MAPS: Robot-Mediated Multimodal Mental Health Screening for Chinese Adolescents

Zhicheng Du[1], Yang Liu[1], Zheng Chen[1], Zhuo Deng[1], Zhaotian Xie[5], Wenji Xi[6], Xi Yuan[2], Zhen Liang[7], Dongmei Yu[2], Chenggang Yan[8], Jiansong Ji[4], Peiwu Qin[2,3,4,*], Zhenglin Chen[2,3,*], Lan Ma[1,*]

[1]Institute of Biopharmaceutics and Health Engineering, Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, Guangdong 518055, China
[2]Chinese Medicine Guangdong Laboratory/Hengqin Laboratory, Hengqin, Guangdong 519031, China
[3]Tianfu Institute of Research and Innovation, Southwest University of Science and Technology, Chengdu 610000, China
[4]Key Laboratory of Imaging Diagnosis and Minimally Invasive Intervention Research, The Fifth Affiliated Hospital of Wenzhou Medical University, Lishui 323000, Zhejiang, China
[5]Institute of Materials Research, Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China
[6]Faculty of Engineering, The University of Sheffield, Sheffield S1 3JD, United Kingdom
[7]School of Biomedical Engineering, Medical School, Shenzhen University, Shenzhen, Guangdong 518055, China
[8]School of Communication Engineering, Hangzhou Dianzi University, Hangzhou, Zhejiang 310018, China
[*]Corresponding authors: pwqin@sz.tsinghua.edu.cn, chenzlin1992@163.com, malan@sz.tsinghua.edu.cn

Abstract

We present an initial study of robot-mediated multimodal mental-health screening for Chinese adolescents aged 12–15. MAPS (Multimodal Adolescent Psychological Screening) combines facial, acoustic, textual, and remotely estimated physiological signals collected through a standardized protocol of questionnaires, emotionally evocative tasks, and question answering. From 3,783 participants, 968 records were retained after consent, completeness, and quality-control filtering. We introduce GAME (Generalized Model with Attention and Multimodal EmbraceNet), a fusion model combining eight unimodal representations with relation-graph and attention features. Separate models are trained for 12 questionnaire- and expert-derived screening targets. Stratified 10-fold cross-validation yields 73.04–92.77% accuracy and 71.32–91.06% weighted F1. Feature ablations and descriptive contribution analyses suggest that representation use varies across targets. This initial study connects a portable interaction protocol with task-specific multimodal modeling in a Chinese school setting and provides a basis for further research on screening-label prediction. Future work will expand data collection, refine the fusion and training procedures, and evaluate generalization and screening utility through external and clinically supervised validation.

1 Introduction

Adolescent mental-health problems often emerge during a period in which communication with adults and access to professional support are constrained, and mental disorders are associated with later suicidal behavior (Gili et al., 2019). Screening instruments and interviews remain important, but they can be affected by recall, stigma, response style, and the limited observability of short encounters. Multimodal interaction offers a complementary perspective: speech prosody, facial behavior, language, and physiological responses can be recorded while a participant engages with the same set of prompts and stimuli. A standardized interaction protocol places these signals in a common elicitation context, enabling analysis of how their predictive contributions differ across screening targets.

Prior work has used language, speech, wearable sensors, and digital phenotyping for mental-health modeling, but many systems focus on a single condition or modality (Moura et al., 2023; Dwyer and Koutsouleris, 2022; Cohen et al., 2020). Multimodal fusion can improve robustness, yet missing or noisy channels, class imbalance, and opaque feature aggregation remain practical obstacles (Huang et al., 2021; Ma et al., 2021). Interactive robots have also been studied as acceptable interfaces for youth mental-health interaction, but their use as a multimodal screening platform is less explored (Rasouli et al., 2022; Abbasi et al., 2022; Li et al., 2023).

We address this problem through an integrated collection and modeling workflow. Our contributions are: (1) MAPS, a Chinese adolescent screening collection with four signal modalities and 12

Proceedings of the 1st Workshop on Multimodal, Multilingual, and Multicultural Mental Health and Psychotherapy (MultiPsyche), pages 69–79
November 9, 2026 ©2026 Association for Computational Linguistics

screening targets; (2) GAME, a fusion model trained independently for each target, combining eight unimodal representations with cross-modal relation and attention features; and (3) a stratified 10-fold evaluation with feature-wise classical comparisons, feature ablations, and target-specific contribution analysis. Together, these components connect the interaction protocol to the evidence used by each screening model.

2 Related Work

Clinical NLP and mental-health assessment. Clinical NLP highlights the importance of label validity and supporting evidence. Expert review can reveal false positives in depression annotations (Milintsevich et al., 2024), while the CLPsych 2024 shared task evaluates evidence supporting suicidality labels (Chim et al., 2024). Recent work also advocates validated, dimensional, and transdiagnostic measures (Shani and Stade, 2025). These studies motivate our distinction between predicting screening labels and establishing clinical diagnoses.

Multimodal modeling. Mental-health research draws on language, speech, facial behavior, and physiological measurements, including non-contact sensing such as PhysioCam (Davila et al., 2017). Wearable approaches include ultrathin ionic electrodes for electrophysiological recording and depression-related EEG analysis (Zhang et al., 2025); our protocol instead derives physiological signals from facial video. Their predictive value can depend on the elicitation context: Mandal et al. (2025) model question-wise modality contributions for depression estimation, and TelME uses cross-modal supervision for emotion recognition (Yun et al., 2024). Our setting combines four signal modalities within a standardized robot interaction protocol. A common fusion architecture is trained separately for each screening target, allowing feature contributions to vary across outcomes.

Cultural and linguistic validity. Mental-health expressions depend on language, school context, and local understandings of distress. Havaldar et al. (2024) model cultural variation with knowledge-guided lexica and document limitations in language models' handling of culturally situated language. Our Chinese school-based collection provides a specific linguistic and cultural setting for multimodal research. Its measurement instrument and recruitment context define the scope within which the reported results should be interpreted.

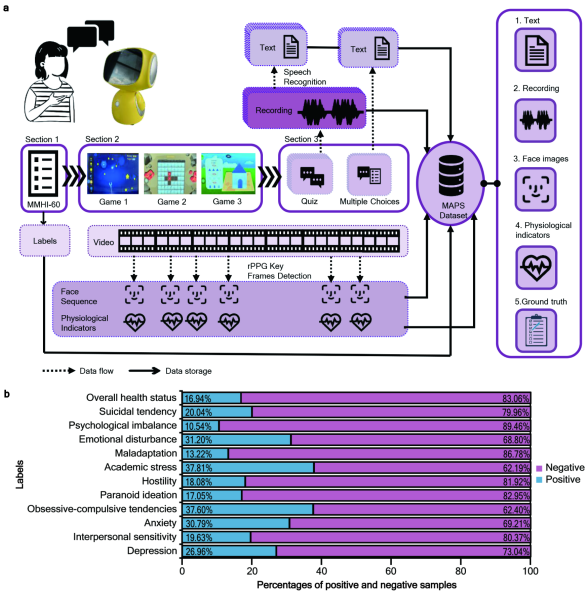


Figure 1: Robot-mediated acquisition and label distribution. The interaction combines questionnaire screening, interactive tasks, and question answering; the resulting collection contains four signal modalities and imbalanced screening labels.

3 System and Dataset

3.1 Robot-mediated interaction

The platform is a portable robot with a touch screen, camera, speaker, and microphone. An Android application presents three consecutive stages: questionnaire-based screening, emotionally evocative tasks, and question answering. The interaction design draws on hyper-emotion theory (Johnson-Laird et al., 2006), which motivates examining responses to shared emotional stimuli. The interface was designed with input from psychologists, school counselors, and parents. Video and audio are recorded during interaction, while the backend extracts features for screening-task prediction rather than diagnosis. The design aims to reduce the formality of data collection in schools.

3.2 MAPS collection

The collection involved 3,783 Chinese middle-school students aged 12–15 recruited across multiple schools in one province of China. The analyzed MAPS subset contains 968 records retained after consent, completeness, and quality-control

filtering. Each record is associated with labels derived from the Mental Health Inventory of Middle School Students (MMHI-60; Wang et al., 1997) and additional expert assessment for suicidal tendency. The 12 binary screening targets are depression, interpersonal sensitivity, anxiety, obsessive-compulsive tendencies, paranoid ideation, hostility, academic stress, maladaptation, emotional disturbance, psychological imbalance, suicidal tendency, and overall mental-health status.

The collection is culturally grounded in a Chinese school context and uses a Chinese adolescent instrument rather than translating a clinical measure developed for another population. This makes it useful for studying cultural and linguistic validity, but it also limits direct transfer to other languages, age groups, and health systems. As shown in Figure 1, positive-label prevalence ranges from 10.54% (psychological imbalance) to 37.81% (academic stress). Positive cases are the minority for every target, motivating stratified evaluation and class-aware interpretation.

3.3 Modalities and preprocessing

The four modalities are face video, audio, text transcripts, and remote physiological signals. Face processing uses facial landmarks and iris tracking; audio is represented by MFCC and wav2vec 2.0 embeddings (Baevski et al., 2020); transcripts are represented by Chinese RoBERTa (Liu et al., 2019) and PERT (Cui et al., 2022), using BERT-family encoders (Devlin et al., 2019); and physiological time series are summarized using time-series features (Christ et al., 2018). Audio recordings are standardized to a 10-second window through truncation or zero padding. All extracted features are z-score normalized before fusion. Open responses are transcribed with the Iflyrec speech-recognition service. The three multiple-choice responses are represented by their full selected wording, yielding eight textual responses alongside five audio responses per participant.

3.4 Interaction protocol and quality control

The protocol was designed to elicit complementary responses without requiring wearable sensors. In the first stage, participants complete MMHI-60 in a classroom setting. In the second stage, they interact with three interactive tasks: Gold Miner, Minesweeper, and a construction task combining elements of the House–Tree–Person and Lowenfeld Mosaic paradigms. The first two tasks expose participants to reward and loss events, while the construction task records choices and visual behavior. In the third stage, the robot presents five open questions and three situational multiple-choice questions (Table 1). These prompts were selected with psychologists and school counselors to cover everyday adolescent experiences without presenting a diagnostic interview.

	Prompt or scenario
O1	Briefly introduce your interests and hobbies.
O2	Who is your favorite idol? Why?
O3	Who do you admire most? Why?
O4	Share one of the happiest things that happened recently.
O5	Share something that recently bothered you or made you unhappy.
S1	No classmates make friends with you: remain alone or initiate contact?
S2	Effort brings poor grades and negative comments: give up or continue working?
S3	Your parents argue every day: remain silent or ask them to stop?

Table 1: Question inventory, condensed from the documented English protocol. Open responses (O) yield audio and text; situational choices (S) are encoded as the full selected wording.

Of 3,783 participants, 891 records were excluded for missing guardian consent, 426 for unavailable facial data, 254 for missing physiological signals, 391 for missing voice recordings, 299 for missing transcripts, and 554 for insufficient quality. The remaining 968 records (25.6% of the initial cohort) form a complete-case subset. Consent, availability, and quality filtering may select participants with different characteristics or recording conditions. The results therefore describe this retained subset, not all 3,783 participants. Facial video is recorded at 30 frames per second during the interactive tasks and question answering. The backend estimates physiological signals from this video using remote photoplethysmography (rPPG), retaining corresponding facial frames and physiological estimates when extraction succeeds. Thus, facial and physiological representations share a recording source; a video-quality failure can affect both.

4 GAME: Task-Specific Multimodal Fusion

4.1 Unimodal representations

For each participant, the fusion model uses eight unimodal representations: expression, expres-

sion nuance, eye movement, physiological signs, MFCC, wav2vec, PERT, and RoBERTa. These representations preserve complementary information at different temporal and semantic scales. The text checkpoints are Chinese-roberta-wwm-ext-large and Chinese-pert-large. Features from eight textual responses and five audio responses are averaged within each participant. We use fixed pretrained feature extractors where applicable and train the downstream fusion and prediction layers separately for each target.

The extracted dimensions before projection are heterogeneous. MFCC is represented as a 13×999 matrix flattened to 12,987 dimensions and wav2vec 2.0 produces a 1,024-dimensional vector. Chinese RoBERTa and PERT each produce 1,024-dimensional representations for the responses. Face landmarks form an $n \times 3 \times 468$ sequence. Averaging landmarks yields a 1,404-dimensional expression vector; averaging differences between successive landmark frames yields a 1,404-dimensional expression-nuance vector. Iris coordinates are zero-padded to 694×4 and flattened to a 2,776-dimensional eye-movement vector. Tsfresh produces a 4,722-dimensional physiological vector. The 8×8 relation matrix is flattened to 64 dimensions; the attention representation has 26,365 dimensions. A multilayer perceptron projects every representation to the common embracement size of 32. The projection makes these different input sizes compatible with the embracement layer.

4.2 Cross-modal relation and attention

The first cross-modal feature is a relation graph. Each unimodal representation is a node, and edges encode pairwise DTW distances. The extracted feature vectors are treated as sequences, and dynamic time warping (DTW) is used to compare their unequal lengths. The second feature is an attention representation. For each unimodal feature in turn, distances to the remaining features are converted to normalized weights, and weighted aligned representations are combined with the reference feature. Concatenating the eight reference-specific outputs yields an additional cross-modal representation.

For two sequences X and Y, we use a squared point-wise distance and dynamic programming to compute the alignment cost:

$$D(i,j) = \|x_i - y_j\|_2^2 + \min_{\ell \in \triangle} D(\ell), \qquad (1)$$

where $\triangle = \{(i-1,j), (i,j-1), (i-1,j-1)\}$ contains the three admissible predecessors. The final DTW distance is $D(m,n)$, with $D(0,0) = 0$ and infinite costs on the remaining zero-index boundaries. For a reference feature, a softmax transformation of its pairwise DTW distances produces normalized weights for the other features. The weighted aligned sequence is added to the reference representation. Repeating this operation for all eight references produces eight aligned views, which are concatenated into the attention feature. This is a heuristic over ordered feature coordinates, not temporal alignment of the original signals. Coordinates from different encoders do not share a semantic axis, and the cost depends on coordinate order and sequence length; z-score normalization alone does not remove these differences. The resulting weights should therefore not be interpreted as calibrated cross-modal similarity. Joint-representation metrics such as MAJORScore (Du et al., 2025) offer an alternative approach to assessing relevance across three or more modalities; their applicability to these heterogeneous screening features remains to be tested.

4.3 EmbraceNet fusion and target-specific prediction

The eight unimodal features and two cross-modal features are projected into a common docking space and fused with EmbraceNet (Choi and Lee, 2019). We train a separate model for each of the 12 binary screening targets. For task t,

$$\hat{y}_t = g_t\left(\mathrm{Embrace}_t\left(\{d_{k,t}\}_{k=1}^{10}\right)\right), \qquad (2)$$

where $d_{k,t}$ denotes a representation docked by the model for target t. The embracement size is 32. The architecture and feature-extraction procedure are reused across targets, but fusion and prediction parameters are learned independently. Target-specific feature use arises through this independent supervised training.

EmbraceNet uses a docking layer for each input and a probability-based embracement layer. For input $x^{(k)}$, the kth docking output is $d^{(k)} = f_a(W^{(k)}x^{(k)} + b^{(k)})$. For each embracement coordinate, a one-hot sample from a multinomial distribution selects one docked representation, and the selected vectors are summed. This stochastic selection provides an interface for future modality dropout. In the present experiments, however, evaluation is performed on the complete-case sub-

set; robustness to systematically missing modalities is therefore an open empirical question rather than an established property of this study.

5 Experimental Setup

5.1 Evaluation protocol

We use stratified 10-fold cross-validation, preserving the class ratio for each target. Accuracy, weighted precision, weighted recall, and weighted F1 are computed for each fold and then averaged. The use of stratification is important because several targets are substantially imbalanced. The reported results are screening-task performance against questionnaire- and expert-derived labels; they do not establish psychiatric diagnosis, treatment benefit, or clinical deployment readiness.

We compare with SVM-RBF, random forest, and gradient boosting; SVM-Polynomial was also evaluated but is omitted from the compact table. Classical results are averages across separate feature-wise experiments using unimodal and cross-modal representations, whereas the fusion model combines all ten representations. These are descriptive comparisons, not matched all-feature baselines, and they cannot isolate the effect of the proposed fusion architecture. Matched concatenation and EmbraceNet-only baselines would further isolate the contribution of the cross-modal features.

5.2 Labels and measurement interpretation

MMHI-60 contains 60 Likert-style items scored from 1 to 5. Ten subscales correspond to depression, interpersonal sensitivity, anxiety, obsessive-compulsive tendencies, paranoid ideation, hostility, academic stress, maladaptation, emotional disturbance, and psychological imbalance. A subscale mean of at least 2 is treated as positive. Overall mental-health status is positive when the sum of the 60 item scores exceeds 120, and suicidal tendency uses the questionnaire together with psychiatrist-provided screening advice. These labels are useful operational targets for a screening study, but they are not equivalent to structured clinical diagnoses. Throughout the paper, a positive label indicates a screening outcome.

5.3 Metrics and statistical protocol

For each target, we calculate accuracy, weighted precision, weighted recall, and weighted F1. If n_c is the number of examples in class c, the weighted score is $\sum_c (n_c/N) m_c$, where m_c is the class-specific metric. We report fold means from ten stratified folds. For class-aware analysis, we additionally use the original confusion matrices, obtained by averaging fold counts and then normalizing over all cells. From their four-decimal entries, we derive sensitivity $TP/(TP+FN)$, specificity $TN/(TN+FP)$, balanced accuracy (their mean), and macro-F1 (the unweighted mean of positive- and negative-class F1). These approximate aggregate metrics differ from means of fold-level scores. We do not claim statistical significance from the cross-validation means because the evaluation does not include independent repeated trials or an external test cohort.

5.4 Main results

Table 2 reports results for all 12 tasks. The fusion model obtains weighted F1 values between 71.32% and 91.06%, with the lowest scores on obsessive-compulsive tendencies and the highest score on psychological imbalance. Accuracy ranges from 73.04% to 92.77%. Across targets, the model exceeds the reported feature-wise classical averages, although the absolute difference varies with prevalence and label separability.

The highest raw GAME accuracy is for psychological imbalance (92.77% accuracy; 91.06% weighted F1), whereas obsessive-compulsive tendencies have the lowest scores. Class prevalence provides essential context: Figure 1 reports 89.46% negative labels for psychological imbalance, so the reported accuracy exceeds an always-negative predictor by only 3.31 percentage points. The class-aware analysis below shows why this difference alone does not establish practical screening benefit.

For comparison, academic stress reaches 74.18% accuracy against 62.19% negative-class prevalence, a gap of 11.99 percentage points. For obsessive-compulsive tendencies, the gap is 10.64 points (73.04% versus 62.40%). These cohort-based reference calculations show that the highest raw accuracy need not indicate the largest improvement over a constant prediction rule.

5.5 Class-aware screening performance

Table 3 summarizes the existing aggregate confusion matrices. Sensitivity ranges from approximately 24.9% to 52.4%, while specificity ranges from 87.4% to 99.9%. Balanced accuracy is 62.4–70.5%, substantially below the raw accuracy

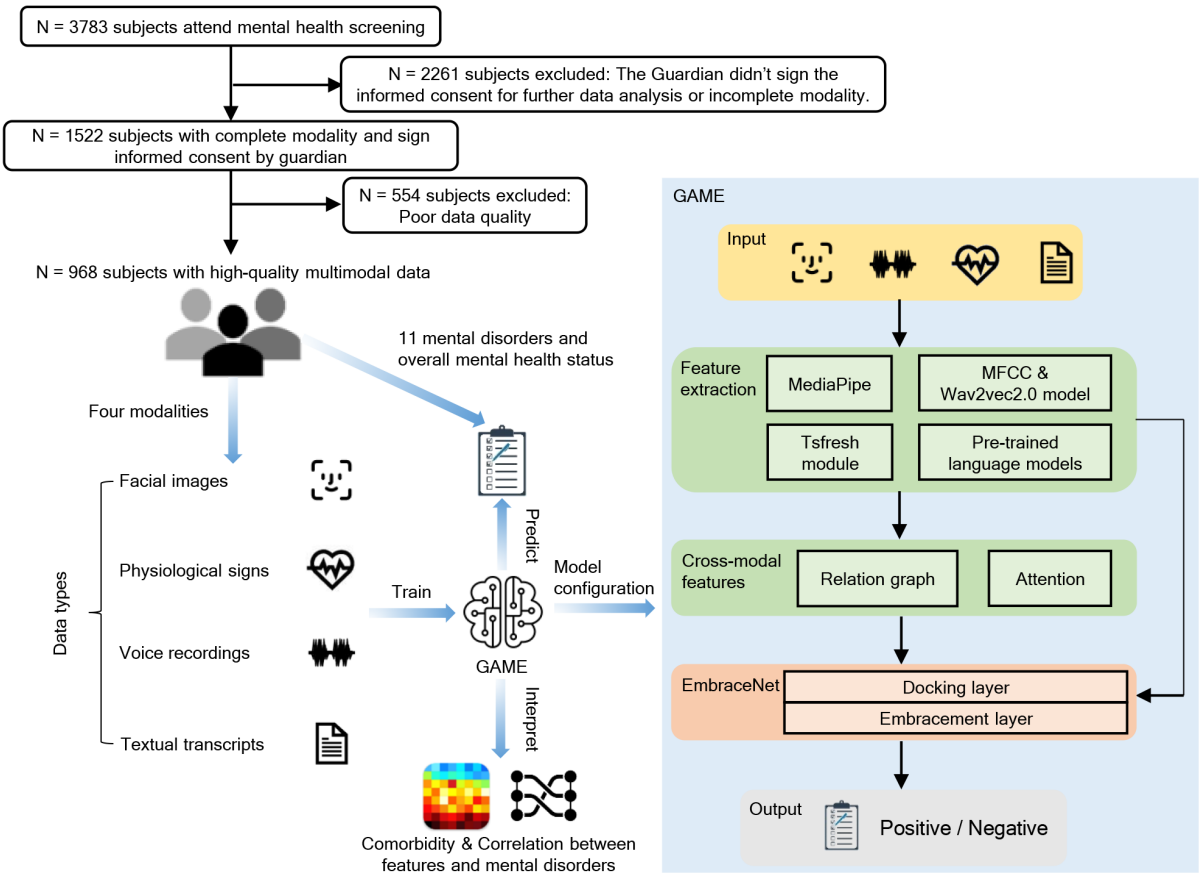


Figure 2: Overview of feature extraction and GAME fusion. Four signal modalities yield eight unimodal representations; relation-graph and attention features are then fused by an EmbraceNet backbone for 12 screening tasks.

range. For suicidal tendency, sensitivity is 30.4%, specificity 99.2%, balanced accuracy 64.8%, and macro-F1 68.6%; about 69.6% of positive screening labels are missed. For psychological imbalance, sensitivity is 33.3% despite 92.77% accuracy. Thus, the evaluated models favor the majority negative class and cannot be regarded as validated stand-alone screening tools.

These values are recomputed from rounded summaries, without retraining or changing the decision rule. Fold-level predictions are needed to quantify uncertainty and compare models reliably; aggregate matrices alone do not recover fold variability. Further evaluation should report confidence intervals and examine sensitivity–specificity trade-offs with threshold selection confined to training/validation data.

6 Analysis and Discussion

6.1 Feature ablation

Removing one representation at a time shows that feature value is task dependent. In accuracy, wav2vec and facial expression features have the largest average effects, followed by RoBERTa, expression nuance, relation graph, eye movement, PERT, attention, physiological signs, and MFCC. For weighted F1, attention and RoBERTa are especially influential. For example, removing attention substantially reduces performance on anxiety and obsessive-compulsive tendencies, while removing wav2vec affects emotional-disturbance prediction. These findings argue against treating one channel as a universal proxy for mental state.

Accuracy and weighted F1 yield different feature rankings (Figure 3). Because weighted F1 weights class-specific scores by support, an improvement does not establish minority-class stability. Ablations describe model dependence on features; they do not identify causal biomarkers or provide a clinical explanation (Chaddad et al., 2023).

6.2 Target-specific feature contributions

The contribution profiles in Figure 3(c) complement the removal experiments. Attention has the highest reported contribution for depression; physiological features for interpersonal sensitiv-

Target	Accuracy (%)				Weighted F1 (%)			
	SVM-RBF	RF	GBDT	GAME	SVM-RBF	RF	GBDT	GAME
Overall mental health	70.15	83.08	82.23	**89.26**	69.78	76.82	76.86	**87.49**
Depression	60.98	72.86	71.30	**80.16**	56.87	63.35	64.19	**76.80**
Interpersonal sensitivity	70.29	80.15	79.07	**85.85**	68.56	72.59	72.88	**82.76**
Anxiety	57.04	68.38	66.44	**77.58**	52.01	58.21	59.49	**74.83**
Obsessive-compulsive	53.67	60.87	59.25	**73.04**	49.44	52.50	54.89	**71.32**
Paranoid ideation	72.17	82.58	81.38	**87.08**	70.13	75.71	75.74	**83.59**
Hostility	70.95	81.47	80.41	**86.78**	69.21	74.24	74.50	**83.54**
Academic stress	57.10	61.11	59.53	**74.18**	49.52	53.02	54.98	**73.06**
Maladaptation	68.16	86.30	84.83	**90.08**	67.17	80.67	80.32	**87.65**
Emotional disturbance	55.67	68.74	66.61	**77.17**	50.56	59.22	59.70	**73.00**
Psychological imbalance	75.15	89.25	88.25	**92.77**	76.19	84.44	84.43	**91.06**
Suicidal tendency	68.45	79.53	78.20	**85.43**	65.71	71.27	71.70	**82.20**

Table 2: Screening performance under stratified evaluation. Classical columns average separate feature-wise experiments; GAME combines all ten representations and reports 10-fold means. Inputs are therefore not matched across columns. All values are percentages.

Target	Sens.	Spec.	Bal. acc.	Macro-F1
Overall mental health	42.1	98.9	70.5	75.5
Depression	33.3	97.5	65.4	67.7
Interpers. sensitivity	30.5	99.4	64.9	68.8
Anxiety	39.3	94.6	66.9	68.6
Obsessive-compulsive	46.2	89.2	67.7	68.4
Paranoid ideation	24.9	99.9	62.4	66.2
Hostility	28.0	99.7	63.9	67.9
Academic stress	52.4	87.4	69.9	70.7
Maladaptation	28.1	99.5	63.8	68.7
Emotional disturb.	30.8	98.2	64.5	65.6
Psychol. imbalance	33.3	99.8	66.5	72.7
Suicidal tendency	30.4	99.2	64.8	68.6

Table 3: Approximate class-aware metrics (%) derived from the original rounded, aggregate confusion matrices. These are post-hoc summaries of the existing predictions, not new experiments or means of fold-level metrics. Positive labels denote screening outcomes.

ity, paranoid ideation, and psychological imbalance; MFCC for anxiety; and PERT for obsessive-compulsive tendencies. RoBERTa ranks highest for hostility, academic stress, and suicidal tendency, while eye movement ranks highest for maladaptation and overall mental-health status. Wav2vec ranks highest for emotional disturbance. These parameter-derived profiles are descriptive summaries; their ranking is not a validated measure of clinical importance.

Contribution ratios and ablation effects answer different questions: the former describe feature use in a fitted model, whereas the latter measure the performance change when an input representation is removed. Their rankings need not coincide. Moreover, removing wav2vec leaves MFCC available, and removing PERT leaves RoBERTa available; these experiments do not remove the entire acoustic or textual modality. The results motivate testing reduced-sensor protocols tailored to a screening objective, with modality-subset evaluation needed to establish which measurements can be omitted.

6.3 Cross-target screening associations

We examine associations through label co-occurrence and cross-target model transfer. Co-occurrence counts participants positive on pairs of screening targets. Transfer applies a model trained for one target to another target's labels, producing a task-association matrix. The first reflects the questionnaire and sampled population; the second reflects predictive similarity in the fitted models.

The analyses identify depression–anxiety, anxiety–emotional disturbance, and associations of suicidal tendency with depression, anxiety, academic stress, and emotional disturbance. Transfer patterns also suggest links involving maladaptation, paranoid ideation, and psychological imbalance. These are hypotheses for clinical follow-up: shared class imbalance and correlated screening labels can produce cross-target accuracy without establishing diagnostic comorbidity. This interpretation follows the emphasis on measurement validity in clinical NLP (Milintsevich et al., 2024; Shani and Stade, 2025).

6.4 Multicultural and multilingual relevance

Chinese transcripts, school-related prompts, and MMHI-60 place the collection in a specific linguis-

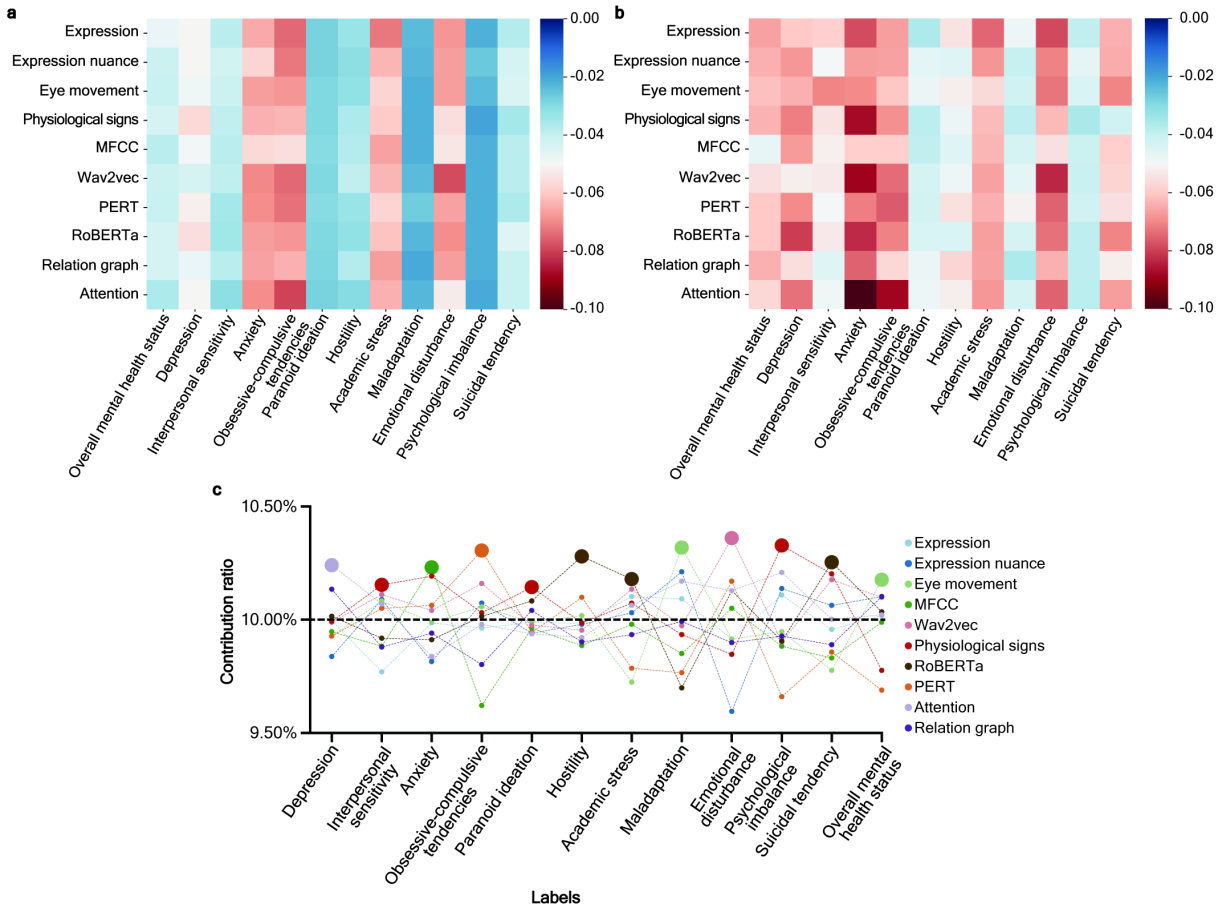


Figure 3: Feature ablation and model contributions. Panels (a) and (b) show changes in accuracy and weighted F1 after removing individual representations; negative values indicate decreases. Panel (c) shows the reported feature contribution ratios for each independently trained target model.

tic and cultural setting. Cross-lingual evaluation, dialect coverage, and measurement-invariance tests remain necessary. Translation alone cannot establish that prompts elicit comparable constructs or that screening thresholds have comparable meaning. Parallel studies should report subgroup performance and examine whether feature contributions and error profiles remain stable across settings (Havaldar et al., 2024).

6.5 Missing modalities and deployment

Complete-case evaluation does not represent deployment-time availability. Head movement, lighting, or occlusion can impair both facial tracking and rPPG; declined recording or failed transcription can remove acoustic or textual inputs. Although EmbraceNet permits modality selection, explicit missing-modality conditions were not evaluated. Future experiments should test modality subsets and realistic availability patterns, subject to consent, and report worst-group as well as average performance.

6.6 Data expansion and model refinement

As an initial study, this work establishes a collection and modeling workflow for subsequent iteration. We plan to extend data collection across schools and improve representation of positive screening cases, subject to consent and ethical approval. Algorithm development will examine class-aware training objectives and threshold selection within training/validation data to improve positive-case detection while tracking false-positive rates. Matched-input baselines and school-held-out evaluation will help assess whether revised fusion methods provide reproducible gains beyond the current cohort. These are planned directions rather than completed experiments.

7 Ethical Considerations

The study was approved by the Life Ethics Committee of Tsinghua Shenzhen International Graduate School, Tsinghua University (Protocol No. 41,

2021; approved June 7, 2021). All participants retained for analysis had written informed consent from both the adolescent and their legal guardian; records lacking guardian consent for further analysis were excluded. Raw audiovisual and psychological data are not publicly released. Access is subject to ethical and privacy review, including for embeddings that may retain identifying information.

Any future use must include human review and an escalation pathway for distress or suicide-related responses. A prediction alone should not determine diagnosis or intervention. Consent, retention, access, and deletion procedures must account for the sensitivity of minors' data and the intended use.

8 Limitations

The data come from Chinese adolescents in a limited geographic and institutional context. Only 25.6% of the initial cohort is retained, and consent or recording-quality selection may alter the population represented. Retained-versus-excluded comparisons should use only data permitted by consent. No school-held-out or external-cohort evaluation is available, so random stratified folds do not establish transfer to new schools, cultures, or age groups.

Labels reflect MMHI-60 and expert screening assessment rather than structured diagnoses. For suicidal tendency, the available description does not establish a reproducible adjudication protocol or inter-rater agreement. Low positive-class sensitivity limits screening utility, while rounded aggregate metrics cannot establish uncertainty. The feature-wise classical comparisons do not isolate architectural gains; matched-input concatenation and EmbraceNet-only comparisons remain necessary.

DTW over heterogeneous feature coordinates is a heuristic, and the contribution profiles and ablations do not establish causal biomarkers. Facial, voice, physiological, and language signals may encode demographic or cultural biases. Missing-modality robustness, temporal stability, threshold calibration, and prospectively supervised clinical use remain untested.

9 Conclusion

We presented an initial study connecting MAPS and GAME: a standardized robot interaction protocol and independently trained multimodal models for 12 adolescent mental-health screening targets. The collection and target-specific analyses provide a foundation for further work in Chinese school settings. Next steps are to expand the cohort, refine the algorithms with attention to positive-case detection, and assess transfer across schools. External validation and clinically supervised evaluation remain necessary to establish practical screening utility.

Acknowledgments

We thank the participants for their time and commitment. This research was supported by the Shenzhen Medical Research Funds (D2301002, C2401007); the National Natural Science Foundation of China (U24A20761, 62473005, 32350410397, 22306048); the Science, Technology and Innovation Commission of Shenzhen Municipality (JCYJ20240813112016022, JCYJ20220530143014032, JCYJ20230807113017035, KCXFZ20211020163813019, JCYJ20230807094803007); the Tsinghua Shenzhen International Graduate School Cross-disciplinary Research and Innovation Fund Research Plan (JC2022009); the Bureau of Planning, Land and Resources of Shenzhen Municipality (Grant (2022) 207); the Lishui Joint Funds of the Zhejiang Provincial Natural Science Foundation of China (LLSQN26H090003); the Sichuan Provincial Natural Science Foundation of China (2026NSFSC1501, 2026NSFSC1221); the Zhejiang Provincial Health and Science Technology Program (General Program) (2025HY1542); the Zhejiang Province Post-Doctoral Research Project Selected Funding (Z12024150); the Shenzhen Nanshan District Artificial Intelligence Scenario Application "Challenge Tackling Mechanism" Project; the Development and Reform Commission of Shenzhen Municipality (F-2024-Z99-503929); the State Key Laboratory of Chemical Oncogenomics, Institute of Biomedical Health Technology and Engineering, Shenzhen Bay Laboratory.

References

Nida Itrat Abbasi, Micol Spitale, Joanna Anderson, Tamsin Ford, Peter B. Jones, and Hatice Gunes. 2022. Can robots help in the evaluation of mental wellbeing in children? an empirical study. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1459–1466. IEEE.

Alexei Baevski, Yuhao 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. Curran Associates, Inc.

Ahmad Chaddad, Jihao Peng, Jian Xu, and Ahmed Bouridane. 2023. Survey of explainable AI techniques in healthcare. *Sensors*, 23(2):634.

Jenny Chim, Adam Tsakalidis, Dimitris Gkoumas, Dana Atzil-Slonim, Yaakov Ophir, Ayah Zirikly, Philip Resnik, and Maria Liakata. 2024. Overview of the CLPsych 2024 shared task: Leveraging large language models to identify evidence of suicidality risk in online posts. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 177–190. Association for Computational Linguistics.

Jun-Ho Choi and Jong-Seok Lee. 2019. EmbraceNet: A robust deep learning architecture for multimodal classification. *Information Fusion*, 51:259–270.

Maximilian Christ, Nils Braun, Julius Neuffer, and Andreas W. Kempa-Liehr. 2018. Time series FeatuRe extraction on basis of scalable hypothesis tests (tsfresh – a Python package). *Neurocomputing*, 307:72–77.

Joshua Cohen, Jennifer Wright-Berryman, Lesley Rohlfs, Donald Wright, Marci Campbell, Debbie Gingrich, Daniel Santel, and John Pestian. 2020. A feasibility study using a machine learning suicide risk prediction model based on open-ended interview language in adolescent therapy sessions. *International Journal of Environmental Research and Public Health*, 17(21):8187.

Yiming Cui, Ziqing Yang, and Ting Liu. 2022. PERT: Pre-training BERT with permuted language model. *arXiv preprint arXiv:2203.06906*.

Maria I. Davila, Gregory F. Lewis, and Stephen W. Porges. 2017. The PhysioCam: A novel non-contact sensor to measure heart rate variability in clinical and field applications. *Frontiers in Public Health*, 5:300.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Zhicheng Du, Qingyang Shi, Jiasheng Lu, Yingshan Liang, Xinyu Zhang, Yiran Wang, and Peiwu Qin. 2025. MAJORScore: A novel metric for evaluating multimodal relevance via joint representation. *arXiv preprint arXiv:2509.21365*.

Dominic Dwyer and Nikolaos Koutsouleris. 2022. Annual research review: Translational machine learning for child and adolescent psychiatry. *Journal of Child Psychology and Psychiatry*, 63(4):421–443.

Margalida Gili, Pere Castellví, Margalida Vives, Alejandro de la Torre-Luque, José Almenara, Maria J Blasco, Ana I Cebrià, Andrea Gabilondo, Mª Angeles Pérez-Ara, A. Miranda-Mendizábal, Carolina Lagares, Oleguer Parés-Badell, José A Piqueras, Tiscar Rodríguez-Jiménez, Jesús Rodríguez-Marín, Victoria Soto-Sanz, Jordi Alonso, and Miquel Roca. 2019. Mental disorders as risk factors for suicidal behavior in young people: A meta-analysis and systematic review of longitudinal studies. *Journal of Affective Disorders*, 245:152–162.

Shreya Havaldar, Salvatore Giorgi, Sunny Rai, Thomas Talhelm, Sharath Chandra Guntuku, and Lyle Ungar. 2024. Building knowledge-guided lexica to model cultural variation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 211–226. Association for Computational Linguistics.

Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. 2021. What makes multi-modal learning better than single (provably). In *Advances in Neural Information Processing Systems*, volume 34, pages 10944–10956. Curran Associates, Inc.

P. N. Johnson-Laird, Francesco Mancini, and Amelia Gangemi. 2006. A hyper-emotion theory of psychological illnesses. *Psychological Review*, 113(4):822–841.

Tony W. Li, Michael Murray, Zander Brumbaugh, Raida Karim, Hanna Lee, Maya Cakmak, and Elin A. Björling. 2023. Tell me about it: Adolescent self-disclosure with an online robot for mental health. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pages 183–187. ACM.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. 2021. SMIL:

Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2302–2310. Association for the Advancement of Artificial Intelligence (AAAI).

Aishik Mandal, Dana Atzil-Slonim, Thamar Solorio, and Iryna Gurevych. 2025. Enhancing depression detection via question-wise modality fusion. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 44–61. Association for Computational Linguistics.

Kirill Milintsevich, Kairit Sirts, and Gaël Dias. 2024. Your model is not predicting depression well and that is why: A case study of PRIMATE dataset. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 166–171. Association for Computational Linguistics.

Ivan Moura, Ariel Teles, Davi Viana, Jean Marques, Luciano Coutinho, and Francisco Silva. 2023. Digital phenotyping of mental health using multimodal sensing of multiple situations of interest: A systematic literature review. *Journal of Biomedical Informatics*, 138:104278.

Samira Rasouli, Garima Gupta, Moojan Ghafurian, and Kerstin Dautenhahn. 2022. Proposed applications of social robots in interventions for children and adolescents with social anxiety. In *Sixteenth International Conference on Tangible, Embedded, and Embodied Interaction*, pages 1–7. ACM.

Chen Shani and Elizabeth Stade. 2025. Measuring mental health variables in computational research: Toward validated, dimensional, and transdiagnostic approaches. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 69–78. Association for Computational Linguistics.

J. S. Wang, Y. Li, and E. S. He. 1997. Development and standardization of mental health inventory of middle-school students in China. *Science of Social Psychology*, (4):15–20. In Chinese; English title translated.

Taeyang Yun, Hyunkuk Lim, Jeonghwan Lee, and Min Song. 2024. TelME: Teacher-leading multimodal fusion network for emotion recognition in conversation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 82–95. Association for Computational Linguistics.

Likun Zhang, Peiwu Qin, Huazhang Ying, Zhicheng Du, Chenying Lu, Minjiang Chen, Liyan Lei, Ziwu Song, Jiaju Chen, Xi Yuan, Canhui Yang, Vijay Pandey, Can Yang Zhang, Dongmei Yu, Peisheng He, Liwei Lin, Wenbo Ding, Xinhui Xing, Chenggang Yan, and 2 others. 2025. A 3.55-μm ultrathin, skin-like mechanoresponsive, compliant, and seamless ionic conductive electrode for epidermal electrophysiological signal acquisition and human-machine interaction. *Exploration*, 5(5):20240232.