kezhi Li, and 1 others. 2024. Large language models for mental health applications: systematic review. *JMIR mental health*, 11(1):e57400.
- M. Hamilton. 1959. The assessment of anxiety states by rating. *British Journal of Medical Psychology*, 32(1):50–55.
- M. Hamilton. 1960. A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry*, 23(1):56–62.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.
- Yining Hua, Fenglin Liu, Kailai Yang, Zehan Li, Hongbin Na, Yi-han Sheu, Peilin Zhou, Lauren V Moran, Sophia Ananiadou, David A Clifton, and 1 others. 2025. Large language models in mental health care: a scoping review. *Current Treatment Options in Psychiatry*, 12(1):1–18.
- Xiaowen Jia, Jingxia Chen, Kexin Liu, Qian Wang, and Jialing He. 2025. Multimodal depression detection based on an attention graph convolution and transformer. *Mathematical biosciences and engineering: MBE*, 22(3):652–676.
- Yunhan Lin, Biman Najika Liyanage, Yutao Sun, Tianlan Lu, Zhengwen Zhu, Yundan Liao, Qiushi Wang, Chuan Shi, and Weihua Yue. 2022. A deep learning-based model for detecting depression in senior population. *Frontiers in psychiatry*, 13:1016676.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.
- Yang Liu, Xiaoyong Lu, Daimin Shi, Jingyi Yuan, Tao Pan, and Haizhen An. 2021. Improved depression recognition using attention and multitask learning of gender recognition. In *2021 International Conference on Asian Language Processing (IALP)*, pages 57–61. IEEE.
- Xinyu Ma, Hongyu Yang, Qiang Chen, Di Huang, and Yunhong Wang. 2016. Depaudionet: an efficient deep model for audio based depression classification. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 35–42. ACM.
- OpenAI. 2024. Hello gpt-4o. <https://openai.com/index/hello-gpt-4o/>. Accessed: 2025-08-04.
- OpenAI. 2025. Introducing openai o3 and o4-mini. <https://openai.com/index/introducing-o3-and-o4-mini/>. Accessed: 2025-08-04.
- C. Ramponi, F. C. Murphy, A. J. Calder, and P. J. Barnard. 2010a. Recognition memory for pictorial material in subclinical depression. *Acta Psychologica*, 135(3):293–301.
- Cristina Ramponi, Simon L Collinson, Richard Worters, and Nicola Breen. 2010b. Picture perception in depression: A systematic review and meta-analysis of neuroimaging studies. *Journal of Psychiatric Research*, 44(15):1002–1014.
- Andrew Rinaldi, Jean E. Fox Tree, and Snigdha Chaturvedi. 2020. Predicting depression in screening interviews from latent categorization of interview

- prompts. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7–18. Association for Computational Linguistics.
- Sabah Mohammed Sarsam, Hosam Al-Samarraie, Ahmed Ibrahim Alzahrani, and Bianca Wright. 2024. Multimodal machine learning in mental health: A survey of data, algorithms, and challenges. *Information Fusion*, 106:102517.
- Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, Wenwu Zhu, and 1 others. 2017. Depression detection via harvesting social media: A multimodal dictionary learning solution. In *IJCAI*, volume 2017, pages 3838–3844.
- Michael Mesfin Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. 2019. Detection of depression-related posts in reddit social media forum. *IEEE Access*, 7:44883–44893.
- Mashrura Tasnim, Malikeh Ehghaghi, Brian Diep, and Jekaterina Novikova. 2022. Depac: a corpus for depression and anxiety detection from speech. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 1–16.
- Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 2016. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*, pages 3–10. ACM.
- Yuxi Wang, Diana Inkpen, and Prasadith Kirinde Gamaarachchige. 2024. Explainable depression detection using large language models on social media data. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 108–126, St. Julians, Malta. Association for Computational Linguistics.
- Yuntao Wei, Yuzhe Zhang, Shuyang Zhang, and Hone Zhang. 2023. Canamrf: An attention-based model for multimodal depression detection. In *Pacific Rim International Conference on Artificial Intelligence*, pages 111–116. Springer.
- Zijian Wu, Leijing Zhou, Shuanglin Li, Changzeng Fu, Jun Lu, Jing Han, Yi Zhang, Zhuang Zhao, and Siyang Song. 2025. Depmgnn: Matrixial graph neural network for video-based automatic depression assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1610–1619.
- Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Jihyeon Yoon, Chae-yeon Kang, Soyeon Kim, and Jaewoo Han. 2022. D-vlog: Multimodal vlog dataset for depression detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12226–12234.
- Ayah Zirikly, Philip Resnik, Ozlem Uzuner, and Kristy Hollingshead. 2019. Clpsych 2019 shared task: Predicting the degree of suicide risk in reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 24–33. ACL.
- Bochao Zou, Jiali Han, Yingxue Wang, Rui Liu, Shenghui Zhao, Lei Feng, Xiangwen Lyu, and Huimin Ma. 2022. Semi-structural interview-based chinese multimodal depression corpus towards automatic preliminary screening of depressive disorders. *IEEE Transactions on Affective Computing*, 14(4):2823–2838.
- William W. K. Zung. 1965. A self-rating depression scale. *Archives of General Psychiatry*, 12(1):63–70.
- William W. K. Zung. 1971. A rating instrument for anxiety disorders. *Psychosomatics*, 12(6):371–379.

A C-MIND Collection Details

A.1 Experimental Materials

1) Interview Task (INT) The task has a practice part and a formal part.

Practice Session

- **Instructions:** “A few questions will show up on the screen. Please say your answers. If you are ready, press the spacebar to start the practice.”
- **Procedure:** First, a dot appears for 500ms. Then, a question appears with a note: “You have 5 seconds to prepare.” This screen shows for 10 seconds. Next, a new screen shows the same question with the note: “Now, please begin your answer.”
- **Material:** The practice question is: Please describe one thing you regret the most.

Formal Session

- **Instructions:** “A few questions will show up on the screen. Please say your answers. If you are ready, please press the spacebar to begin.”
- **Procedure:** Same as the practice part.
- **Materials:** The formal session includes three questions: a) Please describe an event that made you the angriest. b) Please describe an event that you felt was the most meaningful. c) Please describe your favorite food.

2) Picture Description Task (PDT) This task also has a practice and a formal part. The images used are shown in Figure 8.

Practice Session

- *Instructions:* “Next, you will see a picture. Please talk about the feelings or thoughts the picture gives you. Note: Please only talk about the picture itself. If you are ready, press the spacebar to start the practice.”
- *Procedure:* After a 500ms dot, a picture appears. The text above it says: “What thoughts or feelings does this picture give you? You have 5 seconds to prepare.” This screen shows for 10 seconds. Then, the screen shows the same picture with new text: “Now, please begin talking.”
- *Material:* The practice picture is an image of a flying bird (see Figure 8a).

Formal Session

- *Instructions:* “Next, you will see a picture. Please talk about the feelings or thoughts the picture gives you. Note: Please only talk about the picture itself. If you are ready, press the spacebar to start the formal experiment.”
- *Procedure:* Same as the practice part.
- *Materials:* The formal session uses three pictures: a cat (Figure 8b), a car accident (Figure 8c), and a spaceship (Figure 8d).

3) Verbal Fluency Task (VFT) This task asks people to say words from a specific group.

Practice Session

- *Instructions:* “A word for a category will appear on the screen. Please say words from that category. You have 30 seconds. If you are ready, press the spacebar to start.”
- *Procedure:* After a 500ms dot, a screen shows a category word with examples. The text above says: “Now, please state your answer.” This screen is on for 30 seconds. Then, there is a 10-second rest.
- *Material:* The practice category is: Green plants (e.g., you can say ivy, cabbage, etc.).

Formal Session

- *Instructions:* “A word for a category will appear on the screen. Please say words from that category. You have 30 seconds! If you are ready, please press the spacebar to start the formal experiment.”
- *Procedure:* Same as the practice part.
- *Materials:* The categories for the formal experiment are: a) Four-legged animals (e.g., pig, cow, etc.) b) Round fruits (e.g.,

watermelon, apple, etc.) c) Cities with two-character names (e.g., Beijing, Shanghai, etc.) d) Vegetables (e.g., carrot, broccoli, etc.)

A.2 Participant Examples

To give a clear picture of our data, this section shows two examples: one male participant (Figure 9) and one female participant (Figure 10). For each person, we show two images. The first image shows the person in our lab during the experiment. This shows our data collection setup. The second image is a sample of the real, anonymized clinical record. This record is the gold-standard ground truth we use for the final diagnosis label in our study.

A.3 Details of Psychometric Questionnaires

To get a full picture of each participant’s emotional state, we use a set of standard questionnaires to help check for related factors like sleep problems or personality traits. The details of each instrument are as follows.

- 1) **HAMD** Hamilton Depression Rating Scale ([Hamilton, 1960](#)). This is a clinician-rated tool to get an objective evaluation of depression severity. It is based on observed behaviors and what the participant reports.
- 2) **HAMA** Hamilton Anxiety Rating Scale ([Hamilton, 1959](#)). This is also a clinician-rated tool that gives an objective evaluation of how severe a person’s anxiety is.
- 3) **SDS** Self-Rating Depression Scale ([Zung, 1965](#)). This is a self-report tool. It helps capture a person’s own feelings of depression, which might not be clear from an interview alone.
- 4) **SAS** Self-Rating Anxiety Scale ([Zung, 1971](#)). Like the SDS, this is a self-report tool for a person’s own feelings of anxiety.
- 5) **PSQI** Pittsburgh Sleep Quality Index ([Buysse et al., 1989](#)). This scale measures a person’s sleep quality over the last month. We include it because sleep problems are common in mood disorders.
- 6) **16PF** 16 Personality Factor Questionnaire ([Cattell et al., 1970](#)). This tool assesses personality traits. This helps us understand if certain personality types are related to a person’s mental state.



(a) Image: Flying bird

(b) Image: Cat

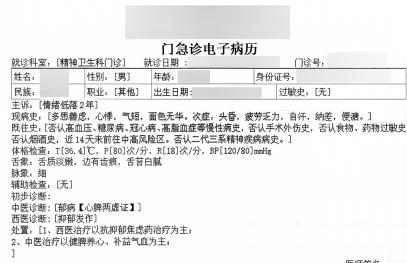
(c) Image: Car accident

(d) Image: Spaceship

Figure 8: Visual stimuli used in the Picture Description Task (PDT). Figure (a) is the image for the practice session, while (b), (c), and (d) are the images for the formal experiment.



(a) A male participant in the data collection environment.

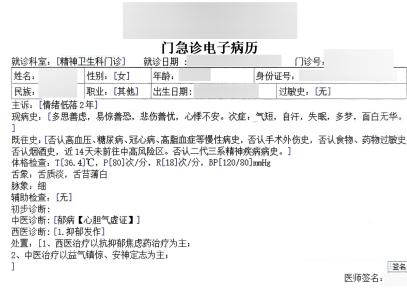


(b) Anonymized clinical diagnosis record.

Figure 9: Example of a male participant.



(a) A female participant in the data collection environment.



(b) Anonymized clinical diagnosis record.

Figure 10: Example of a female participant.

7) SCL-90 Symptom Checklist-90 (Derogatis, 1977). This scale screens for a wide range of psychological symptoms. This helps us check for other conditions that might exist alongside depression.

8) HCL-32 Hypomania Checklist-32 (Angst et al., 2005). This checklist helps identify signs of hypomania. We use it to make sure we do not misdiagnose someone with bipolar disorder as having depression. This improves our diagnostic accuracy.

B Feature Extraction Details

We extract two different sets of features for experiments. The first is a “Classical Feature Set” from common toolkits. The second is a “Foundation Model Feature Set” from large pretrained models.

B.1 Classical Feature Set

1) Audio We use the eGeMAPS feature set from OpenSMILE 3.0 (Eyben et al., 2010) to extract features from the audio clips. This feature set includes statistics for 8 frequency-related features, 3 energy-related features, and 14 spectral features. This gives a total of 88 features for audio.

2) Video We use OpenFace 2.2.0 (Baltrušaitis et al., 2018) to extract facial features from video files. OpenFace is an open-source tool for facial analysis. We extract features like frame number, confidence, eye gaze vectors, eye landmarks, head pose, and facial action units (AUs). We then take frames with a confidence score greater than 0.75 and use MATLAB’s smooth function to smooth the data with a window size of 11. For all processed features, we calculate 7 statistics: minimum, maximum, mean, variance, range, kurtosis, and

skewness. This gives a total of 4,963 features.

3) Transcript We use a Chinese pretrained DeBERTa model (He et al., 2020) to extract 768-dimension feature vector for the text.

4) fNIRS We collect the oxygenated and deoxygenated hemoglobin concentration from 45 channels of the fNIRS device (Cui et al., 2011). For this data, we calculate the same 7 statistics as the video features: minimum, maximum, mean, variance, range, kurtosis, and skewness. This gives a total of 630 features for fNIRS.

B.2 Foundation Model Feature Set

We use large Qwen models to create another feature set. All models run using vLLM for deployment.

1) Audio We use Qwen2-Audio-7B-Instruct (Chu et al., 2023). We take the audio for each of the 10 tasks and cut it into clips shorter than 30 seconds. We feed these clips into the model and take the features from the last hidden layer. We join the features for the same task and then apply global max pooling over the time dimension. This gives a final 4,096-dimension audio feature.

2) Video We use Qwen2.5-VL-72B (Bai et al., 2025). First, we split each participant’s full video into 10 smaller videos based on the 10 task timestamps. We feed each video into the model. Each frame is resized to 224x224 pixels. We use the official `get_video_features` function to get the feature tensor. Then, we apply global max pooling over the time dimension. This gives a final 8,192-dimension video feature.

3) Transcript We use Qwen3-235B-A22B (Yang et al., 2025). We take the speech-to-text transcripts for each of the 10 tasks and feed them into the model. We extract features from the last hidden layer and then apply global max pooling over the sequence dimension. This gives a final 4,096-dimension text feature.

4) fNIRS Due to the lack of a pretrained model for fNIRS signals, we use the same statistical features for fNIRS as in the Classical Feature Set.

C Methodology Implementation Details

C.1 Behavioral Signature Modeling

We train six different learning backbones to model the behavioral signatures. The implementation details for each model are as follows:

1) LSTM For the LSTM model, we first extract features for specific time steps based on the task type and normalize the data for each modality. Each modality passes through a Batch Normalization (BN) layer, followed by a two-layer fully-connected network for dimension reduction (hidden dimensions 512 then 128), with a Dropout rate of 0.2. For single-modality cases, the reduced features are used directly. For multi-modality, all modal features are first concatenated. Then, global average pooling is used to generate one-dimensional gate weights, which are used to fuse the information from different paths in a weighted manner. The fused features then enter a two-layer bidirectional LSTM, with each layer containing 512 hidden units and a Dropout rate of 0.3. This is followed by a single-head self-attention layer with a key dimension of 128 and a dropout rate of 0.3. Afterwards, we apply global average pooling, feed the result into a 128-unit fully-connected layer, and finally use a 1-unit Sigmoid layer to output the binary classification probability. We use the AdamW optimizer for training with a learning rate of 1e-4, weight decay of 1e-4, and a gradient clipping norm of 1.0 with amsgrad enabled. The loss function is binary cross-entropy. We monitor the validation loss and use an early stopping mechanism with a patience of 20 epochs, saving the best model from each training run.

2) CNN To fit the CNN’s input structure, the 512-dimensional fused features are reshaped into a 3D tensor of shape (512, 1, 1). The model first receives this input tensor and passes it through a convolutional layer with 32 filters of size (2, 1), a stride of (2, 1), and ‘same’ padding. This is followed by a batch normalization layer and a ReLU activation function. To mitigate overfitting, a Dropout layer with a rate of 0.3 is added. Next, a Flatten layer is used to flatten the multi-dimensional output, which is then fed into a fully-connected layer with 32 neurons, a ReLU activation, and L2 regularization ($\lambda = 0.01$). Another Dropout layer (rate 0.3) follows. The final output layer has a single neuron with a sigmoid activation function to produce the probability for the positive class. The model is trained using the SGD optimizer with a learning rate of 1e-4 and binary cross-entropy as the loss function. We use callbacks for EarlyStopping, ReduceLROnPlateau, and ModelCheckpoint. The model with the best validation performance is saved.

3) MLP The Multi-Layer Perceptron model has two hidden layers. The first is a fully-connected layer with 64 neurons, followed by a ReLU activation and Dropout (rate 0.3). The second is a 32-neuron fully-connected layer with the same structure. To enhance generalization, we also add an L2 regularization term ($\lambda = 0.01$) to both hidden layers. The output layer uses a Sigmoid activation function for the binary classification task. The model is trained using the SGD optimizer with an initial learning rate of 0.001. It also uses callbacks like EarlyStopping, ReduceLROnPlateau, and ModelCheckpoint to dynamically adjust the learning rate and prevent overfitting. During training, the validation loss is monitored, and the model parameters with the best validation performance are saved.

4) k-NN For the K-Nearest Neighbors model, we perform an automatic selection of the hyperparameter k . We test different numbers of neighbors from 1 to 10 and use five-fold cross-validation to evaluate the accuracy for each. The k value that yields the highest average accuracy is selected to build the final model.

5) RF For the Random Forest model, we use RandomizedSearchCV with five-fold cross-validation to search for the best hyperparameters. We search over two key parameters: the number of trees ($n_estimators$, from 25 to 500) and the maximum depth of the trees (max_depth , from 2 to 50). The final model is then rebuilt, trained, and evaluated using the best parameters found in the search.

6) SVM For the Support Vector Machine model, we use a linear kernel. The regularization strength C is set to 1, and we set $class_weight = balanced$ to address the class imbalance issue in the dataset.

C.2 Psychiatric Reasoning with LLMs

We evaluate several leading LLMs to assess their psychiatric reasoning capabilities. We select seven models for our experiments, including text-based and multimodal models. The models and their specific API parameters used for inference are described below.

1) GPT-4o : A model from OpenAI, released on May 13, 2024. It can process any combination of text, image, and video as input and generate text, audio, and image outputs (OpenAI, 2024).

2) GPT-o3 : A model from OpenAI, released on April 16, 2025. It is described as OpenAI’s most powerful reasoning model, advancing capabilities in coding, math, and visual perception (OpenAI, 2025).

3) DeepSeek-R1 : A reasoning model from the AI company DeepSeek, released on January 20, 2025. Its model weights are open-sourced (Guo et al., 2025).

4) DeepSeek-V3 : An LLM from DeepSeek AI, released and open-sourced on December 26, 2024 (Liu et al., 2024).

5) Qwen3-235B-A22B : A model from Alibaba, released on April 29, 2025. A key innovation is integrating a “thinking mode” for complex reasoning with a “non-thinking mode” for fast responses into one framework. We test both modes, referred to as Qwen3 (T) and Qwen3 (NT) (Yang et al., 2025).

6) Qwen2.5-Omni : An end-to-end omni-modal model from the Qwen team, released on March 27, 2025. It supports simultaneous input of text, image, audio, and video with real-time streaming output (Xu et al., 2025).

All models are called using their official APIs with fixed parameters to ensure deterministic outputs. The prompts for direct prediction, vanilla reasoning and psychiatric reasoning are shown in Table 5. The specific parameters for each model are detailed in Table 4.

D Experimental Results Details

This section provides supplementary tables with detailed results for the experiments. The detailed performance metrics for the behavioral signature modeling, including Precision, Recall, and F1-scores, are presented in Table 6. The detailed results for the modality and task fusion are shown in Table 7 and Table 8. The results for the psychiatric reasoning experiments with LLMs, which show the performance of different models under various reasoning strategies, are available in Table 9.

| Model | Version | Temperature | Top P | Frequency Penalty | Presence Penalty | Stop | Logprobs | Stream | Logit Bias | n |
|--------------|------------|-------------|-------|-------------------|------------------|------|----------|--------|------------|---|
| GPT-4o | 2024-05-13 | 0 | 1.0 | 0 | 0 | none | false | false | none | 1 |
| GPT-o3 | 2025-04-16 | 0 | 1.0 | 0 | 0 | none | false | false | none | 1 |
| DeepSeek-R1 | 2025-01-20 | 0 | 0.95 | 0 | 0 | none | false | false | none | 1 |
| DeepSeek-V3 | 2024-12-26 | 0 | 0.95 | 0 | 0 | none | false | false | none | 1 |
| Qwen3 (NT) | - | 0 | 1.0 | 0 | 0 | none | false | - | none | 1 |
| Qwen3 (T) | - | 0 | 1.0 | 0 | 0 | none | false | - | none | 1 |
| Qwen2.5-Omni | - | 0 | 1.0 | 0 | 0 | none | false | - | none | 1 |

Table 4: API parameters for LLM inference.

| |
|--|
| [Task Setting] A depression detection task is required. Based on the speech-to-text content of subjects in experimental tasks, determine whether they are depression patients (1) or healthy controls (0). This experiment aims to study the linguistic features of subjects under different tasks to distinguish between depression patients and healthy controls. Each subject completes 10 tasks of the following types: 1) Interview Tasks (T1-T3): Subjects answer questions about “an anger-inducing event,” “a meaningful event,” and “their favorite food.” 2) Picture Description Tasks (T4-T6): Subjects describe pictures of a “cat,” a “car accident,” and a “spaceship,” and engage in free association. 3) Verbal Fluency Tasks (T7-T10): Subjects perform free association based on the categories: “four-legged things,” “fruits,” “cities,” and “vegetables.” If a subject’s content for a specific task was not collected, it is marked as “\$missing”. |
| ONLY IN FEW-SHOT SETTINGS |
| [Example Samples] /Depressed Sample/ + /Control Sample/ |
| [Sample to be Judged] /Test Sample/ |
| ONLY FOR PSYCHIATRIC REASONING |
| [Reasoning Process] Think based on the following diagnostic clues: The three types of tasks have their own testing goals. The interview tasks mainly test the subject’s emotional response and self-awareness. The picture description tasks mainly test semantic fluency, imagination, and cognitive flexibility. The verbal fluency tasks are mainly used to evaluate verbal fluency and cognitive flexibility. For patients with depression , they tend to use more negative words and lack detailed descriptions in interview tasks. In picture description tasks, their descriptions are simpler, lean towards negative outcomes, and lack emotional expression. In verbal fluency tasks, they list fewer items and may have repetitions or words irrelevant to the topic. For healthy subjects , they tend to be more rational and specific in their descriptions in interview tasks. In picture description tasks, their descriptions are more detailed, their reflections on events are more diverse, and they show richer emotional expression. In verbal fluency tasks, they tend to list a richer variety of items. |
| [Task Requirements] |
| FOR ZERO-SHOT SETTINGS Your task is to determine whether the subject is a depression patient (labeled 1) or a healthy control (labeled 0) based on the text content from [T1] to [T10]. A judgment must be made regardless of how many “\$missing” entries are present. |
| FOR FEW-SHOT SETTINGS Your task is to determine whether the subject is a depression patient (labeled 1) or a healthy control (labeled 0) based on the text content from [T1] to [T10] in the “Sample to be Judged,” in conjunction with the “Example Samples.” A judgment must be made regardless of how many “\$missing” entries are present. |
| [Output Requirements] |
| FOR DIRECT METHOD If judged as depression, output “1”. If judged as a healthy control, output “0”. Do not output any text other than “1” or “0”. Do not output any other explanatory content. |
| FOR ZERO-SHOT VANILLA/PSYCHIATRIC REASONING First, output the reasoning process, and then on a new line, output only the judgment result. If judged as depression, output “1”; if judged as a healthy control, output “0”. |
| FOR FEW-SHOT VANILLA/PSYCHIATRIC REASONING First, output the reasoning process, summarizing and comparing the depression patient and healthy control in the “Example Samples.” Then, on a new line, output only the judgment result. If judged as depression, output “1”; if judged as a healthy control, output “0”. |

Table 5: Unified prompt template for depression assessment.

| Modality | Feature Set | Metric | Interview (INT) | | | | | | Picture Description (PDT) | | | | | | Verbal Fluency (VFT) | | | | | |
|------------|-------------|---------------|-----------------|-------|-------|-------|-------|-------|---------------------------|-------|-------|-------|-------|-------|----------------------|-------|-------|-------|-------|-------|
| | | | LSTM | CNN | MLP | k-NN | RF | SVM | LSTM | CNN | MLP | k-NN | RF | SVM | LSTM | CNN | MLP | k-NN | RF | SVM |
| Audio | OpenSmile | Macro-F1 | 72.15 | 84.25 | 82.31 | 85.18 | 91.17 | 91.11 | 69.49 | 88.24 | 81.21 | 94.10 | 88.24 | 85.18 | 70.90 | 84.28 | 79.39 | 76.14 | 88.24 | 82.29 |
| | | Accuracy | 72.35 | 84.31 | 82.35 | 85.29 | 91.18 | 91.18 | 70.00 | 88.24 | 81.37 | 94.12 | 88.24 | 85.29 | 71.18 | 84.31 | 79.41 | 76.47 | 88.24 | 82.35 |
| | | Depression-P | 72.08 | 80.76 | 80.08 | 86.43 | 91.32 | 92.50 | 66.18 | 88.24 | 78.22 | 94.74 | 88.24 | 86.43 | 70.19 | 83.26 | 79.86 | 78.02 | 88.24 | 82.81 |
| | | Depression-R | 74.12 | 90.20 | 86.27 | 85.29 | 91.18 | 91.18 | 82.35 | 88.24 | 88.24 | 94.12 | 88.24 | 85.29 | 76.47 | 86.27 | 78.43 | 76.47 | 88.24 | 82.35 |
| | | Depression-F1 | 72.59 | 85.18 | 83.01 | 83.87 | 90.91 | 91.89 | 73.35 | 88.24 | 82.75 | 93.75 | 88.24 | 86.49 | 72.93 | 84.63 | 79.08 | 78.95 | 88.24 | 83.33 |
| | | Control-F1 | 71.72 | 83.32 | 81.62 | 86.49 | 91.43 | 90.32 | 65.63 | 88.24 | 79.66 | 94.44 | 88.24 | 83.87 | 68.87 | 83.94 | 79.71 | 73.33 | 88.24 | 81.25 |
| Qwen-Audio | Qwen-Audio | Macro-F1 | 72.57 | 78.41 | 58.39 | 55.54 | 69.26 | 76.39 | 74.07 | 71.54 | 67.58 | 69.64 | 72.47 | 79.39 | 76.93 | 82.33 | 61.63 | 78.96 | 76.43 | 76.39 |
| | | Accuracy | 72.94 | 78.43 | 58.82 | 55.88 | 69.61 | 76.47 | 74.12 | 71.57 | 68.63 | 70.59 | 72.55 | 79.41 | 77.06 | 82.35 | 62.75 | 79.41 | 76.47 | 76.47 |
| | | Depression-P | 71.83 | 77.30 | 57.24 | 56.07 | 69.92 | 76.84 | 74.71 | 72.22 | 64.74 | 73.52 | 72.83 | 79.51 | 76.70 | 79.92 | 69.47 | 82.20 | 76.66 | 76.84 |
| | | Depression-R | 76.47 | 80.39 | 68.63 | 55.88 | 69.61 | 76.47 | 72.94 | 70.59 | 84.31 | 70.59 | 72.55 | 79.41 | 78.82 | 86.27 | 50.98 | 79.41 | 76.47 | 76.47 |
| | | Depression-F1 | 73.17 | 78.76 | 62.36 | 51.61 | 66.58 | 77.78 | 73.69 | 71.34 | 72.99 | 64.29 | 71.43 | 80.00 | 77.44 | 82.96 | 57.37 | 75.86 | 75.91 | 77.78 |
| | | Control-F1 | 71.96 | 78.06 | 54.41 | 59.46 | 71.93 | 75.00 | 74.45 | 71.75 | 62.18 | 75.00 | 73.52 | 78.79 | 76.41 | 81.69 | 65.89 | 82.05 | 76.95 | 75.00 |
| Video | OpenFace | Macro-F1 | 70.41 | 71.47 | 69.00 | 72.94 | 72.40 | 73.32 | 72.25 | 75.46 | 65.42 | 69.64 | 74.49 | 64.58 | 74.64 | 74.49 | 69.54 | 69.64 | 75.43 | 64.71 |
| | | Accuracy | 70.59 | 71.57 | 69.61 | 73.53 | 72.55 | 73.53 | 72.35 | 75.49 | 65.69 | 70.59 | 74.51 | 64.71 | 74.71 | 74.51 | 69.61 | 70.59 | 75.49 | 64.71 |
| | | Depression-P | 71.50 | 72.53 | 70.87 | 75.76 | 73.01 | 74.29 | 72.80 | 76.23 | 65.59 | 73.52 | 74.56 | 64.91 | 76.55 | 74.56 | 71.29 | 73.52 | 75.77 | 64.71 |
| | | Depression-R | 70.59 | 70.59 | 72.55 | 73.53 | 72.55 | 73.53 | 72.94 | 74.51 | 66.67 | 70.59 | 74.51 | 64.71 | 71.76 | 74.51 | 66.67 | 70.59 | 75.49 | 64.71 |
| | | Depression-F1 | 70.50 | 71.26 | 70.40 | 68.97 | 73.57 | 70.97 | 72.62 | 75.25 | 65.69 | 64.29 | 73.98 | 66.67 | 73.93 | 74.49 | 68.75 | 64.29 | 74.24 | 64.71 |
| | | Control-F1 | 70.32 | 71.67 | 67.61 | 76.92 | 71.23 | 75.68 | 71.88 | 75.66 | 65.15 | 75.00 | 70.01 | 62.50 | 75.36 | 74.49 | 70.32 | 75.00 | 76.61 | 64.71 |
| Qwen-VL | Qwen-VL | Macro-F1 | 79.97 | 83.31 | 76.92 | 85.28 | 84.28 | 85.28 | 79.91 | 86.27 | 80.32 | 88.24 | 83.29 | 85.28 | 78.73 | 84.25 | 74.26 | 67.39 | 81.28 | 85.28 |
| | | Accuracy | 80.00 | 83.33 | 77.45 | 85.29 | 84.31 | 85.29 | 80.00 | 86.27 | 80.39 | 88.24 | 83.33 | 85.29 | 78.82 | 84.31 | 74.51 | 67.65 | 81.37 | 85.29 |
| | | Depression-P | 81.14 | 82.93 | 83.80 | 85.42 | 84.55 | 85.42 | 81.59 | 84.97 | 84.70 | 88.24 | 83.68 | 85.42 | 83.31 | 84.72 | 80.85 | 68.21 | 81.99 | 85.42 |
| | | Depression-R | 78.82 | 84.31 | 72.55 | 85.29 | 84.31 | 85.29 | 77.65 | 88.24 | 74.51 | 88.24 | 83.33 | 85.29 | 72.94 | 84.31 | 64.71 | 67.65 | 81.37 | 85.29 |
| | | Depression-F1 | 79.83 | 83.51 | 76.12 | 84.85 | 83.65 | 85.71 | 79.29 | 86.55 | 79.22 | 88.24 | 82.45 | 84.85 | 77.61 | 84.19 | 71.81 | 64.52 | 79.97 | 84.85 |
| | | Control-F1 | 80.11 | 83.11 | 77.73 | 85.71 | 84.92 | 84.85 | 80.54 | 85.98 | 81.42 | 88.24 | 84.13 | 85.71 | 79.84 | 84.31 | 76.71 | 70.27 | 82.58 | 85.71 |
| Transcript | DeBERTa | Macro-F1 | 60.94 | 64.33 | 51.92 | 46.32 | 58.02 | 52.28 | 66.36 | 62.16 | 48.54 | 55.54 | 52.87 | 51.43 | 57.83 | 56.54 | 56.37 | 62.64 | 56.44 | 55.84 |
| | | Accuracy | 61.18 | 64.71 | 51.96 | 47.06 | 58.82 | 52.94 | 66.47 | 63.73 | 50.00 | 55.88 | 53.92 | 52.94 | 59.41 | 56.86 | 57.84 | 64.71 | 56.86 | 55.88 |
| | | Depression-P | 62.90 | 62.32 | 51.85 | 46.89 | 59.55 | 53.11 | 65.76 | 60.08 | 50.21 | 56.07 | 54.30 | 53.36 | 64.86 | 56.85 | 56.39 | 68.89 | 57.16 | 55.90 |
| | | Depression-R | 54.12 | 74.51 | 54.90 | 47.06 | 58.82 | 52.94 | 69.41 | 82.35 | 64.71 | 55.88 | 53.92 | 52.94 | 41.18 | 56.86 | 74.51 | 64.71 | 56.86 | 55.88 |
| | | Depression-F1 | 58.11 | 67.84 | 53.33 | 52.63 | 63.79 | 57.89 | 67.39 | 69.27 | 56.37 | 59.46 | 59.81 | 60.00 | 50.00 | 56.44 | 64.07 | 53.85 | 52.19 | 54.55 |
| | | Control-F1 | 63.76 | 60.83 | 50.51 | 40.00 | 52.26 | 46.67 | 65.33 | 55.06 | 40.72 | 51.61 | 45.93 | 42.86 | 65.66 | 56.65 | 48.67 | 71.43 | 60.69 | 57.14 |
| Qwen | Qwen | Macro-F1 | 60.57 | 64.69 | 58.16 | 60.92 | 61.37 | 61.73 | 67.69 | 68.38 | 56.81 | 58.88 | 67.58 | 67.39 | 55.07 | 59.29 | 53.83 | 76.14 | 62.13 | 64.21 |
| | | Accuracy | 61.18 | 64.71 | 58.82 | 61.76 | 61.76 | 61.76 | 68.24 | 68.63 | 56.86 | 61.76 | 68.63 | 67.65 | 58.24 | 60.78 | 57.84 | 76.47 | 62.75 | 64.71 |
| | | Depression-P | 65.04 | 64.05 | 57.02 | 62.88 | 62.28 | 61.81 | 70.08 | 70.75 | 56.99 | 66.35 | 71.27 | 68.21 | 57.29 | 59.13 | 53.73 | 78.02 | 63.04 | 65.57 |
| | | Depression-R | 52.94 | 66.67 | 70.59 | 61.76 | 61.76 | 61.76 | 63.53 | 64.71 | 54.90 | 61.76 | 68.63 | 67.65 | 57.65 | 74.51 | 54.90 | 76.47 | 62.75 | 64.71 |
| | | Depression-F1 | 57.65 | 65.32 | 62.99 | 55.17 | 60.21 | 60.61 | 65.71 | 67.08 | 55.88 | 48.00 | 61.79 | 64.52 | 53.26 | 65.35 | 50.43 | 78.95 | 57.88 | 60.00 |
| | | Control-F1 | 63.49 | 64.05 | 53.33 | 66.67 | 62.53 | 62.86 | 69.67 | 69.68 | 57.73 | 69.77 | 73.36 | 70.27 | 56.87 | 53.24 | 57.22 | 73.33 | 66.38 | 68.42 |
| fNIRS | Statistics | Macro-F1 | 62.54 | 71.27 | 59.61 | 43.33 | 73.49 | 61.46 | 65.98 | 69.06 | 54.83 | 42.05 | 73.30 | 45.36 | 62.55 | 64.05 | 65.57 | 50.18 | 75.43 | 46.88 |
| | | Accuracy | 62.94 | 71.57 | 60.78 | 52.94 | 73.53 | 61.76 | 67.65 | 69.61 | 56.86 | 47.06 | 73.53 | 47.06 | 62.94 | 64.71 | 66.67 | 52.94 | 75.49 | 47.06 |
| | | Depression-P | 65.49 | 68.91 | 64.06 | 59.14 | 73.68 | 62.14 | 79.67 | 77.27 | 62.10 | 45.50 | 74.31 | 46.64 | 62.76 | 64.28 | 66.57 | 53.78 | 75.77 | 47.02 |
| | | Depression-R | 54.12 | 78.43 | 60.78 | 52.94 | 73.53 | 61.76 | 47.06 | 56.86 | 37.25 | 47.06 | 73.53 | 47.06 | 63.53 | 68.63 | 72.55 | 52.94 | 75.49 | 47.06 |
| | | Depression-F1 | 59.07 | 72.96 | 59.99 | 20.00 | 72.77 | 58.06 | 58.70 | 65.25 | 45.92 | 25.00 | 70.88 | 35.71 | 62.56 | 65.34 | 67.72 | 38.46 | 74.24 | 43.75 |
| | | Control-F1 | 66.00 | 69.58 | 59.23 | 66.67 | 74.22 | 64.86 | 73.25 | 72.86 | 63.75 | 59.09 | 75.71 | 55.00 | 62.54 | 62.76 | 63.43 | 61.90 | 76.61 | 50.00 |

Table 6: Performance of different models and feature sets, evaluated for each task and modality combination.

| Feature Set | Metric | A | T | V | N | A | A | T | T | T | V | A | T | T | T | T | T | A | V | N |
|-------------|--------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|---|---|---|---|
| Classific | Mean | 72.98 | 68.24 | 73.24 | 66.33 | 74.47 | 68.76 | 68.51 | 75.00 | 65.72 | 74.46 | 75.39 | 75.39 | 71.53 | 76.22 | 76.47 | | | | |
| | Std | ±6.60 | ±1.71 | ±6.01 | ±1.84 | ±1.66 | ±1.78 | ±8.91 | ±7.61 | ±8.07 | ±4.53 | ±6.93 | ±6.93 | ±4.52 | ±0.44 | ±0.00 | | | | |
| Foundation | Mean | 75.43 | 66.05 | 77.30 | 66.38 | 67.53 | 77.27 | 70.43 | 78.39 | 72.50 | 76.38 | 80.36 | 71.10 | 78.42 | 79.30 | 79.39 | | | | |
| | Std | ±6.13 | ±6.84 | ±7.47 | ±4.40 | ±0.24 | ±1.46 | ±2.95 | ±1.74 | ±6.14 | ±2.87 | ±1.67 | ±1.60 | ±1.69 | ±0.08 | ±0.00 | | | | |

Table 7: Performance comparison of modality fusion.

| Feature Set | Metric | INT | PDT | VFT | INT PDT | INT VFT | PDT VFT | INT PDT VFT |
|------------------|--------|-------|-------|-------|------------|------------|------------|-------------------|
| Classific | Mean | 73.49 | 68.51 | 70.31 | 76.37 | 72.43 | 74.32 | 76.46 |
| | Std | ±2.90 | ±8.34 | ±3.29 | ±3.04 | ±4.59 | ±1.58 | ±2.94 |
| Foundation Model | Mean | 76.78 | 78.23 | 77.17 | 78.34 | 77.38 | 77.42 | 79.38 |
| | Std | ±4.90 | ±4.81 | ±4.62 | ±1.70 | ±4.58 | ±4.51 | ±2.98 |

Table 8: Performance comparison of task fusion.

| Strategy | Metric | Zero-Shot | | | | | | | Few-shot | | | | | | |
|-----------------------|---------------|-----------|-------------|-----------|--------|-------------|----------|--------------|----------|-------------|-----------|--------|-------------|----------|--------------|
| | | GPT-4o | DeepSeek-v3 | Qwen2(NT) | GPT-03 | DeepSeek-TI | Qwen3(T) | Qwen2.5-Omni | GPT-4o | DeepSeek-v3 | Qwen2(NT) | GPT-03 | DeepSeek-TI | Qwen3(T) | Qwen2.5-Omni |
| Direct Prediction | Macro-F1 | 53.52 | 52.41 | 53.87 | 52.18 | 53.23 | 55.85 | 60.26 | 52.27 | 59.98 | 51.11 | 51.47 | 60.38 | 58.95 | 38.64 |
| | Accuracy | 53.54 | 52.53 | 54.55 | 52.53 | 53.54 | 56.57 | 62.63 | 57.58 | 60.61 | 51.52 | 51.52 | 61.62 | 60.61 | 45.45 |
| | Depression-P | 63.37 | 63.1 | 70 | 65 | 61.34 | 62.94 | 65.17 | 60 | 66.8 | 63.64 | 60 | 66.25 | 64.44 | 52.00 |
| | Depression-R | 45.61 | 42.11 | 36.84 | 38.6 | 52.63 | 59.65 | 75.44 | 78.95 | 63.16 | 36.84 | 47.37 | 68.42 | 70.18 | 68.42 |
| | Depression-F1 | 53.03 | 50.44 | 48.28 | 48.35 | 56.57 | 61.15 | 69.91 | 68.18 | 64.9 | 46.67 | 52.94 | 67.28 | 67.18 | 59.09 |
| Reasoning | Control-F1 | 54.01 | 54.37 | 59.46 | 56.01 | 49.9 | 50.54 | 50.62 | 36.36 | 55.06 | 55.56 | 50 | 53.48 | 50.71 | 18.18 |
| | Macro-F1 | 59.53 | 49.3 | 55.09 | 52.45 | 54.17 | 57.22 | 45.4 | 59.58 | 50.62 | 54.25 | 64.09 | 37.88 | 60.69 | 43.72 |
| | Accuracy | 59.6 | 49.49 | 55.56 | 52.53 | 54.55 | 57.58 | 45.45 | 59.6 | 51.52 | 54.55 | 64.65 | 56.57 | 61.62 | 45.45 |
| | Depression-P | 71.89 | 59.01 | 62.96 | 63.46 | 62.2 | 64.71 | 53.16 | 70.7 | 65.66 | 62.58 | 70.41 | 57.27 | 66.7 | 52.46 |
| | Depression-R | 49.12 | 42.11 | 56.14 | 42.11 | 54.39 | 57.89 | 40.35 | 50.88 | 33.33 | 50.88 | 66.67 | 96.49 | 66.67 | 54.39 |
| Psychiatric Reasoning | Depression-F1 | 58.31 | 48.86 | 59.25 | 50.6 | 57.95 | 61.11 | 45.83 | 59.15 | 44.13 | 55.79 | 68.45 | 71.85 | 66.65 | 53.35 |
| | Control-F1 | 60.76 | 49.73 | 50.93 | 54.29 | 50.4 | 53.33 | 44.98 | 60.01 | 57.12 | 52.7 | 59.74 | 3.92 | 54.72 | 34.1 |
| | Macro-F1 | 60.53 | 54.69 | 57.54 | 55.68 | 55.56 | 61.6 | 46.36 | 60.61 | 56.48 | 55.91 | 61.06 | 48.42 | 61.42 | 37.94 |
| | Accuracy | 60.61 | 55.56 | 57.58 | 56.57 | 64.55 | 61.62 | 46.46 | 60.61 | 56.57 | 57.58 | 61.62 | 58.59 | 61.62 | 38.38 |
| | Depression-P | 72.52 | 72.58 | 69.23 | 75.19 | 52.63 | 73.08 | 54.96 | 71.43 | 65.79 | 62.26 | 67.94 | 59.28 | 69.36 | 44.44 |
| Reasoning | Depression-R | 50.88 | 36.84 | 47.37 | 36.84 | 57.83 | 52.63 | 42.11 | 52.63 | 50.88 | 66.67 | 63.16 | 89.47 | 59.65 | 26.32 |
| | Depression-F1 | 59.67 | 48.68 | 56.25 | 49.43 | 52.75 | 61.17 | 47.59 | 60.61 | 57.28 | 64.33 | 65.31 | 71.31 | 64.13 | 33.00 |
| | Control-F1 | 61.39 | 60.7 | 58.82 | 61.93 | 55.29 | 62.03 | 45.13 | 60.61 | 55.68 | 47.48 | 56.81 | 25.54 | 58.71 | 42.88 |

Table 9: Performance of different LLMs and reasoning strategies.