ParaCalib: Semantically Calibrated Paralinguistic Modeling for Depression Detection

Yuxin Li[1,†], Yifei Li[1,†], Yi-Wen Chao[1], Xiangyu Zhang[2]
Eng Siong Chng[1], Cuntai Guan[1,*]

[1]Nanyang Technological University, Singapore
[2]University of New South Wales, Sydney, Australia

[†]These authors contributed equally.
[*]Corresponding author: CTGuan@ntu.edu.sg

Abstract

Vocal behavior provides important signals for speech-based depression detection, but its interpretation often depends on what is being said and how it functions in context. However, existing methods typically treat acoustic cues as context-independent markers, making it difficult to distinguish vocal form from its context-dependent communicative function. We propose ParaCalib, a framework that semantically calibrates paralinguistic behavior by interpreting vocal patterns relative to utterance-level semantic context and inferred communicative function and representing them in a comparable state space. Concretely, ParaCalib uses an Audio-Language Model (ALM) to generate contextualized vocal descriptions and an LLM-based Paralinguistic State Extractor (PSE) to map these descriptions into a structured Semantically Calibrated Paralinguistic (SC-Para) representation. ParaCalib achieves the highest mean Macro-F1 among the evaluated methods, reaching 71.9% on DAIC-WOZ and 90.5% on MODMA. Our controlled analysis provides direct evidence of semantic calibration at the PSE stage: under a fixed caption-level acoustic description, varying the accompanying semantic context changes the inferred depression-related paralinguistic evidence. Our exploratory analysis further identifies recurring configurations of paralinguistic states associated with depression labels rather than a single uniformly dominant state.

1 Introduction

Depression is a prevalent mental disorder associated with substantial disability and impaired daily functioning (World Health Organization, 2025). Early and scalable screening is central to computational mental health research (Galatzer-Levy and Onnela, 2023). Speech is a promising modality for this purpose because it is non-invasive, relatively easy to collect during clinical interviews or remote interactions, and psychologically informative, capturing both what a person says and how they say it (Cummins et al., 2015; Li et al., 2025b).

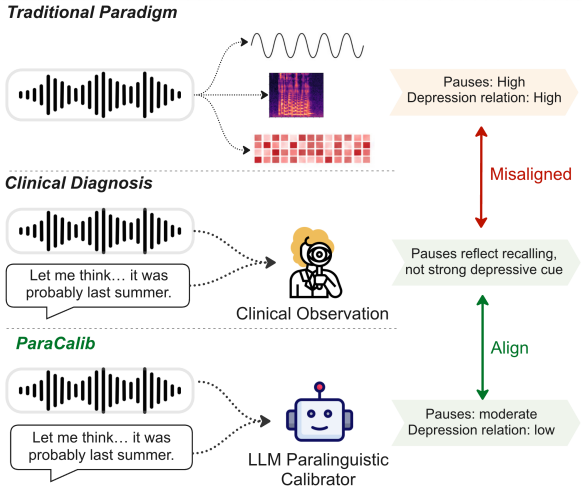

Figure 1: Illustration of the mismatch between context-independent acoustic-marker modeling and semantic calibration of paralinguistic cues. The same acoustic pattern can receive different interpretations depending on utterance-level semantic context and inferred communicative function.

Most speech-based depression detection methods follow what we refer to as an acoustic-marker paradigm: they seek vocal patterns that are statistically associated with depression and use them as participant-level indicators. Early approaches rely on low-level descriptors (LLDs), including prosodic, spectral, voice-quality, and temporal features, motivated by clinical links between depression and psychomotor slowing, longer pauses, reduced vocal energy, monotonic prosody, and altered voice quality (Ellgring and Scherer, 1996; Cummins et al., 2015; Yamamoto et al., 2020). More recent end-to-end and self-supervised (SSL) models learn representations from raw waveforms, time-frequency inputs, or large unlabeled speech corpora (Ma et al., 2016; Chen et al., 2022b;

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 27197–27214
October 24-29, 2026 ©2026 Association for Computational Linguistics

Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2022a; Wu et al., 2023; Li et al., 2025a). Despite their architectural differences, these approaches generally map acoustic form directly to depression-related evidence, without explicitly modeling how utterance-level semantic context and inferred communicative function shape the interpretation of vocal behavior.

This mismatch arises for two reasons. First, depression is clinically heterogeneous: major depressive episodes may involve either psychomotor retardation or agitation, suggesting that depression-related speech need not follow a single monotonic acoustic profile (American Psychiatric Association, 2022; Alpert et al., 2001). For example, depressive speech may be slower and more paused in some cases, but tense or accelerated in others. Second, the same acoustic form can serve different communicative functions. A long pause, flattened pitch contour, reduced intensity, or slowed speaking rate may reflect depressive psychomotor changes, but may also result from hesitation, emphasis, syntactic planning, turn-taking, topic shifts, or the semantic content being expressed (Cole, 2015; Wagner and Watson, 2010). In short, depression can manifest through different vocal patterns, while the same vocal pattern can serve different communicative functions. As illustrated in Figure 1, an acoustic pattern that appears strongly depression-related under the traditional paradigm may receive a different interpretation when considered together with the utterance and its communicative context.

Motivated by this observation, we propose ParaCalib,[1] a framework for representing utterance-level vocal behavior as semantically calibrated paralinguistic states. We define semantic calibration as interpreting vocal behavior relative to utterance-level semantic context and inferred communicative function and expressing the resulting interpretation in a structured, comparable state space. ParaCalib operationalizes this process by combining ALM-generated contextualized vocal descriptions with PSE mapping into predefined SC-Para dimensions. For each participant, the valid utterance-level SC-Para vectors form an instance bag, which is aggregated by an attention-based MIL classifier for participant-level prediction.

Experiments on DAIC-WOZ and MODMA show that ParaCalib achieves the highest mean Macro-F1 among the evaluated methods, outperforming conventional acoustic and SSL speech representations while providing inspectable intermediate paralinguistic states. Controlled analysis further shows that, under a fixed caption-level acoustic description, the PSE-derived evidence changes with semantic context, providing direct evidence that semantic calibration is operationalized at the PSE stage. Exploratory analysis identifies specific combinations of states associated with depression labels. Together, these findings support structured, context-sensitive, and compositional modeling of vocal behavior for speech-based depression detection.

2 Related Work

2.1 Depression Detection from Speech

The AVEC challenges have established widely used benchmarks for computational depression assessment (Valstar et al., 2013; Ringeval et al., 2019). Early systems primarily relied on handcrafted prosodic, spectral, temporal, and voice-quality descriptors (Li et al., 2025b). Their dependence on manually selected features motivated CNN-, RNN-, and Transformer-based models that learn representations directly from waveforms or time–frequency inputs (Chlasta et al., 2019; Dubagunta et al., 2019; Saidi et al., 2020; Othmani et al., 2021; Vázquez-Romero and Gallardo-Antolín, 2020; Wang et al., 2022; Salekin et al., 2018; Zhao et al., 2020; Muzammel et al., 2020).

Pretrained speech encoders, including wav2vec 2.0 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), WavLM (Chen et al., 2022a), and Whisper (Radford et al., 2023), further improve representation transfer in data-limited settings (Zhang et al., 2021; Toto et al., 2021; Chen et al., 2022b; Ravi et al., 2022; Wu et al., 2023; Li et al., 2025a; Zhang et al., 2025b). Despite their different architectures, these approaches generally use learned representations as generic features for participant-level prediction rather than explicitly modeling how vocal behavior should be interpreted in context.

Several studies incorporate linguistic information or examine its potential confounding effects. Syntactic information has been used to guide acoustic feature extraction and multimodal fusion (Morales et al., 2018). DepFlow addresses semantic shortcut reliance by synthesizing examples

[1] Code is available at https://github.com/yifeili-13/ParaCalib.

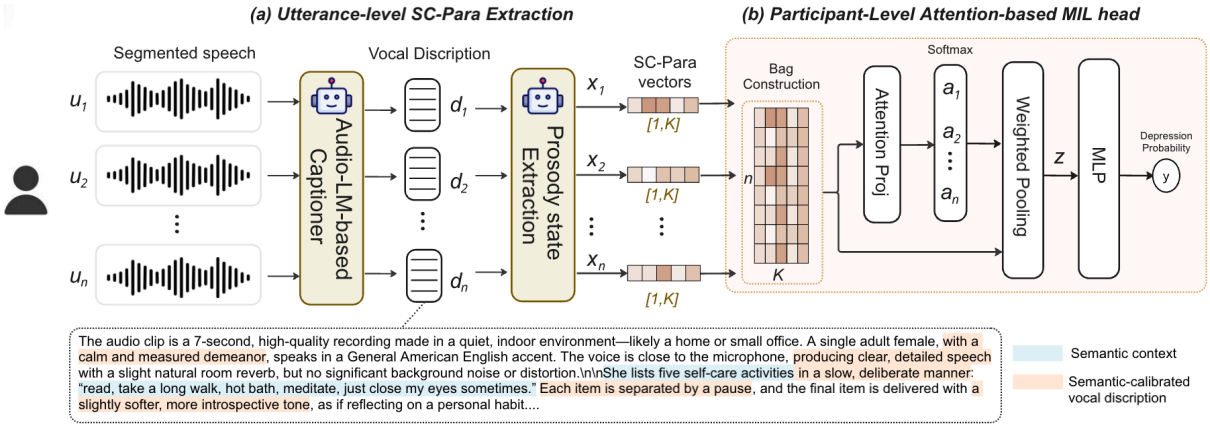

Figure 2: Overview of ParaCalib. A frozen ALM captioner converts utterance-level speech into contextualized vocal descriptions, and PSE structures those descriptions into utterance-level SC-Para vectors. The attention-based MIL classifier uses an attention projection only to estimate utterance-level instance weights, forms the participant-level representation as a weighted sum of the original seven-dimensional SC-Para vectors, and feeds this pooled representation to an MLP for depression-label prediction.

that decouple depression-conditioned vocal expression from linguistic sentiment (Li et al., 2026). These studies demonstrate that linguistic content can provide relevant contextual information while also introducing spurious label associations. However, they do not explicitly represent how the meaning of a vocal pattern changes with utterance-level semantic context and inferred communicative function.

2.2 Audio-Language Models for Affective Speech Understanding

Audio-Language Models (ALMs), such as SALMONN, Qwen-Audio, Qwen2-Audio, and GAMA, connect audio encoders with pretrained language models to support general audio understanding (Tang et al., 2024; Chu et al., 2023, 2024; Ghosh et al., 2024). In affective speech, these models have been applied to closed-set emotion recognition, free-form vocal captioning, and prediction refinement grounded in transcripts and acoustic cues (Zhang et al., 2025a; Xu et al., 2024). Related work has also explored acoustic-grounded reasoning (Tian et al., 2025), ordinal modeling of affective dimensions (Chandra et al., 2025), and multimodal LLMs for depression assessment (Zhang et al., 2024, 2025b; Zhao et al., 2025; Chen et al., 2026).

These approaches show that ALMs can generate semantically rich descriptions of vocal behavior and support flexible affective reasoning. However, free-form descriptions are difficult to compare consistently across utterances and participants, while closed-set affective classifiers often compress vocal behavior into coarse categories. Existing systems primarily optimize prediction or explanation generation rather than constructing a structured representation that preserves context-dependent paralinguistic distinctions.

ParaCalib addresses this gap by mapping contextualized vocal descriptions into fixed Semantically Calibrated Paralinguistic (SC-Para) dimensions. This design combines the contextual sensitivity of ALM-generated descriptions with the comparability of a structured state space. Unlike generic acoustic representations, SC-Para explicitly conditions the interpretation of vocal behavior on utterance-level semantic context and inferred communicative function. Unlike free-form captions or explanations, it supports participant-level aggregation and representation-level analysis.

3 Methodology

3.1 Overview

As illustrated in Figure 2, ParaCalib consists of three stages: ALM-based vocal captioning, PSE-based SC-Para extraction, and attention-based MIL aggregation. Given an utterance-level audio segment, the frozen ALM generates a natural-language description that captures both the spoken content and the accompanying vocal behavior. The frozen PSE then converts this description into a fixed-dimensional SC-Para vector. Finally, the attention-based MIL classifier aggregates utterance-level SC-Para vectors into a participant-level representation for depression-

label prediction. The ALM and PSE remain frozen throughout training, and only the downstream MIL classifier is optimized.

3.2 Utterance-level ALM Captioning

Let u_i denote the i-th utterance-level audio segment. We use a frozen Qwen3-Omni-30B-A3B Captioner (Xu et al., 2025)[2] to generate a natural-language vocal description:

$$d_i = \text{ALM}(u_i). \tag{1}$$

The ALM processes the audio input as a unified audio-language signal. It can infer both what is being said and how it is being said, allowing it to describe vocal behavior in relation to the utterance content. The generated description may include observable properties such as pitch variability, speaking rate, rhythm, pause pattern, vocal energy, and voice quality, together with their interpretation in the context of the utterance.

The captioner receives only the utterance-level audio; no external text prompt is supplied because the captioning instruction is built into the specialized checkpoint. The ALM is not fine-tuned on either dataset and does not receive participant-level depression labels. The checkpoint and decoding settings are provided in Appendix A.3.

3.3 SC-Para Extraction

Although ALM-generated descriptions provide rich information about vocal behavior, their free-form language is difficult to compare consistently across utterances and participants. We therefore use a frozen DeepSeek-R1-Distill-Qwen-32B model (Guo et al., 2025)[3] as the PSE to convert each description d_i into a fixed SC-Para schema:

$$x_i = \text{PSE}(d_i), \tag{2}$$

where $x_i \in \mathbb{R}^K$ and $K = 7$.

In the main evaluation, the frozen PSE receives only the ALM-generated utterance description; it is not provided with participant identifiers, PHQ scores, or dataset labels.

The SC-Para schema contains five constituent paralinguistic dimensions and two holistic dimensions. The five constituent dimensions describe pitch variability, speaking rate, rhythm regularity, pause pattern, and vocal energy. These dimensions reflect paralinguistic phenomena commonly examined in clinical and affective speech research (Cummins et al., 2015; Koops et al., 2023). Each constituent dimension is assigned one of three discrete states and encoded numerically as 0, 1, or 2. The two holistic dimensions are a depression-related paralinguistic evidence score and a confidence score. The evidence score is generated on a 0–100 scale and normalized to $[0, 1]$, while the confidence score is represented directly on $[0, 1]$. These scores are PSE-generated intermediate variables rather than utterance-level gold depression labels, and provide additional information derived from the PSE's pretrained linguistic and affective knowledge. Their contribution is evaluated through component-level ablations and prompt-sensitivity analyses in Appendix B.

For each utterance u_i, the PSE produces

$$x_i = [x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(K)}]. \tag{3}$$

If the PSE assigns `unknown` to any categorical dimension because the description provides insufficient vocal evidence, the utterance is excluded before participant-level aggregation. The two holistic scores are retained for all utterances with valid categorical states. Discard rates are reported in Table 13 of Appendix C. The PSE prompt template and an abbreviated output schema are provided in Appendix D.

3.4 Attention-based MIL Classification

For participant b, let

$$S_b = \{x_{b,1}, x_{b,2}, \ldots, x_{b,n_b}\} \tag{4}$$

denote the set of valid utterance-level SC-Para vectors, where n_b may vary across participants. Each participant is treated as a bag of utterance-level instances.

The attention-based MIL classifier assigns an importance score to each instance using an additive attention function:

$$e_{b,i} = v_{\text{att}}^{\top} \tanh\left(W_{\text{att}} x_{b,i} + b_{\text{att}}\right), \tag{5}$$

where W_{att}, b_{att}, and v_{att} are learnable parameters. The attention weights are normalized across the utterances belonging to the same participant:

$$\alpha_{b,i} = \frac{\exp(e_{b,i})}{\sum_{j=1}^{n_b} \exp(e_{b,j})}. \tag{6}$$

[2] https://huggingface.co/Qwen/Qwen3-Omni-30B-A3B-Captioner

[3] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-32B

The utterance-level vectors are pooled into a participant-level representation:

$$z_b = \sum_{i=1}^{n_b} \alpha_{b,i} x_{b,i}. \tag{7}$$

The attention module estimates the relative contribution of each utterance, while the readout preserves the original seven-dimensional SC-Para representation. The pooled representation z_b is then passed to a two-hidden-layer multilayer perceptron (MLP), which produces a scalar logit s_b for binary depression-label classification:

$$\hat{y}_b = \sigma(s_b), \tag{8}$$

where $\sigma(\cdot)$ denotes the sigmoid function. The architecture and training hyperparameters are reported in Appendix A, Table 3.

3.5 Optimization Objective

The ALM and PSE remain frozen, and only the attention module and downstream MLP are optimized. We use weighted binary cross-entropy with positive-class weight $w_+ = N_{\text{neg}}/N_{\text{pos}}$ and negative-class weight $w_- = 1$. The remaining optimization settings are provided in Appendix A.

4 Data and Evaluation

4.1 Datasets

We evaluate ParaCalib on two speech-based depression datasets: DAIC-WOZ (Gratch et al., 2014) and MODMA (Cai et al., 2022).

DAIC-WOZ contains 189 English-speaking participants and provides an official participant-disjoint training, development, and test split. We define depression-label status using a PHQ-8 threshold of 10 (Kroenke et al., 2009). Following prior preprocessing (Bailey and Plumbley, 2021), we correct the label assignment of participant 409. The resulting cohort contains 57 depression-label and 132 non-depression-label participants.

MODMA contains 52 Mandarin-speaking participants, with 18 interview utterances per participant. Although the original participant groups were established through psychiatric assessment, including the MINI under DSM-IV criteria, our experiments use PHQ-9-defined depression labels with a threshold of 10 (Kroenke et al., 2001). Following prior preprocessing (Li et al., 2025a), this criterion reassigns participant 2010037 to the non-depression-label group, resulting in 22 depression-label and 30 non-depression-label participants. Thus, the MODMA results concern PHQ-9-based labels rather than the original diagnostic grouping. Label definitions and preprocessing details are provided in Appendix A.

4.2 Evaluation Protocol

For DAIC-WOZ, we use the official training, development, and test partitions. For MODMA, we perform participant-level five-fold cross-validation and repeat each fold with the same five random seeds. Within each fold, all preprocessing, model selection, and classifier training are conducted without access to the held-out participants.

We report participant-level Macro-F1 as the primary metric, together with sensitivity, specificity, and area under the receiver operating characteristic curve (AUC). Unless otherwise stated, thresholded predictions use a fixed decision threshold of 0.5. All results are reported as mean and standard deviation (SD) in percentage points. Specifically, DAIC-WOZ results are aggregated over five matched random seeds. For MODMA, metrics are first averaged over the five seeds within each fold, and the reported mean and SD are then computed across the five participant-level fold means. Reproduced acoustic, end-to-end, and self-supervised baselines, including SpeechFormer, use the same participant-level splits, preprocessing procedure, model-selection protocol, and held-out evaluation protocol (Ma et al., 2016; Wang et al., 2023; Wu et al., 2023; Chen et al., 2022b; Li et al., 2025a). Implementation and model settings are provided in Appendix A.

4.3 Statistical and Diagnostic Analyses

Representation comparisons evaluate complete representation strategies under the same attention-based MIL classifier and training protocol. For each comparison, we conduct a two-sided paired t-test across the five matched evaluation units defined above and report the 95% confidence interval for the mean paired difference. These comparisons characterize variation across matched runs or fold-level evaluation units rather than uncertainty obtained by resampling participants. When multiple comparisons are performed within the same analysis family, we apply Holm correction to the corresponding p-values. Additional details of the statistical tests are provided in Appendix A.

We perform several analyses to examine the properties of SC-Para. First, the semantic-context

Dataset	Method	Input	Macro-F1 (↑)	Sensitivity (↑)	Specificity (↑)	AUC (↑)
DAIC-WOZ	DepAudioNet (Ma et al., 2016)*	MFbanks	50.7 (1.3)	67.1 (8.1)	44.8 (5.1)	54.1 (2.6)
	NUSD (Wang et al., 2023)*	Raw speech	51.4 (4.6)	30.0 (13.7)	74.5 (10.2)	57.5 (3.4)
	Wu et al.-style (Wu et al., 2023)*	WavLM	53.5 (5.3)	32.8 (24.0)	78.2 (18.0)	54.7 (7.5)
	SpeechFormer (Chen et al., 2022b)*	Wav2Vec 1.0	58.0 (3.5)	40.0 (3.9)	75.8 (4.3)	54.1 (2.4)
	HAREN-CTC (Li et al., 2025a)*	WavLM	57.9 (2.1)	41.4 (12.0)	**83.6** (7.0)	55.6 (1.0)
	ParaCalib	SC-Para	**71.9** (4.9)	**77.1** (15.5)	73.3 (6.6)	**83.3** (2.6)
MODMA	Wu et al.-style (Wu et al., 2023)*	WavLM	59.7 (4.9)	50.0 (14.0)	70.0 (6.2)	64.8 (6.6)
	DepAudioNet (Ma et al., 2016)*	MFbanks	69.5 (6.3)	52.7 (10.0)	86.0 (4.9)	74.7 (6.0)
	HAREN-CTC (Li et al., 2025a)*	WavLM	73.1 (5.9)	61.4 (8.3)	86.0 (9.2)	76.4 (7.6)
	ParaCalib	SC-Para	**90.5** (11.2)	**91.0** (12.5)	**90.0** (14.9)	**96.4** (6.9)

Table 1: Main participant-level depression-label classification results on DAIC-WOZ and MODMA. * indicates reproduced baselines. Open-source baselines use the released implementations and reported configurations. Wu et al.-style denotes pretrained WavLM without the sub-dialogue augmentation used in the original study. All reproduced methods follow our common preprocessing, model-selection, and held-out test protocol. Results are reported as mean (SD) in percentage points.

control holds the caption-level acoustic description fixed while varying the accompanying context. We compare three conditions: acoustic-only context, non-depression-related context, and depression-related context. The analysis uses 100 utterances from 66 participants. The construction of the context conditions and the corresponding inferential procedure are described in Appendix B.3.

Second, prompt sensitivity is evaluated on the same 100 captions using two meaning-preserving prompt paraphrases, while keeping the SC-Para schema, numerical mappings, and output format fixed. Third, we conduct an exploratory utterance-level contingency-table analysis of associations between categorical SC-Para states and depression labels. Because multiple utterances share the same participant label, the adjusted significance values are used only to screen descriptive patterns and are not interpreted as participant-level confirmatory inference. The construction of the contingency tables and the screening criteria are described in Appendix B.3.

We additionally evaluate speaker-identity leakage using a participant-disjoint speaker-verification probe and assess attribute-wise removal through retrained MIL classifiers. Finally, additive-noise robustness is evaluated by mixing background-noise recordings from the MUSAN noise partition (Snyder et al., 2015) into DAIC-WOZ test utterances at 20, 10, and 5 dB signal-to-noise ratios. These perturbations are applied only during robustness evaluation and do not affect model training or model selection. Additional results are reported in Appendices B and C.

5 Results and Analysis

5.1 Depression Detection Performance

Table 1 reports participant-level classification results on DAIC-WOZ and MODMA. ParaCalib achieves the highest mean Macro-F1 among the evaluated methods on both datasets, reaching 71.9% on DAIC-WOZ and 90.5% on MODMA. It also obtains the highest sensitivity and AUC on both datasets. Relative to HAREN-CTC, ParaCalib improves mean Macro-F1 by 14.0 percentage points on DAIC-WOZ and 17.5 percentage points on MODMA, based on the unrounded matched-unit means. The paired evaluation-unit comparisons yield Holm-adjusted p-values of .0232 and .0734, respectively (Table 4). Split-wise results and exploratory severity-stratified analyses are reported in Appendix A and Appendix C.

5.2 Representation Comparison

Table 2 compares complete representation strategies under the same attention-based MIL evaluation protocol. Discretizing eGeMAPS descriptors improves Macro-F1 from 43.9% to 49.0% on DAIC-WOZ and from 53.4% to 67.8% on MODMA. The strongest generic SSL representations reach 46.3% and 68.8%, respectively, whereas SC-Para reaches 71.9% and 90.5%.

On DAIC-WOZ, transcript-based representations achieve Macro-F1 values between 46.1% and 63.1%. Transcript-masked Caption-PSE reaches 64.6% on DAIC-WOZ and 60.6% on MODMA, indicating that vocal-delivery information remains predictive after explicit lexical content is removed. The larger reduction under content mask-

Dataset	Representation type	Representation	Macro-F1 (↑)	Sensitivity (↑)	Specificity (↑)	AUC (↑)
DAIC-WOZ	LLD-based	LLDs	43.9 (6.4)	35.7 (29.0)	62.4 (34.9)	48.1 (4.4)
		Discretized LLDs	49.0 (6.2)	40.0 (21.2)	61.2 (14.4)	50.4 (6.9)
	SSL-based	Wav2Vec 2.0	41.8 (10.5)	40.0 (25.6)	53.3 (33.2)	42.3 (4.6)
		HuBERT	46.3 (5.9)	37.1 (12.8)	57.0 (6.6)	40.1 (3.5)
		WavLM	41.1 (7.3)	28.6 (34.6)	66.7 (36.9)	46.6 (5.7)
	Transcript-based	Transcript	63.1 (13.3)	52.9 (30.1)	79.4 (12.6)	73.5 (3.2)
		LLD-PSE w/ transcript	54.5 (8.2)	28.6 (18.9)	**84.2** (15.4)	64.2 (6.6)
		Transcript + WavLM	46.1 (11.1)	20.0 (22.2)	81.2 (29.2)	49.7 (12.4)
	Caption-based	Transcript-masked Caption-PSE	64.6 (3.8)	51.4 (9.3)	78.2 (7.8)	77.9 (2.0)
		Caption Embedding	66.6 (14.2)	68.6 (39.3)	75.8 (14.1)	77.1 (12.0)
		SC-Para	**71.9** (4.9)	**77.1** (15.5)	73.3 (6.6)	**83.3** (2.6)
MODMA	LLD-based	LLDs	53.4 (11.5)	35.4 (14.1)	79.1 (12.3)	57.8 (13.4)
		Discretized LLDs	67.8 (19.1)	56.4 (25.6)	82.0 (11.9)	78.7 (19.4)
	SSL-based	Wav2Vec 2.0	55.6 (7.7)	62.6 (24.0)	59.2 (16.8)	63.7 (7.4)
		HuBERT	61.6 (24.7)	62.8 (19.1)	75.1 (18.3)	71.8 (22.3)
		WavLM	68.8 (20.5)	63.2 (25.8)	79.1 (15.9)	76.5 (19.6)
	Caption-based	Transcript-masked Caption-PSE	60.6 (19.8)	58.2 (14.9)	71.9 (17.0)	69.2 (13.3)
		Caption Embedding	85.7 (10.0)	86.4 (9.0)	85.3 (11.0)	94.3 (7.0)
		SC-Para	**90.5** (11.2)	**91.0** (12.5)	**90.0** (14.9)	**96.4** (6.9)

Table 2: Representation analysis of ParaCalib across alternative representation strategies. Results are reported as mean (SD) in percentage points. Shaded cells indicate the best performance within each representation type, while bold values denote the overall best result per dataset and metric. Transcript-based baselines are omitted for MODMA due to the lack of available transcripts. Because the representations differ in upstream pretraining, dimensionality, and extraction complexity, this table compares complete representation strategies under a common downstream evaluation protocol rather than parameter-matched models. Additional extraction configurations and dimension specifications are given in Appendix A.

ing on MODMA indicates a stronger contribution from semantic context on that dataset. Accordingly, ParaCalib integrates contextual semantic and vocal-delivery information rather than isolating a purely acoustic marker. The content-masking prompt is provided in Appendix D.

Direct Caption Embedding reaches 66.6% Macro-F1 on DAIC-WOZ and 85.7% on MODMA. SC-Para further improves the mean Macro-F1 to 71.9% and 90.5%, corresponding to gains of 5.3 and 4.8 percentage points, respectively; the Holm-adjusted paired comparisons yield $p_{\text{adj}} = .7269$ on both datasets. Across both datasets, SC-Para therefore achieves higher mean performance while converting free-form captions into a compact, fixed-dimensional, and inspectable representation. This pattern suggests that structuring contextualized vocal descriptions into semantically calibrated states preserves and may sharpen task-relevant information while enabling dimension-level analysis.

5.3 Semantic-Context Control

We next examine whether semantic context affects PSE interpretation independently of the vocal-behavior description. The controlled analysis holds this description fixed and compares acoustic-only, non-depression-related, and depression-related context conditions. Depression-related context increases the PSE evidence score by 8.6 points relative to non-depression-related context (participant-clustered 95% CI $[6.7, 10.6]$). The corresponding increase relative to acoustic-only input is 8.7 points, whereas non-depression-related context differs from acoustic-only input by only 0.1 points ($p = .92$). These results provide direct evidence that semantic calibration is operationalized at the PSE stage: the same caption-level acoustic description yields different depression-related paralinguistic evidence under different semantic contexts. Additional test statistics are provided in Appendix B.3.

5.4 Component and State-Level Analysis

Attentive MIL achieves the highest Macro-F1 among the evaluated participant-level classifiers on both datasets; the complete classifier comparison is provided in Appendix A.

Figure 3 summarizes the metric-wise performance changes after removing each SC-Para di-

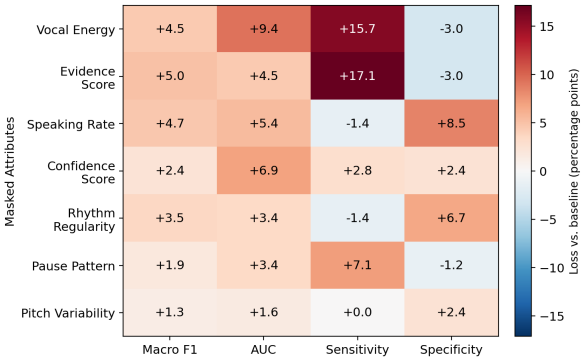

Figure 3: Metric-wise performance changes on DAIC-WOZ after removing one SC-Para dimension and retraining the attentive MIL classifier. Values are calculated as the five-seed mean performance of the full model minus that of the corresponding ablated model and are reported in percentage points. Positive values indicate performance degradation after removal, whereas negative values indicate improvement.

mension and retraining the attentive MIL classifier on DAIC-WOZ. The largest Macro-F1 decreases occur after removing the PSE-derived evidence score (5.0 percentage points), speaking rate (4.7 points), and vocal energy (4.5 points). Their effects differ across metrics: vocal energy and the evidence score contribute primarily to sensitivity, whereas speaking rate contributes more strongly to specificity.

Beyond individual-dimension removal, the five categorical dimensions alone achieve 64.0% Macro-F1, while the two holistic scores alone achieve 64.7% and the evidence score alone achieves 63.2%, compared with 71.9% for the complete representation. Replacing the depression-specific holistic scores with a generic affect/arousal prompt reduces Macro-F1 from 64.7% to 61.3% and AUC from 69.6% to 61.9%. These results suggest that the categorical and holistic dimensions provide complementary predictive information.

Pairwise Spearman correlations among the five categorical dimensions range from -0.400 to 0.703 (Table 8), indicating that the dimensions are related but not redundant.

State-Level and Pairwise Associations. In this exploratory utterance-level analysis, marginal associations are selective rather than uniformly monotonic across states. Low energy, flat pitch, moderate pauses, and specific speaking-rate states show positive associations with depression labels relative to their corresponding reference states. Seven specific combinations of states met the pre-specified frequency and multiplicity-adjusted display criteria; several involve low energy together with slow speech, steady rhythm, few pauses, or flat pitch. Because utterances are clustered within participants, these patterns are descriptive associations rather than participant-level confirmatory evidence. The displayed associations are shown in Figure 4, and the screening criteria are given in Appendix B.3.

5.5 Robustness and Error Analysis

We conduct additional analyses of speaker identity, extraction stability, acoustic robustness, and prediction errors. Additional protocols and results are reported in Appendices C and B.

A participant-disjoint speaker-verification probe obtains an AUC of 52.9% for the full seven-dimensional SC-Para representation, compared with 78.7% for WavLM embeddings. This result suggests that SC-Para contains substantially less speaker-identity information than the generic SSL representation. In an exploratory DAIC-WOZ test-set analysis, depression-label participants showed lower attention-weighted SC-Para dispersion than non-depression-label participants (mean difference $= -0.0327$; Appendix C.3).

Repeated-decoding analysis yields 94–99% exact agreement across SC-Para dimensions. Prompt paraphrase analysis also shows high consistency, with correlations of $r \geq .841$. In contrast, additive acoustic noise degrades performance: mixing background noise at 5 dB SNR reduces Macro-F1 from 71.9% to 62.6%.

In a small internal human validation, both raters judged 18 of 20 captions and 17 of 20 constituent-state sets as consistent with perceived vocal evidence; details are provided in Appendix C.4 and Table 14.

Finally, consistent false positives tend to occur near the PHQ-8 cutoff, with a mean PHQ-8 score of 6.4 compared with 2.6 for reliably classified controls. Consistent false negatives contain fewer valid utterances, averaging 66 utterances compared with 119 for reliably classified true positives. These descriptive error patterns are summarized in Appendix C.

6 Discussion

The results support semantic calibration as a useful modeling principle for depression-related paralin-

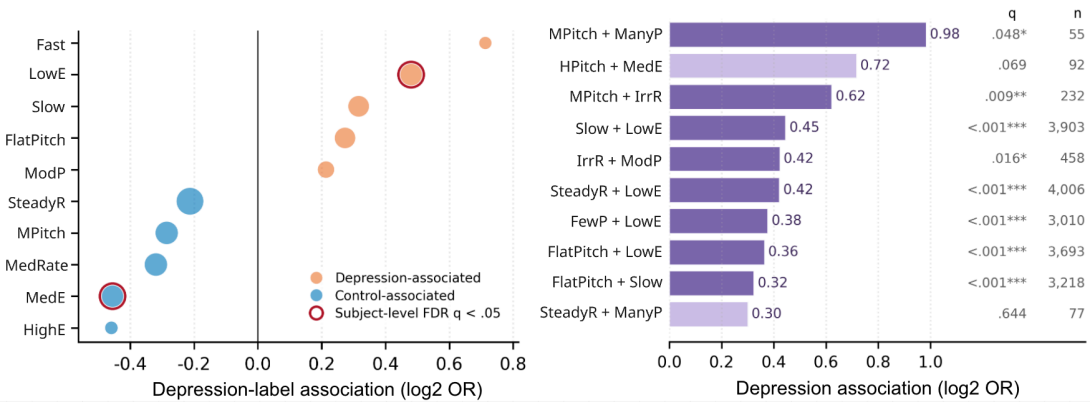

Figure 4: Exploratory utterance-level associations between categorical SC-Para states and participant-level depression labels. Left: marginal $\log_2$ odds ratios. Right: the ten highest-ranked state pairs. We interpret only the seven pairs satisfying $n \geq 100$ and $q < .05$; the remaining three are shown descriptively. The q values do not account for within-participant clustering.

guistic evidence. Rather than treating vocal patterns as context-independent markers, ParaCalib interprets them jointly with utterance-level semantic information and maps the resulting interpretation into a structured state space. Under a fixed caption-level acoustic description, changing the accompanying semantic context shifts the PSE-derived evidence score, demonstrating context sensitivity at the PSE stage.

The representation comparisons further clarify the role of SC-Para. SC-Para substantially outperforms the evaluated handcrafted and SSL-based representations and achieves higher mean Macro-F1 than direct Caption Embedding on both datasets. These consistent numerical improvements indicate that the structured state representation preserves and may refine the task-relevant information contained in free-form captions. Importantly, SC-Para also provides a compact and fixed-dimensional representation whose intermediate paralinguistic dimensions can be compared across utterances and participants. The component analyses further suggest that categorical vocal states and PSE-derived holistic scores provide complementary information.

This structured representation enables predictions to be inspected at the utterance level before participant-level aggregation. State-level associations, attribute-wise removal, and attention-weighted dispersion offer complementary descriptions of the patterns used by the classifier. However, these analyses should not be interpreted as causal feature importance or as direct evidence of clinically observed symptoms. They characterize the behavior of the extracted representation and the downstream classifier under the evaluated datasets and protocols.

7 Conclusion

ParaCalib operationalizes semantic calibration by converting contextualized vocal descriptions into a compact and inspectable SC-Para representation for participant-level depression-label classification. It achieves the highest mean Macro-F1 among all evaluated representation strategies on both DAIC-WOZ and MODMA, yielding higher mean performance than conventional acoustic, SSL-based, transcript-based, and direct caption-embedding alternatives. Controlled analysis further shows that the PSE-derived evidence is sensitive to semantic context under a fixed caption-level acoustic description. Together, these findings support structured, context-sensitive, and compositional modeling of vocal behavior.

8 Limitations

We note several considerations for interpreting and extending ParaCalib. First, as with other speech-based approaches, performance depends on recording conditions and the amount of available speech. Performance decreases under severe additive noise, and participants with fewer valid utterances provide less evidence for participant-level aggregation. These observations motivate future work on noise-robust extraction and uncertainty-aware prediction.

Second, our experimental targets are binary labels derived from PHQ-8 and PHQ-9 cutoffs. Accordingly, the reported results should be inter-

preted as depression-label prediction rather than clinical diagnosis. Further validation on prospectively collected data and independently adjudicated clinical outcomes would be needed before considering clinical deployment.

Third, ParaCalib uses frozen generative models for captioning and state extraction. Although the extraction is stable under repeated decoding and prompt paraphrasing, future work should examine robustness across alternative ALMs, PSEs, recording conditions, and prompting choices. The current `unknown`-filtering procedure also warrants continued evaluation; its discard rates are reported in Table 13.

Fourth, the state-association analysis is exploratory because multiple utterances from the same participant share one label. The displayed q values do not account for this within-participant clustering and therefore cannot be interpreted as participant-level confirmatory significance. A participant-clustered or participant-level reanalysis is needed to confirm these patterns.

Finally, DAIC-WOZ and MODMA are evaluated independently. The consistent results across these datasets do not constitute a cross-dataset or cross-lingual transfer evaluation. Moreover, the paired comparisons in Table 4 are based on five matched runs or fold-level means and therefore characterize evaluation-unit variation rather than participant-resampling uncertainty. Larger and more diverse cohorts would enable stronger analyses of demographic robustness, generalizability, and real-world screening utility.

Ethics Statement

This study follows the ACL Ethics Policy and uses existing speech datasets for computational mental health research.

Data Access and Original Consent We obtained DAIC-WOZ (Gratch et al., 2014) and MODMA (Cai et al., 2022) through their official distribution channels and used them in accordance with the applicable research-access conditions. The original dataset documentation describes the corresponding ethical-approval and informed-consent procedures. Our study involved no new participant recruitment, participant contact, or prospective data collection.

Privacy and Data Protection Although direct identifiers had been removed by the dataset providers, speech recordings remain sensitive and potentially identifiable biometric data. We therefore treated the recordings and participant-level labels as protected research data. Access was limited to the research workflow, and we made no attempt to identify speakers, recover personal information, or link participants to external records. The ALM and PSE models were run locally; raw audio and participant-level data were not submitted to third-party hosted model services. Generated captions and SC-Para representations were used only for the analyses described in this study.

Human Validation The small-scale validation described in Appendix C.4 was conducted internally by two graduate researchers. No external annotators were recruited and no additional compensation was provided. Annotators evaluated only whether the generated captions and constituent paralinguistic states were consistent with perceived vocal evidence; they were not asked to infer participant identity, life circumstances, mental state, or clinical diagnosis.

Intended Use and Clinical Interpretation ParaCalib is a research framework for studying context-sensitive paralinguistic representations. It is not a medical device and should not be used for diagnosis, triage, treatment decisions, eligibility determinations, or covert monitoring. The prediction targets in this study are binary labels derived from PHQ-8 and PHQ-9 thresholds. These labels indicate questionnaire-defined symptom burden and are not equivalent to independently established clinical diagnoses. Any prospective clinical use would require external validation, appropriate regulatory and ethical review, and qualified human oversight.

Bias and Generalizability Speech-based prediction can be affected by language, accent, dialect, age, gender, culture, recording conditions, and access to care. DAIC-WOZ and MODMA represent limited English- and Mandarin-speaking cohorts and were evaluated independently; the experiments do not establish cross-lingual or cross-population generalizability. ALM-generated descriptions and PSE-derived states may also reflect biases or errors inherited from their pretrained models. Accordingly, the extracted states should be interpreted as model-derived representations rather than clinically observed symptoms. Evaluation across larger and more diverse populations is necessary before

considering broader use.

Acknowledgments

This work was supported by the Lien Foundation, Singapore.

References

Murray Alpert, Enrique R Pouget, and Raul R Silva. 2001. Reflections of depression in acoustic measures of the patient's speech. *Journal of affective disorders*, 66(1):59–69.

American Psychiatric Association. 2022. *Diagnostic and statistical manual of mental disorders*, 5th ed., text rev. edition. American Psychiatric Association.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460.

Andrew Bailey and Mark D Plumbley. 2021. Gender bias in depression detection using audio features. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pages 596–600. IEEE.

Hanshu Cai, Zhenqin Yuan, Yiwen Gao, Shuting Sun, Na Li, Fuze Tian, Han Xiao, Jianxiu Li, Zhengwu Yang, Xiaowei Li, Qinglin Zhao, Zhenyu Liu, Zhijun Yao, Minqiang Yang, Hong Peng, Jing Zhu, Xiaowei Zhang, Guoping Gao, Fang Zheng, and 8 others. 2022. A multi-modal open dataset for mental-disorder analysis. *Scientific Data*, 9(1).

Shreeram Suresh Chandra, Lucas Goncalves, Junchen Lu, Carlos Busso, and Berrak Sisman. 2025. Emotionrankclap: Bridging natural language speaking styles and ordinal speech emotion via rank-n-contrast. *arXiv preprint arXiv:2505.23732*.

Jiuyi Chen, Mingkui Tan, Haifeng Lu, Qiuna Xu, Zhihua Wang, Runhao Zeng, and Xiping Hu. 2026. Towards stable cross-domain depression recognition under missing modalities. *Pattern Recognition*, 177:113367.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022a. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Weidong Chen, Xiaofen Xing, Xiangmin Xu, Jianxin Pang, and Lan Du. 2022b. Speechformer: A hierarchical efficient framework incorporating the characteristics of speech. *arXiv preprint arXiv:2203.03812*.

Karol Chlasta, Krzysztof Wołk, and Izabela Krejtz. 2019. Automated speech-based screening of depression using deep convolutional neural networks. *Procedia Computer Science*, 164:618–628.

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

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. *Preprint*, arXiv:2311.07919.

Jennifer Cole. 2015. Prosody in context: A review. *Language, Cognition and Neuroscience*, 30(1-2):1–31.

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.

S Pavankumar Dubagunta, Bogdan Vlasenko, and Mathew Magimai Doss. 2019. Learning voice source related information for depression detection. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6525–6529. IEEE.

Heiner Ellgring and Klaus R. Scherer. 1996. Vocal indicators of mood change in depression. *Journal of Nonverbal Behavior*, 20(2):83–110.

Florian Eyben, Klaus R. Scherer, Björn W. Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. 2016. The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202.

Isaac R Galatzer-Levy and Jukka-Pekka Onnela. 2023. Machine learning and the digital measurement of psychological health. *Annual Review of Clinical Psychology*, 19:133–154.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024. Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6288–6313.

Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. 2014. The distress analysis interview corpus of

human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3123–3128, Reykjavik, Iceland. European Language Resources Association (ELRA).

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

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.

Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613.

Kurt Kroenke, Tara W Strine, Robert L Spitzer, Janet BW Williams, Joyce T Berry, and Ali H Mokdad. 2009. The phq-8 as a measure of current depression in the general population. *Journal of affective disorders*, 114(1-3):163–173.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626. Association for Computing Machinery.

Yuxin Li, Eng Siong Chng, and Cuntai Guan. 2025a. Hierarchical self-supervised representation learning for depression detection from speech. *arXiv preprint arXiv:2510.08593*.

Yuxin Li, Sinchana Kumbale, Yanru Chen, Tanmay Surana, Eng Siong Chng, and Cuntai Guan. 2025b. Automated depression detection from text and audio: A systematic review. *IEEE Journal of Biomedical and Health Informatics*.

Yuxin Li, Xiangyu Zhang, Yifei Li, Zhiwei Guo, Haoyang Zhang, Eng Siong Chng, and Cuntai Guan. 2026. Depflow: Disentangled speech generation to mitigate semantic bias in depression detection. *arXiv preprint arXiv:2601.00303*.

Xingchen 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.

Michelle Morales, Stefan Scherer, and Rivka Levitan. 2018. A linguistically-informed fusion approach for multimodal depression detection. In *proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic*, pages 13–24.

Muhammad Muzammel, Hanan Salam, Yann Hoffmann, Mohamed Chetouani, and Alice Othmani. 2020. Audvowelconsnet: A phoneme-level based deep cnn architecture for clinical depression diagnosis. *Machine Learning with Applications*, 2:100005.

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 *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part II*, pages 5–19. Springer.

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.

Vijay Ravi, Jinhan Wang, Jonathan Flint, and Abeer Alwan. 2022. A step towards preserving speakers' identity while detecting depression via speaker disentanglement. In *Interspeech*, volume 2022, page 3338.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3982–3992. Association for Computational Linguistics.

Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, Siyang Song, Shuo Liu, Ziping Zhao, Adria Mallol-Ragolta, Zhao Ren, Mohammad Soleymani, and Maja Pantic. 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*, AVEC '19, page 3–12, New York, NY, USA. Association for Computing Machinery.

Afef Saidi, Slim Ben Othman, and Slim Ben Saoud. 2020. Hybrid cnn-svm classifier for efficient depression detection system. In *2020 4th International Conference on Advanced Systems and Emergent Technologies (IC_ASET)*, pages 229–234. IEEE.

Asif Salekin, Jeremy W Eberle, Jeffrey J Glenn, Bethany A Teachman, and John A Stankovic. 2018. A weakly supervised learning framework for detecting social anxiety and depression. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 2(2):1–26.

David Snyder, Guoguo Chen, and Daniel Povey. 2015. MUSAN: A music, speech, and noise corpus.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. Salmonn: Towards generic hearing abilities for large language models. In *International Conference on Learning Representations*, volume 2024, pages 16607–16629.

Fei Tian, Xiangyu Tony Zhang, Yuxin Zhang, Haoyang Zhang, Yuxin Li, Daijiao Liu, Yayue Deng, Donghang Wu, Jun Chen, Liang Zhao, Chengyuan Yao, Hexin Liu, Eng Siong Chng, Xuerui Yang, Xiangyu Zhang, Daxin Jiang, and Gang Yu. 2025. Step-audio-r1 technical report. *Preprint*, arXiv:2511.15848.

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.

Michel Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic. 2013. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 3–10.

Adrián Vázquez-Romero and Ascensión Gallardo-Antolín. 2020. Automatic detection of depression in speech using ensemble convolutional neural networks. *Entropy*, 22(6):688.

Michael Wagner and Duane G. Watson. 2010. Experimental and theoretical advances in prosody: A review. *Language and Cognitive Processes*, 25(7-9):905–945.

Dong Wang, Yanhui Ding, Qing Zhao, Peilin Yang, Shuping Tan, and Ya Li. 2022. Ecapa-tdnn based depression detection from clinical speech. In *Interspeech*, pages 3333–3337.

Jinhan Wang, Vijay Ravi, and Abeer Alwan. 2023. Non-uniform speaker disentanglement for depression detection from raw speech signals. In *Interspeech*, volume 2023, page 2343.

World Health Organization. 2025. Depressive disorder (depression). WHO Fact Sheet, accessed 2026.

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, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025. Qwen3-omni technical report. *Preprint*, arXiv:2509.17765.

Yaoxun Xu, Hangting Chen, Jianwei Yu, Qiaochu Huang, Zhiyong Wu, Shi-Xiong Zhang, Guangzhi Li, Yi Luo, and Rongzhi Gu. 2024. Secap: Speech emotion captioning with large language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19323–19331.

Masahiro Yamamoto, Akihiro Takamiya, Kazutaka Sawada, Masaru Yoshimura, Masahiro Kitazawa, Kuang Liang, Takashi Fujita, Masaru Mimura, and Taishiro Kishimoto. 2020. Using speech recognition technology to investigate the association between timing-related speech features and depression severity. *PLOS ONE*, 15(9):e0238726.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Pingyue Zhang, Mengyue Wu, Heinrich Dinkel, and Kai Yu. 2021. Depa: Self-supervised audio embedding for depression detection. In *Proceedings of the 29th ACM international conference on multimedia*, pages 135–143.

Wenyu Zhang, Yingxu He, Geyu Lin, Zhuohan Liu, Shuo Sun, Bin Wang, Xunlong Zou, Jeremy H. M. Wong, Qiongqiong Wang, Hardik Bhupendra Sailor, Nancy F. Chen, and AiTi Aw. 2025a. Beyond classification: Towards speech emotion reasoning with multitask AudioLLMs. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1132–1148, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Xiangyu Zhang, Hexin Liu, Kaishuai Xu, Qiquan Zhang, Daijiao Liu, Beena Ahmed, and Julien Epps. 2024. 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.

Xiangyu Zhang, Hexin Liu, Qiquan Zhang, Beena Ahmed, and Julien Epps. 2025b. Speecht-rag: Reliable depression detection in llms with retrieval-augmented generation using speech timing information. *arXiv preprint arXiv:2502.10950*.

Xiangyu Zhao, Yaling Shen, Yiwen Jiang, Zimu Wang, Jiahe Liu, Maxmartwell H Cheng, Guilherme C

Oliveira, Robert Desimone, Dominic Dwyer, and Zongyuan Ge. 2025. It hears, it sees too: Multimodal llm for depression detection by integrating visual understanding into audio language models. *arXiv preprint arXiv:2511.19877*.

Ziping Zhao, Zhongtian Bao, Zixing Zhang, Nicholas Cummins, Haishuai Wang, and Björn Schuller. 2020. Hierarchical attention transfer networks for depression assessment from speech. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 7159–7163. IEEE.

A Experimental Setup & Implementation Details

A.1 Dataset Details and Preprocessing

We correct two known annotation inconsistencies. For DAIC-WOZ, we relabel participant 409 as depressed following Bailey et al. (Bailey and Plumbley, 2021). For MODMA, we use PHQ-based binary labels rather than the original clinician-defined grouping and relabel participant 2010037 as non-depression based on the PHQ score, following Li et al. (Li et al., 2025a).

For DAIC-WOZ, raw audio is segmented into participant utterances using transcript timestamps. We correct transcript–audio misalignment by shifting timestamps by 34.000 s for participant 318, 3.355 s for participant 321, 6.070 s for participant 341, and 16.540 s for participant 362. MODMA is distributed with sentence-level segmentation.

A.2 Training Details, Architectures, and Splits

All compared classifiers are trained under the same optimization, batching, device, and early-stopping settings on a single NVIDIA RTX 6000 GPU.

Architecture Details: The MLP baseline applies a two-hidden-layer classifier (∼2.7K params). The LSTM applies a two-layer LSTM (∼58K). The Bi-LSTM uses two bidirectional layers (∼147K). The Transformer uses two encoder layers and mean pooling (∼106K). Attentive MIL uses attention pooling followed by the same MLP (∼3.3K parameters in total). A full performance breakdown is shown in Table 5.

Significance and Split-wise Performance: We compare macro-F1 scores using a two-sided paired t-test across the five matched runs described in Section 4 (Table 4). These tests characterize matched run- or fold-level variation and are not participant-resampling analyses. Table 6 reports the training, development, and test performance under the evaluated splits.

Category	Hyperparameter & Value
Input	Unit: SC-Para vector (Dim 7) Grouping: Valid vectors per participant
Attention	Projection: Linear, $7 \rightarrow 64$, Tanh Query: Learned 64-d vector Weights: Softmax over utterances Readout: Weighted sum of 7-d vectors
Classifier	Architecture: 2-layer MLP (64, 32) Norm: BatchNorm (after 1st layer) Act./Dropout: ReLU / 0.3
Optim.	Optimizer: AdamW (5×10^{-4} / 1×10^{-4}) Pos. Weight: $N_{\text{neg}}/N_{\text{pos}}$, Clip: 1.0 LR scheduler: ReduceLROnPlateau
Training	Max epochs/batch: 50 / 4 Early stopping patience: 10

Table 3: Implementation and training configuration.

Comparison	Δ Macro-F1	95% CI	p_{adj}
SC-Para vs. HAREN-CTC			
DAIC-WOZ	14.03	[6.76, 21.30]	.0232
MODMA	17.46	[3.68, 31.24]	.0734
SC-Para vs. Caption Embedding			
DAIC-WOZ	5.37	[-9.18, 19.92]	.7269
MODMA	4.80	[-15.23, 24.84]	.7269

Table 4: Paired Macro-F1 comparisons across the five matched evaluation units. Confidence intervals are unadjusted 95% paired-difference intervals; p-values are adjusted using the Holm procedure across the four comparisons.

A.3 Model and Reproducibility Settings

We use Qwen3-Omni-30B-A3B-Captioner (Xu et al., 2025) and DeepSeek-R1-Distill-Qwen-32B (Guo et al., 2025) locally with an OpenAI-compatible vLLM backend (Kwon et al., 2023) in `bfloat16` precision. Generation uses temperature $= 0$, top-$p = 1.0$. Table 3 records the downstream optimizer, learning rates, batch size, scheduler, and early-stopping settings; Section 4 specifies the seed and fold aggregation used for evaluation.

The continuous LLD baseline uses 88 utterance-level eGeMAPSv02 functionals (Eyben et al., 2016). The discretized baseline maps five functionals to acoustic proxies (Table 7). For SSL representations, we use the hidden states from the last Transformer layer of Wav2Vec2-large-960h, HuBERT-large-ls960-ft, and WavLM-large. Transcript rep-

Dataset	Classifier	Macro-F1	Sensitivity	Specificity	AUC
DAIC-WOZ	Bi-LSTM	49.6 (3.3)	40.0 (13.0)	61.2 (10.4)	56.8 (3.8)
	LSTM	42.2 (5.0)	18.6 (14.8)	70.9 (24.5)	45.7 (3.7)
	Transformer w/ Pos. Enc.	40.2 (10.2)	22.9 (43.3)	**79.4** (44.4)	55.8 (3.6)
	Transformer w/o Pos. Enc.	60.6 (5.5)	**85.7** (8.7)	50.9 (11.0)	73.2 (3.7)
	MLP	65.5 (5.0)	58.6 (9.3)	74.5 (2.7)	73.7 (4.2)
	Attentive MIL	**71.9** (4.9)	77.1 (15.5)	73.3 (6.6)	**83.3** (2.6)
MODMA	LSTM	79.3 (7.5)	74.0 (9.4)	85.3 (14.1)	94.2 (4.2)
	Transformer w/ Pos. Enc.	77.6 (9.1)	71.4 (9.4)	86.4 (16.7)	94.4 (7.1)
	Transformer w/o Pos. Enc.	81.3 (8.1)	77.0 (7.3)	86.0 (14.8)	95.7 (6.0)
	Bi-LSTM	82.8 (7.5)	80.6 (6.8)	85.5 (11.7)	94.2 (8.1)
	MLP	84.7 (13.1)	90.0 (10.3)	81.9 (18.8)	90.5 (10.9)
	Attentive MIL	**90.5** (11.2)	**91.0** (12.5)	**90.0** (14.9)	**96.4** (6.9)

Table 5: Classifier comparison using the same utterance-level SC-Para vectors.

Dataset	Split	Accuracy	Macro-F1	Sensitivity	Specificity	AUC
DAIC-WOZ	Train	74.9 (3.2)	70.1 (4.2)	61.3 (8.7)	80.3 (2.6)	78.0 (6.7)
	Dev	74.9 (6.5)	71.0 (7.6)	56.7 (12.4)	84.4 (5.8)	83.0 (6.9)
	Test	74.5 (4.0)	71.9 (4.9)	77.1 (15.5)	73.3 (6.6)	83.3 (2.6)
MODMA	Train	92.9 (1.9)	92.8 (1.9)	91.5 (3.2)	94.0 (3.1)	98.5 (0.7)
	Dev	91.4 (10.5)	91.1 (10.6)	88.4 (10.5)	93.3 (13.1)	95.8 (7.2)
	Test	90.7 (11.1)	90.5 (11.2)	91.0 (12.5)	90.0 (14.9)	96.4 (6.9)

Table 6: Training, development, and test performance of ParaCalib across splits.

resentations encode utterances with a frozen all-MiniLM-L6-v2 (Reimers and Gurevych, 2019). For Caption Embedding, we encode each complete ALM-generated caption using the same frozen all-MiniLM-L6-v2 encoder and use the resulting 384-dimensional mean-pooled sentence embedding as the utterance-level MIL instance.

B Additional Ablation and Sensitivity Analyses

B.1 Correlations among Categorical SC-Para Dimensions

Table 8 reports the pairwise Spearman correlations among the five categorical SC-Para dimensions on DAIC-WOZ, and Table 9 reports their extraction stability under repeated decoding. Correlation magnitudes reach up to .70, indicating that the dimensions are related but not redundant.

B.2 PSE Capacity and Stability

Replacing DeepSeek-32B with the smaller Qwen3-8B (Yang et al., 2025) achieves 68.2 (6.3)% Macro-F1 on DAIC-WOZ, remaining above HAREN-CTC at 57.9 (2.1)% and the Caption Embedding mean of 66.6 (14.2)%.

Repeated decoding shows high exact agreement across all SC-Para dimensions (Table 9). Sensitivity to meaning-preserving prompt paraphrases yields a mean evidence-score difference of 2.30 points, with correlations ranging from .841 to .952.

B.3 Methodological Details for Statistical and Controlled Analyses

Semantic-Context Control: Because changing lexical content would also change the waveform, while speech resynthesis could introduce acoustic artifacts, we isolate semantic calibration at the PSE stage. The analysis uses 100 utterances from 66 participants. For each utterance, we hold the ALM-generated caption-level acoustic description fixed and compare three prompt conditions: acoustic-only, non-depression-related context, and depression-related context.

Depression-related context increases the PSE-derived evidence score by 8.6 points relative to non-depression-related context. The utterance-paired comparison yields $t = 8.64$, $p < .001$, and $d_z = .86$; the corresponding participant-clustered 95% confidence interval is $[6.7, 10.6]$. Depression-related context also increases the score by 8.7 points relative to acoustic-only input, whereas non-depression-related context differs from acoustic-only input by only 0.1 points ($p = .92$). This experiment characterizes semantic-context sensitivity at the PSE stage while keeping the caption-level acoustic description unchanged.

SC-Para proxy	eGeMAPSv02 functional	Discretization
Pitch var.	`F0semitoneFrom27.5Hz_sma3nz_stddevNorm`	Tertiles
Speak rate	`VoicedSegmentsPerSec`	Tertiles
Rhythm	`StddevVoicedSegmentLengthSec`	Median
Pauses	`MeanUnvoicedSegmentLength`	Tertiles
Energy	`loudness_sma3_amean`	Tertiles

Table 7: Mapping used by the discretized LLD baseline.

	Pitch	Rate	Rhythm	Pauses	Energy
Pitch	1.000	.586	-.236	-.265	.674
Rate	.586	1.000	-.336	-.400	.703
Rhythm	-.236	-.336	1.000	.589	-.296
Pauses	-.265	-.400	.589	1.000	-.349
Energy	.674	.703	-.296	-.349	1.000

Table 8: Pairwise Spearman correlations among the five categorical SC-Para dimensions on DAIC-WOZ.

SC-Para dimension	Exact agreement
Pitch variability	99 / 100 (99%)
Speaking rate	99 / 100 (99%)
Rhythm regularity	99 / 100 (99%)
Pause pattern	98 / 100 (98%)
Vocal energy	98 / 100 (98%)
Depression evidence	94 / 100 (94%)
Extraction confidence	95 / 100 (95%)

Table 9: Exact agreement of PSE outputs under repeated decoding for 100 randomly sampled ALM-generated captions.

State Associations and Benjamini–Hochberg Correction: Figure 4 displays the ten highest-ranked state pairs for descriptive comparison. Seven pairs satisfy both prespecified screening criteria: $n \geq 100$ utterances and a Benjamini–Hochberg-adjusted value of $q < .05$. The three lower-support configurations ($n = 55$, 77, and 92) are shown only for context and are not interpreted. Because multiple utterances from each participant share one depression label, these utterance-level q values do not account for within-participant clustering and must not be interpreted as participant-level confirmatory significance.

C Robustness, Bias, and Error Analysis

C.1 Speaker-Identity and Additive Noise

A speaker-verification probe on DAIC-WOZ (Table 10) yields ROC-AUC values close to chance for SC-Para (52.8–52.9%), whereas WavLM reaches 78.7%.

Mixing background noise from MUSAN (Snyder et al., 2015) at 20, 10, and 5 dB SNR decreases Macro-F1 from 71.9% to 68.6%, 63.9%, and 62.6%, respectively.

Representation	Macro-F1	AUC
SC-Para (5 states)	51.9 (0.8)	52.8 (0.9)
SC-Para (7 dimensions)	52.0 (0.8)	52.9 (0.9)
WavLM (1024 dims)	71.7 (0.3)	78.7 (0.4)

Table 10: Speaker-verification probe on DAIC-WOZ.

DAIC-WOZ			MODMA		
Severity	N	Acc (%)	Severity	N	Acc (%)
Minimal	22	82.7	Minimal	26	92.3
Mild	11	54.5	Mild	4	100.0
Mod.	5	80.0	Mod.	3	93.3
Mod-Sev.	7	68.6	Mod-Sev.	10	80.0
Severe	2	100.0	Severe	9	95.6

Table 11: Post-hoc prediction accuracy across PHQ severity strata for DAIC-WOZ official test participants and the complete MODMA out-of-fold cohort. Values are mean accuracies across the same five random seeds used in the main evaluation. N denotes the number of unique participants in each severity stratum.

Stage	Time/utt (s)	Total GPU (h)	Peak Mem (GiB)
Captioner	1.64	10.5	87.5
PSE	1.38	8.8	87.1
ParaCalib (Total)	3.0	19.3	87.5
SSL baselines	0.019–0.027	0.12–0.18	1.4–1.5

Table 12: Representation-extraction cost on DAIC-WOZ.

C.2 Error Analysis and Computational Cost

Consistent false positives lie closer to the PHQ-8 cutoff than reliable controls (mean PHQ-8 of 6.4 vs 2.6). Consistent false negatives contain fewer valid utterances on average than reliably classified true positives (66 vs. 119). Table 11 reports a post-hoc severity-stratified analysis. Table 12 summarizes computational costs. Table 13 reports discard rates due to `unknown` assignments, which do not substantially alter class composition.

Dataset	Split	Overall	Utterance discard rate		Per-participant discard rate	
			Non-dep.	Dep.	Non-dep.	Dep.
DAIC-WOZ	Train	5.7%	6.0%	4.9%	6.1 (3.8)%	4.4 (3.5)%
	Dev	4.7%	4.6%	4.8%	5.0 (3.0)%	4.5 (2.2)%
	Test	6.2%	6.2%	6.2%	6.0 (3.7)%	7.0 (3.0)%
MODMA	Train	1.9%	1.4%	2.6%	1.4 (3.9)%	2.6 (4.7)%
	Dev	2.5%	2.8%	2.2%	2.8 (2.8)%	2.3 (2.8)%
	Test	0.6%	0.0%	1.1%	0.0 (0.0)%	1.1 (2.2)%

Table 13: Discard rates after PSE filtering by split and label group.

Validation Target	Consistency Metrics			Error Counts	
	Pos. Judgments	Strict Pass	Rater Agree.	Inconsist. by ≥ 1	Inconsist. by both
Caption consistency	38/40 (95%)	18/20 (90%)	18/20 (90%)	2/20	0/20
PSE constituent states	36/40 (90%)	17/20 (85%)	18/20 (90%)	3/20	1/20

Table 14: Human validation of generated captions and SC-Para states on 20 sampled utterances.

C.3 Attention-Weighted Dispersion Measure

To assess the internal consistency of participant-level representations, we computed an attention-weighted dispersion measure across valid utterance-level SC-Para vectors. Lower values indicate that the utterances receiving higher attention weights have more similar categorical SC-Para patterns within a participant. Group differences were evaluated using a subject-level permutation test, with a bootstrap confidence interval for the mean difference, a Mann–Whitney U test, and Cliff's delta as an effect-size estimate.

In the DAIC-WOZ test split, depression-label participants had lower mean attention-weighted dispersion than non-depression-label participants (0.3617 vs. 0.3945; $n = 14$ and $n = 33$, respectively). The mean difference was -0.0327 (bootstrap 95% CI $[-0.0593, -0.0081]$; subject-level permutation $p = .0189$). The result was consistent with the Mann–Whitney U test ($p = .0467$), with Cliff's $\delta = -0.3723$. Given the small test-set groups, this analysis is considered exploratory.

C.4 Human Validation

Two annotators (graduate researchers proficient in the evaluated languages) evaluated 20 randomly sampled utterances. Since this small-scale validation was conducted internally as part of routine research activities, no external recruitment or additional financial compensation was involved. The generated captions and the five PSE-derived constituent states were judged consistent with perceived vocal evidence in 18/20 and 17/20 samples, respectively (Table 14).

D Prompts and Output Schema

Table 15 presents the content-masking prompt, an abbreviated SC-Para feature schema, and the PSE extraction template used in ParaCalib.

Prompt 1: Content Masking (Lexical Cue Removal)

```
You are an audio description filtering expert. Your task is to filter out speech content
↪  transcription from audio descriptions and keep only the audio feature descriptions.
Content to REMOVE:
1. Specific words and phrases spoken by the speaker (direct quotes)
2. Detailed descriptions of conversation content
3. Explanations of semantic meaning and topics discussed

Content to KEEP:
1. Speaker's voice characteristics (timbre, tone, pitch, voice quality)
2. Speaking patterns (pace, rhythm, pauses, cadence)
...
Return ONLY the filtered description without any explanations or additional comments. If
↪  the entire description is transcription content, return an empty string.
```

Prompt 2: SC-Para Feature Definitions & Schema

```
Feature definitions:
prosody.pitch_variability:
   - "flat": reduced pitch variation, monotonous or restricted intonation.
   - "moderate": normal or balanced pitch variation after semantic calibration.
   - "high": large pitch variation that cannot be explained only by transcript context.
   - "unknown": insufficient evidence.
... [Other definitions omitted for brevity: speaking_rate, rhythm, pauses, energy] ...

depression_related_paralinguistic_evidence.value:
   - A calibrated depression-related paralinguistic evidence score from 0 to 100.
depression_related_paralinguistic_evidence.confidence:
   - A value from 0.0 to 1.0. Higher confidence means the evidence is clear.
```

Prompt 3: PSE Extraction Template

```
You are an expert in paralinguistic analysis of depression-related speech patterns.
IMPORTANT CONSTRAINTS:
- You CANNOT infer identity, mental diagnosis, or life situation.
- You are NOT making a medical diagnosis.
- Your task is to estimate the strength of depression-related paralinguistic evidence at
↪  the utterance level ONLY.

Input:
A caption generated by an audio-language model for a single utterance.

Your task:
1. Extract acoustically observable vocal and paralinguistic information.
2. Based on vocal and paralinguistic patterns, estimate the presence and strength of
↪  depression-related speech characteristics.
3. If vocal information is sparse or ambiguous, use "unknown" and reduce confidence.

Output STRICTLY in the following JSON format. Each JSON corresponds to ONE utterance.
{
  "prosody": {
    "pitch_variability": "flat/moderate/high/unknown",
    ...
  },
  "depression_related_paralinguistic_evidence": {
    "value": 0-100,
    "confidence": 0.0-1.0
  }
}
```

Table 15: Selected prompts and an abbreviated output schema used for content masking and PSE extraction.