It Hears, It Sees too: Multi-Modal LLM for Depression Detection By Integrating Visual Understanding into Audio Language Models

Xiangyu Zhao[1], Yaling Shen[1], Yiwen Jiang[1], Zimu Wang[2], Jiahe Liu[1], Maxmartwell H Cheng[1], Guilherme C Oliveira[1], Robert Desimone[5], Dominic Dwyer[1,3,4]*, Zongyuan Ge[1]*

[1]Monash University, [2]University of Liverpool,
[3]Orygen, [4]University of Melbourne, [5]McGovern Institute for Brain Research, MIT

Correspondence: dominic.dwyer@orygen.org.au, zongyuan.ge@monash.edu

Abstract

Depression is one of the most prevalent mental health disorders globally. In recent years, multi-modal data, such as speech, video, and transcripts, has been increasingly used to develop AI-assisted depression assessment systems. Large language models (LLMs) have further advanced this field due to their strong language understanding and generalization capabilities. However, conventional LLMs remain text-centric and cannot process the rich nonverbal cues found in audio and visual modalities, which are critical components in mental health evaluation. While multi-modal LLMs offer a promising direction, few are tailored for psychological applications. In this study, we propose a novel multi-modal LLM framework for depression detection. Our approach augments an audio language model with visual understanding and aligns audio-visual features at the timestamp level. This fine-grained alignment improves modeling of temporal dynamics across modalities while reducing the need for extensive training data and computational resources. Experiments on the DAIC-WoZ and E-DAIC datasets demonstrate that our model consistently outperforms both single-modality approaches and previous multi-modal methods. Moreover, the proposed framework can be extended to incorporate additional physiological signals, paving the way for broader clinical applications beyond mental health.

1 Introduction

Depression has emerged as a critical concern in the field of mental health, affecting a broad population across various age groups. Particularly, the incidence of depression among adolescents has surged over the past decade, raising significant social and public health concerns (Thapar et al., 2022). Diagnosing and treating depression often entails substantial labor and financial costs for both families and healthcare systems. With the advancement of natural language processing (NLP), increasing attention has been given to automated approaches for depression detection, reducing human intervention. Large language models (LLMs) have demonstrated remarkable capabilities across a wide array of NLP tasks (Naveed et al., 2023), which has sparked interest in their application to mental health screening (Hengle et al., 2024; Xu et al., 2024). Despite their success, a fundamental limitation of conventional LLMs lies in their confinement to textual inputs, lacking the capacity to interpret multi-modal signals such as speech and facial expressions that are also indicative of depressive symptoms (Koops et al., 2023; Krause et al., 2021).

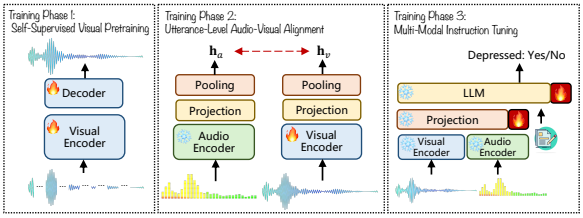


Figure 1: The training scheme of the proposed multi-modal LLM for depression detection.

Multi-modal data, including acoustic and visual cues, can significantly enhance the accuracy of depression detection. Prior studies have shown that individuals at high risk of depression often exhibit reduced facial expressiveness, diminished vitality, and weakened responses to external stimuli such as decreased eye contact (Perez and Riggio, 2003; Waxer, 1974). Similarly, specific acoustic features, such as monotonous tone, slow speech rate, disfluency, and low vocal energy, have been linked to depressive states (Koops et al., 2023). These behavioral signals offer valuable complementary information beyond what can be derived from text alone. Multi-modal large language models (MLLMs) offer an ideal solution to the integration of text and multi-modal data, which shows great promise in

*Corresponding authors.

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 4185–4196
November 6-10, 2026 ©2026 Association for Computational Linguistics

a lot of downstream tasks (Zhang et al., 2024a). However, current MLLMs face several limitations that hinder their application to depression detection. First, depression detection relies heavily on temporal data such as audio and video, yet most existing MLLMs are limited to static images (Caffagni et al., 2024). Furthermore, due to the relatively small size of depression-related datasets compared to standard NLP corpora, developing MLLMs for this domain demands careful consideration of model complexity to mitigate overfitting and ensure training efficiency.

To address these limitations, we propose a simple yet effective framework that adapts a multimodal large language model for depression detection. Our method builds upon a pretrained audio language model (ALM) and augments it with visual understanding capabilities, forming a truly multimodal system. This design leverages the shared temporal structure of audio and visual modalities, allowing for the alignment at the timestamp level. By incrementally integrating visual modules into the ALM with self-supervised visual pretraining and parameter-efficient fine-tuning (PEFT) (Hu et al., 2022), our approach maintains the efficiency and modularity of the base model while enhancing its multi-modal capacity. This strategy also reduces the number of trainable parameters and mitigates the need for large-scale pretraining, making it efficient in data usage and computational requirements. Experiments on two public depression detection datasets, DAIC-WoZ and E-DAIC (Gratch et al., 2014; Ringeval et al., 2019), confirm the effectiveness of our approach, highlighting its potential for practical applications in mental health assessment.

In summary, the contributions of this work consist of the following aspects:

- We develop a multi-modal large language model for depression detection based on the Qwen2-Audio (Chu et al., 2024) model by integrating a self-supervised vision encoder with parameter-efficient fine-tuning. To the best of our knowledge, this is the first study to propose **multi-modal depression detection using LLM across text, audio, and video modalities**;
- We implement a timestamp-level alignment strategy that enables fine-grained temporal fusion across modalities. This design leverages the inherent temporal characteristics of both audio and video signals, enhancing the model's capacity to capture subtle behavioral cues indicative of depression.
- We validate our approach by the comparison with single-modality methods and previous LLM-based state-of-the-art methods on both the DAIC-WoZ and E-DAIC databases (Gratch et al., 2014; Ringeval et al., 2019). The experimental results demonstrate that our approach yields superior performance (achieving F1-scores of 0.825 and 0.857, respectively) at a smaller model scale (7B versus 13B), compared with pioneering multi-modal LLMs.

2 Method

2.1 Overview of the Framework

We propose a multi-modal large language model (MLLM) for depression detection, constructed upon a pretrained audio language model (ALM) as the backbone. As depicted in Figure 2, the framework consists of three key components: (1) an **audio encoder** that processes raw audio signals and extracts temporal embeddings; (2) a **visual encoder** that receives video frames and produces visual embeddings aligned with the audio stream at the timestamp level; (3) a **large language model** that integrates the audio-visual features along with textual inputs to perform depression classification.

The training process is divided into three sequential stages. First, the visual encoder is pretrained using a self-supervised learning strategy inspired by masked autoencoders (He et al., 2022), which enhances its capacity to capture rich visual representations. In the second stage, the visual encoder is fine-tuned on a contrastive alignment task designed to match visual and audio embeddings at the utterance level, thereby improving cross-modal temporal synchronization. Finally, the projection layer and LLM are trained using parameter-efficient fine-tuning (PEFT) techniques to effectively incorporate the visual modality while minimizing additional computational overhead.

2.2 Model Components

2.2.1 Audio Language Model

We adopt Qwen2-Audio (Chu et al., 2024) as the foundation of our framework. This model integrates Whisper-large-v3 (Radford et al., 2023) as the audio encoder and Qwen2-7B as the language model. The audio encoder processes raw wave-

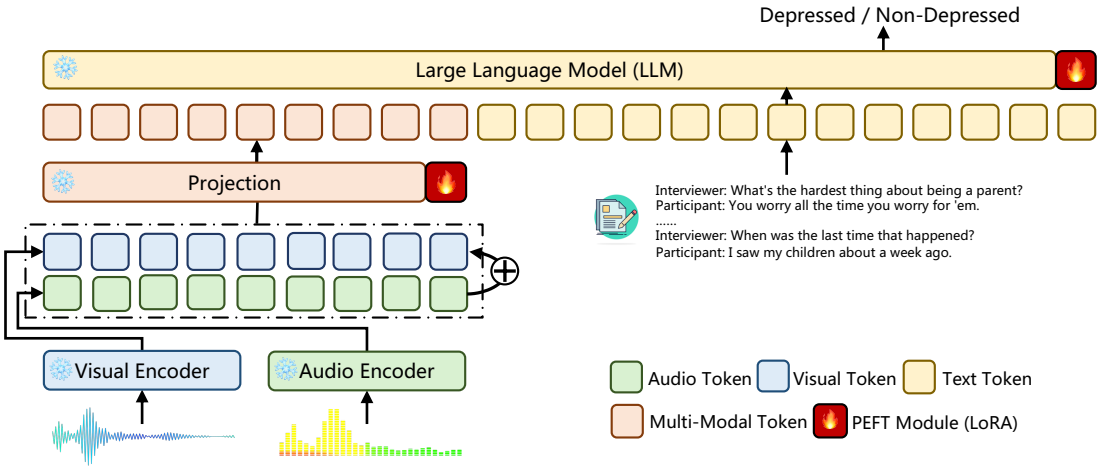


Figure 2: The framework of the proposed multi-modal large language model. The model includes an audio encoder, a visual encoder, and an LLM for detection.

forms resampled to 16 kHz and converts them into 128-channel Mel-spectrograms, with each frame representing a 10 ms segment. These spectrograms are subsequently downsampled via strided convolutions and average pooling, resulting in encoder outputs where each frame corresponds to a 40 ms segment of the original waveform. To ensure the universality of our method, we retain the pretrained weights of Qwen2-Audio throughout the initial stages and apply PEFT-based adaptation only in the final training phase. Notably, our framework is modular and can be extended to other audio language models, provided their audio encoders output sequences aligned with fixed temporal intervals.

2.2.2 Visual Encoder

The visual encoder is designed to extract visual embeddings that align temporally with the audio encoder outputs. To ensure architectural compatibility and ease of alignment, its design mirrors the Whisper encoder, comprising a strided convolutional embedding layer, a stack of Transformer encoder layers, and an output average pooling layer. Initially, visual features are resampled to match the temporal resolution of the audio Mel-spectrograms and are projected into the embedding space via 1D convolutions. This embedding process includes striding, reducing the temporal resolution to 20 ms per token. The resulting features are then processed by the Transformer layers and further downsampled through average pooling to match the final 40 ms resolution of the audio encoder outputs. As a result, both audio and visual embeddings are temporally synchronized, as illustrated in Figure 3.

2.2.3 Audio-Visual Projection

After obtaining audio and visual embeddings, the next step is to fuse them into a unified representation for input into the LLM. While a common fusion strategy involves concatenating modality embeddings along the sequence dimension (Xu et al., 2025), this approach is suboptimal for integrating new modalities into pretrained LLMs, as it disrupts the expected sequence length and can interfere with positional encoding. To preserve compatibility with pretrained LLMs, we propose a simple yet effective fusion method—element-wise addition of audio and visual embeddings, which is illustrated in Figure 2. This is feasible due to our explicit timestamp-level synchronization, ensuring both sequences share the same temporal structure. Moreover, our three-stage training strategy progressively aligns the modalities, enabling effective fusion without representation collapse.

2.3 Timestamp-Synchronized Data Augmentation

Depression corpora typically consist of participant–interviewer interviews, which present two challenges: (1) severe class imbalance, as healthy controls far outnumber depressed individuals, and (2) limited data volume, despite long session durations. To alleviate these issues, we adopt subdialogue shuffling based on Wu et al. (2023), segmenting lengthy interviews into shorter, contiguous exchanges. This increases sample size per participant and enables flexible resampling for class balancing.

Building on Wu et al. (2023), we enhance the method by ensuring timestamp alignment across transcript, audio, and visual modalities. Each subdi-

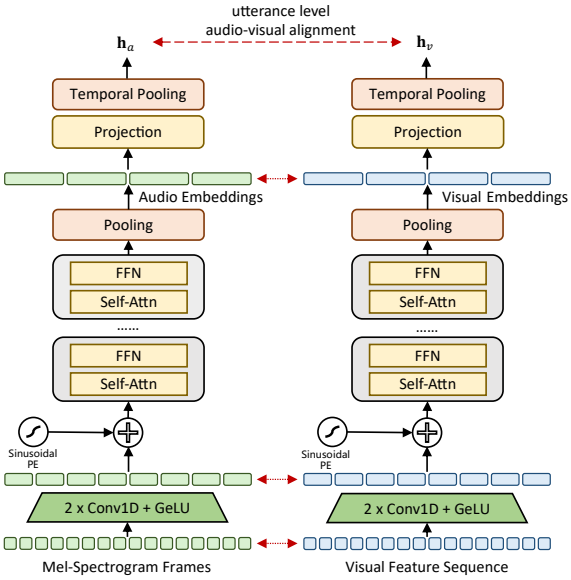


Figure 3: The scheme of utterance-level audio-visual alignment. Audio and visual inputs are strided simultaneously, ensuring synchronization on timestamps.

alogue is constrained to start with an interviewer's utterance and end with the participant's response, maintaining contextual coherence and narrowing the domain gap between LLM pretraining and depression detection. We then discard interviewer audio and corresponding visual frames, retaining only participant segments, while preserving interviewer transcripts. This choice reflects two considerations: interviewer speech carries little acoustic value for mental state assessment, yet their utterances are essential for conversational coherence. Although removing interviewer segments inevitably discards some multimodal information, the trade-off between information reduction and coherence is analyzed in Section 3.3.3. Further augmentation details are provided in the Appendix.

2.4 Training

The training pipeline of our framework is divided into three sequential stages, as shown in Figure 1. The first two stages focus on training the visual encoder, while the final stage involves fine-tuning the LLM.

2.4.1 Self-Supervised Visual Pretraining

To enhance the visual representation capability of the encoder, we first conduct self-supervised pretraining. Instead of learning directly from raw video data, we opt to pretrain on pre-extracted visual features, as raw video files may contain sensitive content and are often unavailable in commonly used depression-related corpora. This not only addresses potential privacy concerns but also reduces computational overhead, making the approach more generalizable to other time-series modalities such as physiological signals (e.g., rPPG and ECG).

Inspired by the masked autoencoder (MAE) framework (He et al., 2022), we design a reconstruction task where the encoder learns to recover masked portions of the input time series. Specifically, given a sequence input $\mathbf{x} = (x_1, x_2, ..., x_T) \in \mathbb{R}^{T\times d}$, we randomly mask K frames of the input and use a learnable token $x_{mask} \in \mathbb{R}^d$ shared across all masked frames. The indices for masked tokens are denoted as $\mathcal{M}$, and the indices for unmasked tokens are denoted as $\mathcal{V}$. Obviously $\mathcal{V} \cup \mathcal{M} = \{1, 2, \ldots, T\}$. The unmasked sequence $\mathbf{x}_{in} = \{x_i | i \in \mathcal{V}\} \in \mathbb{R}^{K\times d}$ are fed to the visual encoder to acquire the latent representation $\mathbf{h} \in \mathbb{R}^{T-K\times d}$. Then, the latent representation $\mathbf{h}$ and masked frames are concatenated together to acquire the input sequence $\mathbf{z} \in \mathbb{R}^{T\times d}$, which are fed to the decoder to obtain the reconstructed input sequence $\hat{\mathbf{x}} \in \mathbb{R}^{T\times d}$. The objective is to minimize the mean squared error (MSE) between the reconstructed and original sequences within the masked regions:

$$\min \frac{1}{|\mathcal{M}|} \sum_{i\in\mathcal{M}} ||\hat{x}_i - x_i||_2^2 \qquad (1)$$

This approach allows the model to capture temporal dependencies and improve robustness in downstream tasks.

2.4.2 Utterance Level Audio-Visual Alignment

After the visual pretraining in the first stage, the visual encoder is enabled to extract visual embeddings from input visual feature sequences for downstream tasks. However, the visual comprehension of the visual encoder is not aligned with the audio encoder. To reduce the training gap between both encoders, we design a proxy downstream task with contrastive learning to align the visual encoder with the audio encoder at the utterance level. As illustrated in Figure 3, we add a projection layer to each encoder, respectively, and pool the outputs in the time dimension to obtain the utterance level representations. Given a mini-batch of audio outputs $\mathbf{h}_a \in \mathbb{R}^{N\times d}$ and visual outputs $\mathbf{h}_v \in \mathbb{R}^{N\times d}$, where N denotes the batch size, we obtain a similarity matrix $\mathbf{Sim} = \mathbf{h}_a\mathbf{h}_v^T \in \mathbb{R}^{N\times N}$. The learning objective is to find the correct match of each

audio-visual pair for utterance level audio-visual alignment:

$$\min \mathcal{L}_{ce}(\mathbf{Sim}/\tau, \mathbf{I}_N) \quad (2)$$

where $\mathcal{L}_{ce}$ denotes cross-entropy loss, $\mathbf{I}_N$ denotes the identity matrix, and τ is the temperature parameter.

During this stage, we freeze the entire audio encoder and the lower layers of the visual encoder to preserve the representations learned in the initial stage. Only the upper layers of the visual encoder receive gradient updates, ensuring stability and preventing catastrophic forgetting.

2.4.3 Multi-Modal Instruction Tuning

In the final stage, we integrate the pretrained visual encoder with the audio language model to construct a multi-modal large language model tailored for depression detection, which is illustrated in Figure 2. Since traditional LLMs are not inherently designed to process visual information, additional instruction tuning is required to adapt the model to this task. We employ Low-Rank Adaptation (LoRA) (Hu et al., 2022) to update the parameters of both the LLM and the modality projection layer. As audio and visual features have been temporally synchronized and aligned at the utterance level in previous stages, the complexity of cross-modal fusion is substantially reduced.

2.5 Multi-Scale Sliding-Window Inference

Since our model is trained on subdialogues rather than entire conversations, we adopt a multi-scale sliding-window inference strategy to derive a final prediction for each full conversation. This approach aggregates predictions from multiple subdialogue segments extracted at different temporal scales. Specifically, for each conversation, we generate a fixed number (200) of subdialogues at three predefined durations: 30s, 75s, and 120s. This multi-scale design ensures that each temporal resolution contributes equally to the final decision, capturing both short-term and long-term behavioral cues. The overlap between adjacent subdialogues is dynamically adjusted based on the conversation length and the total number of segments per setting. Each time-scale configuration yields an independent conversation-level prediction, and the final prediction is determined by majority voting across the three settings.

3 Experiments

3.1 Database and Implementation Details

We utilize two widely recognized benchmarks to develop and evaluate our proposed multi-modal LLM: the DAIC-WoZ database (Gratch et al., 2014) and its extended version, E-DAIC (Extended DAIC-WoZ) (Ringeval et al., 2019).Both datasets include interview audio, transcripts, and derived visual features with PHQ-8 ≥ 10 indicating depression; E-DAIC covers Wizard-of-Oz and autonomous interviews. The DAIC-WoZ dataset consists of 189 participants, divided into a training set of 107 subjects, a development set of 35 subjects, and a test set of 47 subjects. To further verify the robustness and generalization of our method, we also employ the E-DAIC dataset, which expands the original corpus to a total of 275 participants, including a training set of 163 samples, a development set of 56 samples, and a test set of 56 samples. Following standard protocols (Wu et al., 2023; Zhang et al., 2024b), we train on the respective training sets and, leveraging the released golden labels, report performance on both the development and test sets to ensure comprehensive comparison.

For data augmentation, we employ a timestamp-synchronized strategy with a maximum subdialogue length of 120 seconds, generating 1,000 samples per depressed case to balance data diversity and overfitting risks. These augmented features facilitate self-supervised visual pretraining, utterance-level audio-visual alignment, and final multi-modal instruction fine-tuning. Our model is built upon the Qwen2-Audio-7B-Instruct backbone and trained using 2 NVIDIA H200 GPUs. Detailed hyperparameters are provided in the Appendix.

3.2 Results

Following established evaluation protocols, we adopt the F1 score as the primary metric. To investigate the effectiveness of our framework components and benchmark against baseline approaches, we conduct ablation studies and comprehensive comparisons with single-modal methods, conventional multi-modal approaches, and existing multi-modal LLMs primarily on the DAIC-WoZ development set. The detailed results of these analyses are presented in Table 1 and Table 2. Furthermore, to demonstrate the robustness and superior generalization of our method, we report comparative results against current state-of-the-art approaches on the test sets of both the DAIC-WoZ and E-DAIC

Modality	Models	F1
Text	RoBERTa 2022	0.602
	Llama2-7B 2024b	0.578
	Llama2-13B 2024b	**0.636**
	Qwen2-7B 2024	0.564
	GPT4 2024b	0.571
Audio	HuBERT 2023	0.640
	WavLM 2023	**0.720**
	SpeechFormer 2022	0.694
	SpeechFormer++ 2023	0.709
	Whisper-v3 2023	0.694
Video	GSM 2016	0.530
	SSL + CLS	**0.668**
A+T	AudiBERT 2021	0.709
	TOAT 2022	0.741
	LSTM 2018	**0.770**
A+T+V	C-CNN 2018	0.769
	ConvBiLSTM 2022	0.70*
	Ours w/o MS	0.789
	Ours	**0.844**

Table 1: The performance comparison of our method and other approaches on DAIC-WoZ development set. "*" denotes that the original results are reported with 2 significant digits. "MS" denoting the multi-scale strategy in our inference.

datasets, as illustrated in Table 3.

Evaluation on DAIC-WoZ Dev Set We present a comprehensive comparison between our proposed multi-modal LLM and previous methods on the DAIC-WoZ development set in Table 1. Additionally, we evaluate the contribution of each individual module in our framework, including the Qwen2-7B model, the Whisper-v3 audio encoder, and a self-supervised vision encoder. Overall, our multi-modal model achieves superior classification performance on the development set of the DAIC-WoZ dataset, consistently outperforming all single-modality baselines.

Text-based models show that Llama2-13B (Touvron et al., 2023; Zhang et al., 2024b) performs best among text-only models, likely due to its larger parameter scale. Among smaller models, Qwen2-7B and Llama2-7B exhibit similar performance but fall short of the 13B variant. Interestingly, GPT-4, despite its scale and zero-shot capabilities, underperforms relative to Llama2-13B. Likewise, RoBERTa surpasses GPT-4 despite its significantly smaller size as well. A similar phenomenon has been observed in Zhang et al. (2024b). This performance gap may be attributed to the nature of depression detection, which emphasizes representation learning over generative modeling, making encoder-based models more suitable.

Audio-based models generally outperform text-only models, suggesting that acoustic cues carry richer information for detecting depressive symptoms. In addition, the performance of audio models could benefit from downstream tasks such as speech recognition or emotion recognition (Wu et al., 2023). Notably, WavLM fine-tuned for emotion recognition shows superior performance, surpassing even Whisper-v3-large. This suggests that tasks closely related to depression, such as emotion recognition and ASR, provide transferable knowledge useful for this application.

For video models, our finetuned visual encoder with a classification head achieves the best performance. The main factor that could affect video-based models is the choice of visual feature sets. Since raw videos are not available at the DAIC-WoZ database, only facial feature sets, such as landmarks and action units, are available for depression detection. As the feature set could be rather redundant, the performance of video models could even deteriorate if the feature set selection is inappropriate. Self-supervised pretraining alleviates the issue significantly, as masked autoencoders are designed for images, which possess a redundant nature, and are suitable in our scenario.

Multi-modal approaches that incorporate modalities generally outperform single-modal baselines. In particular, the inclusion of audio features often leads to significant performance improvements, highlighting the importance of acoustic information in depression detection. Compared with other multi-modal methods, our proposed framework consistently achieves superior results, demonstrating the effectiveness of timestamp-level alignment and the synergy of modality-specific encoders in capturing clinically relevant cues.

Comparison with Multi-Modal LLMs Table 2 presents the performance comparison between our method and existing multi-modal LLMs. Together with Table 1, the results demonstrate that incorporating audio significantly enhances the classification performance of LLMs. For instance, augmenting Llama2-13B with acoustic landmarks improves its F1 score from 0.636 to 0.695. A similar trend is observed with Qwen2-7B, where the inclusion of audio elevates the F1 score from 0.578 to 0.720. Our proposed multi-modal framework, which jointly models text, audio, and visual signals, achieves the highest F1 score of 0.789, validat-

Model	Base Model	F1
Acoustic LLM (Zhang et al., 2024b)	7B	0.545
	7B-Chat	0.500
	13B	0.695
	13B-Chat	0.666
Qwen2-Audio (Chu et al., 2024)	7B	0.650
	7B-Instruct	0.720
Ours w/o audio	7B	0.617
	7B-Instruct	0.643
Ours	7B	0.709
	7B-Instruct	**0.789**

Table 2: The performance comparison of our method and multi-modal LLMs on DAIC-WoZ development set. Note that for fair comparison we do not employ model ensemble or multi-scale inference.

ing the benefit of integrating visual cues alongside audio and language inputs. This underscores the advantage of leveraging complementary modalities for capturing the complex and multi-faceted nature of depressive symptoms. Notably, both our approach and Qwen2-Audio variants outperform LLMs with acoustic landmarks, despite relying on smaller language backbones (7B vs 13B). This suggests that native multi-modal architectures might be more adept at interpreting raw sensory inputs. While acoustic landmarks serve as a lightweight representation of audio, they may omit subtle prosodic or emotional cues that are preserved in the original waveforms. In contrast, models trained end-to-end on raw audio exhibit stronger modality comprehension and more effective feature fusion.

Comparison with Previous State-of-the-Art We compare our method with previous state-of-the-art approaches on both the DAIC-WoZ and E-DAIC benchmarks, as presented in Table 3. On the DAIC-WoZ dataset, single-modal baselines (e.g., GloVe-CNN (Campbell et al., 2022) and TOAT (Guo et al., 2022)) yield relatively lower performance with F1 scores below 0.70. In contrast, multi-modal approaches integrating Audio, Text, and Video information, such as HiQuE (Jung et al., 2024) and MultiDepNet (Kumar et al., 2025), achieve significantly better results, reaching F1 scores of 0.79 and 0.785, respectively. Our method outperforms these existing baselines, achieving a superior F1 score of 0.825. Furthermore, on the E-DAIC benchmark, our approach demonstrates robust generalization capabilities. While recent methods like MLlm-DR (Zhang et al., 2025) and LLM w/ Facial Feat (Sadeghi et al., 2024) achieve competitive F1 scores of 0.79 and 0.743, our method sets a new state-of-the-art with an F1 score of 0.857. Overall, our proposed framework consistently outperforms previous single-modal and multi-modal approaches across both datasets.

3.3 Ablation Studies and Discussion

In this section, we analyze the source of performance gain in our framework, including the contribution of each modality and the selection of the base model. In addition, we discuss the effectiveness of our proposed timestamp-synchronized data augmentation upon the removal of the interviewer's utterance and context length in subdialogues. The experiments are all conducted on the development set of DAIC-WoZ.

3.3.1 The Contribution of Each Modality

We further investigate the individual contribution of each modality within our framework. As shown in Table 1, both audio and video modalities enhance depression detection performance. The baseline Qwen2-7B model achieves an F1 score of 0.564 using text alone. Introducing audio features leads to a substantial improvement, raising the F1 score to 0.720. Further incorporation of video features elevates the performance to 0.789. Additionally, our proposed multi-scale sliding-window strategy contributes to model performance significantly, improving the F1 score to 0.844.

An interesting observation is that the addition of audio yields a greater performance gain compared to the inclusion of video, in both instruction-tuned and pre-trained variants. This discrepancy can be attributed to two primary factors. First, as pre-extracted visual features rather than raw video data are utilized in our framework, the model may face information loss, leading to reduced expressive power. Second, our model is fundamentally built upon an audio language modeling architecture. Removing audio embeddings may disrupt the alignment mechanism across modalities, thereby compromising the model's ability to integrate non-verbal cues effectively.

3.3.2 The Choice of Base Model

Since both pretrained model and instruction-tuned model are available in Qwen2-Audio families, we compare the performance of these two model variants as the base model. The results in Table 2 indicate that the instruction-tuned model provides higher detection performance. The findings in our

Dataset	Models	Modality	F1
DAIC-WoZ	GloVe-CNN (Campbell et al., 2022)	Text	0.68*
	TOAT (Guo et al., 2022)	Audio	0.647
	EmoAudioNet (Othmani et al., 2021)	Audio	0.66*
	HiQuE (Jung et al., 2024)	A+T+V	0.79*
	MultiDepNet (Kumar et al., 2025)	A+T	0.785
	Ours	A+T+V	**0.825**
E-DAIC	HiQuE (Jung et al., 2024)	A+T+V	0.70*
	LLM w/ Facial Feat (Sadeghi et al., 2024)	A+T+V	0.743
	MultiDepNet (Kumar et al., 2025)	A+T	0.768
	MLlm-DR (Zhang et al., 2025)	A+T+V	0.79*
	Ours	A+T+V	**0.857**

Table 3: The performance comparison of our method and previous approaches on DAIC-WoZ and Extend-DAIC database. "*" denotes that the original results are reported with 2 significant digits.

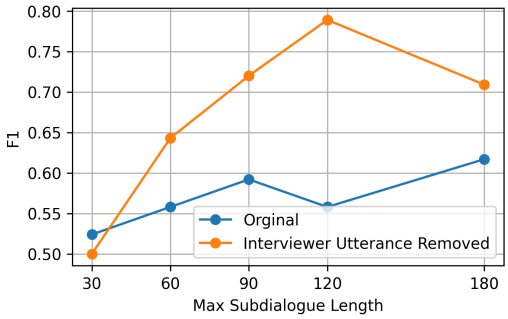


Figure 4: The depression detection performance on subdialogue length with or without interviewer's utterances.

research are different from previous work (Zhang et al., 2024b), where instruction tuning leads to significant performance deterioration compared with the pretrained model. The reasons for the inconsistency could be the difference in instruction tuning in general LLMs and audio language models. Depression detection involves the analysis of both audio and text; a similar task has been used to finetune the model in instruction tuning. Thus, the instruction-tuned model could be better at the audio analysis task.

3.3.3 The Effect of Context Length and Interviewer Utterance Removal

The length of subdialogues plays a crucial role in our framework, as longer contexts generally provide richer cues for depression detection. However, longer subdialogues do not necessarily improve the performance for detection, as the audio records for the interviewer do not contribute to the decision, but even interfere with the depression detection. To address this constraint, we propose to remove the interviewer's utterances during data augmentation, allowing more content from the participant to be retained within the fixed audio window. While this enhances the availability of participant-specific acoustic cues, it also results in the loss of visual information associated with the removed segments. To explore this trade-off, we conduct an ablation study under varying subdialogue lengths, as shown in Figure 4. When the maximum subdialogue length is constrained to 30 seconds, removing the interviewer's speech leads to degraded performance. In this setting, the entire subdialogue can be encoded without truncation, and discarding the interviewer's turns causes unnecessary loss of visual cues, thus impairing multi-modal inference. In contrast, as the subdialogue length increases beyond the model's audio capacity, the removal of interviewer utterances proves beneficial. By prioritizing participant speech within the fixed input window, the model gains access to more relevant acoustic information, leading to improved detection accuracy. However, when the context length becomes excessively long, the performance gain diminishes. This is likely due to reduced dialogue diversity and increased risk of overfitting, as longer subdialogues tend to be less variable.

4 Conclusion

In this study, we propose a novel multi-modal large language model for depression detection, augmenting an audio-based backbone with timestamp-aligned visual capabilities. Extensive experiments on both the DAIC-WoZ and E-DAIC benchmarks demonstrate the superiority and robustness of our framework over existing state-of-the-art methods. To the best of our knowledge, this is the first work to effectively integrate textual, audio, and visual modalities within an LLM for this task. Overall, this work presents a robust solution for adapting multi-modal LLMs to mental health assessment, with promising potential for broader clinical appli-

cations.

Limitations

Despite the promising results achieved by our framework, there are several limitations to this study that should be acknowledged. Although our framework incorporates utterance-level audio-visual alignment, it assumes a consistent temporal correspondence between modalities, which may fall short in capturing intricate or long-range multimodal interactions that are indicative of depressive symptoms. Furthermore, our proposed method has not been validated in the scenarios where more than one extra modality is introduced to the base model. Extending the framework to accommodate multiple additional physiological signals (e.g., EEG or ECG) simultaneously remains a direction for future exploration.

Ethics Statement

This study uses the DAIC-WoZ and E-DAIC databases, which the data provider distributes to academic and other non-profit researchers upon application because of consent constraints. The research release includes participant audio, transcripts, and derived visual features; its controlled access helps protect participant privacy. We obtained DAIC-WoZ under its End-User License Agreement and adhere to its terms of use. Our model is built upon the Qwen2-Audio architecture, and all research activities related to it comply with the Apache-2.0 license under which the model is released. While our method achieves state-of-the-art performance on the DAIC-WoZ benchmark, it is intended for research purposes only and should not be used for clinical diagnosis, treatment, or intervention of depression. We further acknowledge that, like many large language models, our framework may be vulnerable to hallucinations, harmful outputs, or systemic biases. We disclaim responsibility for any misuse, misinterpretation, or unintended consequences resulting from the deployment of this model outside its intended research context.

Data Availability

The datasets utilized in this study, DAIC-WoZ and E-DAIC, are access-controlled research resources available by application to academic and other non-profit researchers. Access to these datasets can be obtained upon request via the following URL: https://dcapswoz.ict.usc.edu/.

Use of AI Tools

In the preparation of this manuscript, we utilized AI assistants, specifically ChatGPT and Gemini 3 Pro, for grammatical polishing and editing to enhance the clarity and readability of the text. The scientific concept, experimental design, data analysis, and conclusions remain the sole work and responsibility of the authors.

References

Tuka Al Hanai, Mohammad M Ghassemi, and James R Glass. 2018. Detecting depression with audio/text sequence modeling of interviews$f$. In *Interspeech*, pages 1716–1720.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2024. The revolution of multimodal large language models: a survey. *arXiv preprint arXiv:2402.12451*.

Edward L Campbell, Laura Docıo-Fernández, Nicholas Cummins, and Carmen Garcıa-Mateo. 2022. Speech and text processing for major depressive disorder detection. *Training*, 31:76.

Weidong Chen, Xiaofen Xing, Xiangmin Xu, Jianxin Pang, and Lan Du. 2022. Speechformer: A hierarchical efficient framework incorporating the characteristics of speech. In *Proc. Interspeech 2022*, pages 346–350.

Weidong Chen, Xiaofen Xing, Xiangmin Xu, Jianxin Pang, and Lan Du. 2023. Speechformer++: A hierarchical efficient framework for paralinguistic speech processing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:775–788.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

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.

Yanrong Guo, Chenyang Zhu, Shijie Hao, and Richang Hong. 2022. A topic-attentive transformer-based model for multimodal depression detection. *arXiv preprint arXiv:2206.13256*.

Albert Haque, Michelle Guo, Adam S Miner, and Li Fei-Fei. 2018. Measuring depression symptom severity from spoken language and 3d facial expressions. *arXiv preprint arXiv:1811.08592*.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009.

Amey Hengle, Atharva Kulkarni, Shantanu Patankar, Madhumitha Chandrasekaran, Sneha D'silva, Jemima Jacob, and Rashmi Gupta. 2024. Still not quite there! evaluating large language models for comorbid mental health diagnosis. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16698–16721.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Juho Jung, Chaewon Kang, Jeewoo Yoon, Seungbae Kim, and Jinyoung Han. 2024. Hique: Hierarchical question embedding network for multimodal depression detection. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 1049–1059.

Sanne Koops, Sanne G Brederoo, Janna N de Boer, Femke G Nadema, Alban E Voppel, and Iris E Sommer. 2023. Speech as a biomarker for depression. *CNS & Neurological Disorders-Drug Targets-CNS & Neurological Disorders)*, 22(2):152–160.

Fernando C Krause, Eftihia Linardatos, David M Fresco, and Michael T Moore. 2021. Facial emotion recognition in major depressive disorder: A meta-analytic review. *Journal of affective disorders*, 293:320–328.

Puneet Kumar, Shreshtha Misra, Zhuhong Shao, Bin Zhu, Balasubramanian Raman, and Xiaobai Li. 2025. Multimodal interpretable depression analysis using visual, physiological, audio and textual data. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5305–5315. IEEE.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

Alice Othmani, Daoud Kadoch, Kamil Bentounes, Emna Rejaibi, Romain Alfred, and Abdenour Hadid. 2021. Towards robust deep neural networks for affect and depression recognition from speech. In *International conference on pattern recognition*, pages 5–19. Springer.

John E Perez and Ronald E Riggio. 2003. Nonverbal social skills and psychopathology. *Nonverbal behavior in clinical settings*, pages 17–44.

Rafał Poświata and Michał Perełkiewicz. 2022. OPI@LT-EDI-ACL2022: Detecting signs of depression from social media text using RoBERTa pre-trained language models. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 276–282, Dublin, Ireland. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, and 1 others. 2019. Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop*, pages 3–12.

Misha Sadeghi, Robert Richer, Bernhard Egger, Lena Schindler-Gmelch, Lydia Helene Rupp, Farnaz Rahimi, Matthias Berking, and Bjoern M Eskofier. 2024. Harnessing multimodal approaches for depression detection using large language models and facial expressions. *npj Mental Health Research*, 3(1):66.

Anita Thapar, Olga Eyre, Vikram Patel, and David Brent. 2022. Depression in young people. *The Lancet*, 400(10352):617–631.

Ermal Toto, ML Tlachac, and Elke A Rundensteiner. 2021. Audibert: A deep transfer learning multimodal classification framework for depression screening. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 4145–4154.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Peter Waxer. 1974. Nonverbal cues for depression. *Journal of Abnormal Psychology*, 83(3):319.

Ping-Cheng Wei, Kunyu Peng, Alina Roitberg, Kailun Yang, Jiaming Zhang, and Rainer Stiefelhagen. 2022. Multi-modal depression estimation based on sub-attentional fusion. In *European Conference on Computer Vision*, pages 623–639. Springer.

James R Williamson, Elizabeth Godoy, Miriam Cha, Adrianne Schwarzentruber, Pooya Khorrami, Youngjune Gwon, Hsiang-Tsung Kung, Charlie

Dagli, and Thomas F Quatieri. 2016. Detecting depression using vocal, facial and semantic communication cues. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*, pages 11–18.

Wen Wu, Chao Zhang, and Philip C Woodland. 2023. Self-supervised representations in speech-based depression detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

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*.

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K Dey, and Dakuo Wang. 2024. 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*, 8(1):1–32.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengping Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024a. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*.

Wei Zhang, Juan Chen, En Zhu, Wenhong Cheng, YunPeng Li, and Yanbo J Wang. 2025. Mllm-dr: Towards explainable depression recognition with multimodal large language models. *arXiv preprint arXiv:2507.05591*.

Xiangyu Zhang, Hexin Liu, Kaishuai Xu, Qiquan Zhang, Daijiao Liu, Beena Ahmed, and Julien Epps. 2024b. When llms meets acoustic landmarks: An efficient approach to integrate speech into large language models for depression detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 146–158.

A Implementation Details

Our framework is implemented using the HuggingFace *transformers* library with PyTorch 2.1. The full hyperparameter configurations used during training are summarized in Table 4. We adopt the AdamW optimizer (Loshchilov and Hutter, 2017) for model optimization. To improve training speed without compromising performance, we enable TensorFloat32 (TF32) computation and apply automatic mixed-precision training using BFloat16 (BF16). For parameter-efficient fine-tuning (PEFT) of the Qwen2-Audio model on the depression detection task, we employ QLoRA (Dettmers et al., 2023), which compresses the base model to 4-bit precision to reduce memory usage and improve computational efficiency. The full training process requires approximately 90+ GPU hours on an NVIDIA H200 141GB GPU. This includes around 40 hours for self-supervised visual pretraining, 20 hours for utterance-level audio-visual alignment, and 30 hours for multimodal instruction tuning. Early stopping is applied in all stages when training loss plateaus.

	Stage I	Stage II	Stage III
Optimizer		AdamW	
Learning Rate	1.5e-4	1e-6	3e-6
β_1		0.9	
β_2	0.95	0.999	
Weight Decay	0	0.001	
Batch Size	128	64	8
Grad Accum Steps	8	16	8
Scheduler		Cosine LR	
Num Epochs	50	20	3
Warm Up Epochs	5	2	0.1
Max Grad Norm	1.0	0.5	
BF16		True	
TF32		True	

Table 4: Training hyperparameters.

B Details of Timestamp-Synchronized Data Augmentation

Following the approach of Wu et al. (2023), we generate subdialogues from the original interview transcripts to mitigate class imbalance and expand the size of the training set. In our data augmentation pipeline, we enforce strict synchronization among transcripts, audio, and video to ensure precise timestamp-level alignment. However, due to varying frame rates across modalities, achieving synchronization presents a technical challenge. For example, audio recordings are typically captured at a 16,000 Hz sampling rate and later converted into Mel-spectrograms with a frame rate of 100 Hz, while video recordings are collected at 30 frames per second (FPS). To address this discrepancy, we constrain the start and end timestamps of each subdialogue to align with whole seconds (i.e., integer-second boundaries).

Additionally, we require each subdialogue to begin with an utterance from the interviewer and conclude with a response from the participant. This design choice ensures that each subdialogue forms a complete and contextually coherent conversational

Algorithm 1 Time-Sync Data Augmentation

$N^+ \leftarrow$ Number of positive samples in the training set
$N^- \leftarrow$ Number of negative samples in the training set
Set number of subdialogues per positive sample M^+
Set minimum length of subdialogue in seconds d_{min}
Set maximum length of subdialogue in seconds d_{max}
$M^- = N^-/N^+ \times M^+ \leftarrow$ Number of sub-dialogues per negative sample
for Dialogue $X^{(n)} = (T^n, A^n, V^n), n = 1, 2, ..., N$ **do**
  $D \leftarrow$ Dialogue length in seconds
  $\{\varepsilon_i\} \leftarrow$ Interviewer utterance start timestamps
  $\{\varepsilon_p\} \leftarrow$ Participant utterance end timestamps
  if $X^{(n)}$ is positive **then**
    $M \leftarrow M^+$
  else
    $M \leftarrow M^-$
  end if
  for Sub-dialogue $X^{(n)m}$, $m = 1$ to M **do**
    Sample length d uniformly from (d_{min}, d_{max})
    Sample start timestamp $\varepsilon'_s \in \{\varepsilon_i\}$ from range (0, $D - d$)
    Round the start timestamp to its closet integer second $\varepsilon_s \leftarrow \lfloor \varepsilon'_s \rfloor$
    $\varepsilon_{tmp} = \varepsilon_s + d \leftarrow$ Raw end timestamp
    Sample end timestamp $\varepsilon'_e \in \{\varepsilon_p\}$ and $\min |\varepsilon'_e - \varepsilon_{tmp}|$
    Round the end timestamp to its closet integer second $\varepsilon_e \leftarrow \lceil \varepsilon'_e \rceil$
    Generate subdialogue $T^{(n)m} \leftarrow T^{(n)}_{\varepsilon_s:\varepsilon_e}$
    Obtain the raw audio segment $A'^{(n)m} \leftarrow A^{(n)}_{\varepsilon_s:\varepsilon_e}$
    Obtain the raw visual segment $V'^{(n)m} \leftarrow V^{(n)}_{\varepsilon_s:\varepsilon_e}$
    Remove the interviewer utterances $A^{(n)m} \leftarrow A'^{(n)m}$ and $V^{(n)m} \leftarrow V'^{(n)m}$
    Subdialogue $X^{(n)m} = (T^{(n)m}, A^{(n)m}, V^{(n)m})$
  end for
end for

unit, with a clear initiation and response structure. Such a constraint preserves the semantic continuity and logical flow within each segment, making them more suitable for downstream tasks that rely on natural discourse patterns. Moreover, this structure aligns with the training paradigm of large language models, which are typically pretrained on large-scale dialogue corpora. By maintaining this dialogue consistency, we enhance the model's ability to interpret the subdialogues effectively within a familiar conversational framework.

C The Prompt Design for Instruction Tuning

During instruction tuning, we design a system prompt to guide the behavior of the language model. Given that the Qwen2-Audio-Instruct model has been fine-tuned on audio analysis tasks, we adopt a chat-based prompt template to elicit model responses. Notably, we use the same prompt design for both the Qwen2-Audio family and our multimodal LLM. This consistency is based on our integration strategy, where visual embeddings are directly added to the audio embeddings without modifying the model architecture. Therefore, we assume that the model can still function effectively even without explicitly referencing visual information in the prompt.

System Prompt `Below is a conversation between an interviewer and a participant. Please analyze the transcripts and audio, and find whether the participant is affected by depression.`

Instructions `Audio: {audio} \n Interview conversation: {transcripts} \n Response: \n`