Rethinking Depression Prediction from a Fine-Grained Subscore Modeling Perspective via Multi-Task Learning

Zhenguang Wang[1], Bo Li[1], Wenhui Tan[2], Peng Cao[1,†], Yang Wang[3], Jia Duan[3], Fei Wang[3,†], Osmar Zaiane[4]

[1]School of Computer Science and Engineering, Northeastern University, Shenyang, China,
[2]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China,
[3]Early Intervention Unit, Department of Psychiatry,
The Affiliated Brain Hospital of Nanjing Medical University, Nanjing, China,
[4]Department of Computing Science, University of Alberta, Edmonton, Canada

Abstract

Standard depression assessment relies on instruments such as the clinician-rated Hamilton Depression Rating Scale (HAMD) and the patient-reported Patient Health Questionnaire (PHQ-8), but manual scoring is time-consuming and subject to inter-rater variability. Prior automated approaches typically regress a single total score or coarse severity category, lacking the fine-grained subscore-level supervision needed for precise clinical diagnosis. To address this, we propose MTSP (**M**ulti-**T**ask **S**ubscore **P**rediction), a fine-grained model for subscore prediction via multi-task learning. MTSP achieves state-of-the-art MAE on the public E-DAIC dataset (MAE 3.48) with comparable RMSE (4.57) and generalizes well to the public PDCH and a large-scale private clinical dataset (CIDH), outperforming total score regression baselines and Qwen3-14B direct scoring. We further show that multi-task learning is essential, subscore-level supervision improves assessment by better capturing symptom-cluster structure, and prior constraints plus task-level self-paced learning enhance robustness to subscore difficulty and annotation noise. Our code is available at https://github.com/wxcwzg/MTSP/.

1 Introduction

Depression is a prevalent mental health disorder affecting approximately 332 million people worldwide (World Health Organization, 2025). Accurate and timely assessment of depression severity is essential for personalized treatment and longitudinal monitoring. In clinical practice, severity is typically quantified with standardized rating scales. Clinical instruments such as the Hamilton Depression Rating Scale (Hamilton, 1960) and brief self-report questionnaires such as the 8-item Patient Health Questionnaire (PHQ-8) (Kroenke et al., 2009) provide clinically validated measures of depressive symptomatology and are widely used to evaluate symptom severity. However, their repeated use is time- and resource-intensive, relies on trained psychologists or patient engagement, and remains vulnerable to subjectivity and inter-rater or self-report variability, demanding the development of *automated, reliable and exact* depression assessment systems.

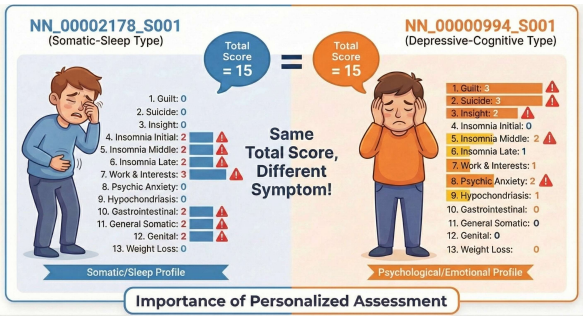


Figure 1: **Illustration of individual-level heterogeneity.** Two representative patients from the CIDH dataset, `NN_00002178_S001` (Patient A) and `NN_00000994_S001` (Patient B), share the same total score (15) but exhibit distinct symptom patterns: Patient A shows a "Somatic–Sleep" pattern characterized by high scores in *Insomnia* (2, 2, 2), *Work & Interests* (3), *GI Symptoms* (2), and *Somatic Symptoms* (2), whereas Patient B presents a "Depressive–Cognitive" profile dominated by *Guilt* (3), *Suicide* (3), and *Psychic Anxiety* (2).

Existing methods predominantly formulate depression assessment as a *coarse-grained* mapping $X \rightarrow y$, where y is a single global target such as the total score or a binary severity category (Ji et al., 2021; Zhao et al., 2025). We argue that this formulation suffers from intrinsic limitations because it overlooks the complex heterogeneity of depressive symptomatology (Buch and Liston, 2020). Patients with identical total scores can exhibit drastically different symptom profiles. For example, Figure 1 shows two representative patients with the same total score but qualitatively distinct patterns

[†]Corresponding authors.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 39659–39672

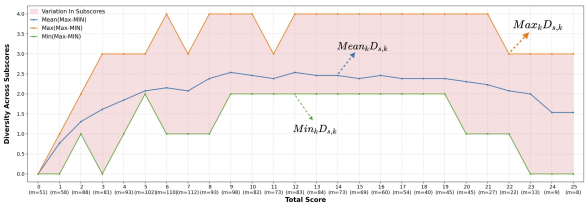


Figure 2: **Statistical discrepancy on the CIDH dataset** The x-axis represents each total score s (with sample size n_s). The y-axis shows the *diversity across subscores*, defined as $D_{s,k} = \max_{i \in \mathcal{P}_s} y_{i,k} - \min_{i \in \mathcal{P}_s} y_{i,k}$, where $y_{i,k}$ denotes the score of the k-th sub-score for the i-th sample, and $\mathcal{P}_s$ is the sample set with the same total score s. Based on this, we compute $\max_k D_{s,k}$, $\text{mean}_k D_{s,k}$, and $\min_k D_{s,k}$, which respectively indicate the max, mean, and min variation across sub-scores with the same total score.

of task-level severity. When a model is trained solely to predict the total score, such individualized patterns inevitably make it difficult to capture clinically meaningful differences between patients. On the CIDH dataset, our statistical analysis (Figure 2) further demonstrates that such discrepancies are not anecdotal but systemic. The resulting curves reveal that, across different levels of depressive severity, even patients who share the same total score exhibit substantial task-level variability, with discrepancies being especially pronounced for those with total scores between 12 and 20.

To bridge this gap, we propose **MTSP** (**M**ulti-**T**ask **S**ubscore **P**rediction), a unified framework that shifts the paradigm from global coarse scoring to fine-grained subscore prediction by formulating it as a multi-task learning (MTL) problem. However, designing an effective MTL framework for depression assessment poses two key challenges.

Challenge 1: How to build the correlation across subscores (tasks) Depression assessment is inherently a multi-task prediction problem: subscores are not independent, but instead have correlation. Treating each subscore as an isolated task neglects these cross-task dependencies and can lead to clinically inconsistent predictions.

Challenge 2: How to solve the inconsistent task difficulty. Different subscores exhibit markedly distinct learning difficulties. Our preliminary experiments (in Fig.4) using Qwen3-14B with zero-shot prompting (non-reasoning mode) for direct PHQ-8 scoring reveal that even state-of-the-art LLMs show highly uneven performance across subscores. This suggests that depression assessment subscores possess inherently different difficulty levels, and that a uniform training schedule can cause difficult tasks to interfere with the learning of easier ones. To address these challenges, MTSP employs a shared-specific task prediction framework. We construct a clinically informed *task correlation graph* that encodes prior knowledge about symptom correlations, and introduce prior constraints that penalize deviations among subscores within the same clinically defined clusters, encouraging predictions consistent with these graph-based priors. In addition, we further incorporate a **Task-level Self-Paced Learning (T-SPL)** strategy that operates on tasks rather than individual samples, adaptively reweighting task losses according to their learning difficulty. Our contributions are summarized as follows:

- **Fine-grained subscore modeling formulation.** We propose MTSP, a fine-grained framework that jointly predicts individual subscores of clinical scales to better describe patients' profiles and predict symptoms.
- **Multi-task learning for subscores.** We capture the structured dependencies among clinical subscores using a MTL mechanism that explicitly models and regularizes the relationships among subscores. Meanwhile, we introduce a task-level self-paced learning strategy that explicitly accounts for subscore difficulty.
- **State-of-the-art performance.** On the public E-DAIC dataset, MTSP achieves state-of-the-art MAE of 3.48 (with comparable RMSE to the best prior work), outperforming both conventional total score regression baselines and direct scoring by the Qwen3-14B large language model. It also excels on the PDCH and CIDH datasets, demonstrating its robustness across different datasets.

2 Related Work

2.1 Natural Language Processing for Clinical Mental Health

Natural language processing (NLP) for mental health assessment has attracted sustained interest. Early work typically relied on feature engineering and rule- or lexicon-based pipelines to extract clinically meaningful signals from text (Gkotsis et al., 2017; Coppersmith et al., 2014; Choudhury et al., 2013). With the rise of pre-trained language models (PLMs), recent approaches reduce reliance on manual features and improve generalization. (Ji et al.,

2021) introduce *MentalBERT*, a domain-specific PLM for mental healthcare that consistently improves mental disorder detection benchmarks. In parallel, general-purpose LLMs such as ChatGPT have been explored in healthcare, and systematic reviews (e.g., (Li et al., 2023)) summarize both their emerging use cases and their limitations in safety-critical clinical scenarios.

A growing line of work focuses on *clinical interviews* and explainable depression assessment (Zhang et al., 2025; Zhao et al., 2025; Mandal et al., 2025; Bi et al., 2025). For example, (Zhang et al., 2025) propose a personalized retrieval-augmented generation framework that retrieves evidence snippets from interview transcripts and generates natural-language explanations for depression decisions. (Mandal et al., 2025) design a question-wise multimodal fusion model that predicts PHQ-8 item scores to enhance interpretability, and (Bi et al., 2025) adopt a multi-agent LLM framework to operationalize structured psychiatric interviews and generate psychometric reasoning traces. Other efforts explore different modeling strategies for text-based depression severity prediction: (Xezonaki et al., 2020) apply hierarchical attention networks with affective conditioning, (Qureshi et al., 2020) jointly learn depression estimation and emotion intensity, (Milintsevich et al., 2023) propose symptom-level prediction through hierarchical regression, and (Agarwal et al., 2024) incorporate psychiatric expertise into symptom-based estimation. Clinical rating instruments in routine practice require *fine-grained level scoring* over multiple correlated symptom dimensions. Most prior work either predicts a single total score or treats symptom-related targets independently, lacking explicit mechanisms to jointly model cross-task structure.

2.2 Multi-Task Learning in Clinical Assessment

Beyond predicting a single total score, a few recent studies have begun to operate at the subscore level. (Steijn et al., 2022) use a multi-target regression framework to predict PHQ-8 item scores as interpretable symptom-level outputs that can be aggregated into overall severity. (Mandal et al., 2025) propose QuestMF, a question-wise multimodal fusion framework that performs ordinal classification for each PHQ-8 item and derives the global PHQ-8 score by summing these predictions. In parallel, (Zhang et al., 2025) formulate EMDRC as an explainable multimodal depression recognition task and introduce PHQ-aware multi-task methods that use PHQ-8 item labels as auxiliary supervision for utterance-level classification and dialogue-level symptom summarization. However, these approaches either treat PHQ subscores independently or do not consider the inherent task relations.

3 Method

3.1 Problem Formulation

Given a clinical interview transcript $\mathbf{x}_i$, our goal is to predict depression subscores $\mathbf{y}_i = (y_{i1}, \ldots, y_{iK})$ in a multi-task regression setting (Fig.3(a)), where K is the number of subscores and $y_{ik} \in \mathbb{R}$ is the score for subscore k of sample i. We support two depression scales: **PHQ-8** is a self-report instrument containing $K{=}8$ subscores (each 0–3), and **HAMD-13** is a clinician-rated scale with $K{=}13$ text-assessable subscores (excluding 4 items requiring behavioral observation). See Appendix B for detailed subscore definitions.

3.2 MTSP

3.2.1 Text Preprocessing

Clinical interview transcripts are first segmented into utterances using a hierarchical strategy: (1) split by newlines, (2) split by speaker tags (e.g., "doctor:", "patient:"), (3) split by sentence delimiters, or (4) split by fixed-length chunks if no natural boundaries exist. Each utterance is then cleaned by removing non-text tags, filler words, and repeated characters. This preprocessing step is crucial for handling long transcripts.

3.2.2 Base Language Model Encoding

The framework architecture is shown in Fig.3(b). First, we use domain-specific pre-trained BERT models as our base encoder to process the segmented utterances. For the English E-DAIC dataset, we employ MentalBERT (Ji et al., 2021), pre-trained on mental health-related text. For the Chinese HAMD-13 datasets (CIDH and PDCH), we use MedBERT-Chinese (Yang et al., 2021), pre-trained on Chinese medical corpora. For each utterance u_j in the transcript, we obtain a contextualized embedding $\mathbf{h}_j = \text{BERT}(u_j)[\text{CLS}]$, where $\mathbf{h}_j \in \mathbb{R}^{d_{\text{bert}}}$ is the [CLS] token embedding that summarizes utterance u_j, and $d_{\text{bert}} = 768$ is the hidden dimension. To improve efficiency, we encode all utterances across samples in a mini-batch, rather than encoding them individually. The sequence of utter-

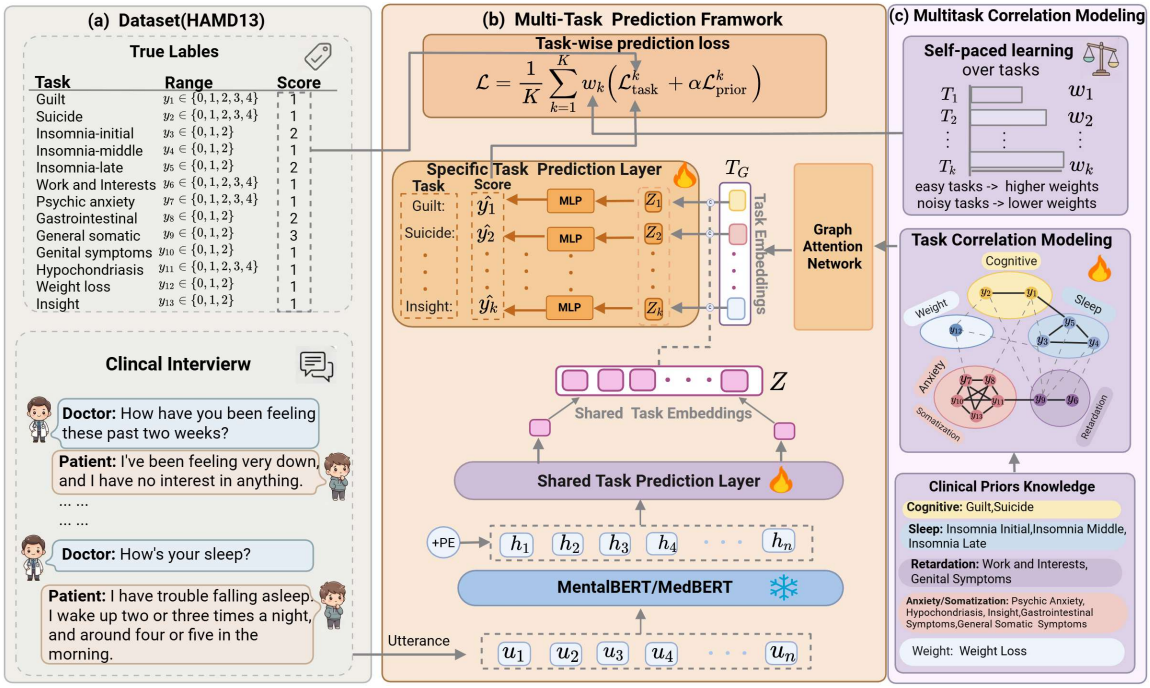


Figure 3: Overview of the proposed MTSP framework (illustrated with the HAMD-13 scale). (a) The illustration of a clinical interview in our dataset. (b) **Multi-task subscore prediction framework**. Each clinical interview is segmented into utterances, which are then encoded by a BERT-based encoder (MentalBERT / MedBERT) to obtain contextualized embeddings. Finally, these embeddings are passed through a shared task encoder and task-specific layers to jointly predict all subscores in a multi-task fashion. (c) **Multitask Correlation Modeling**: **1. task correlation modeling** constructs a clinically informed task correlation graph over symptom subscores and incorporates clinical prior knowledge about symptom clusters as constraints to explicitly regularize relationships among subscores (tasks), while **2. task-level self-paced learning mechanism** adaptively reweights subscore losses according to their difficulty for making our model aware of the task difficulty.

ance embeddings $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_{N_i}] \in \mathbb{R}^{N_i \times d_{\text{bert}}}$ is fed into the Transformer encoder, where N_i is the number of utterances in transcript x_i (padded to the maximum sequence length in each batch).

3.2.3 Shared Transformer Encoder

We apply a shared Transformer encoder to the sequence of utterance embeddings to capture temporal dependencies and long-range interactions: $\mathbf{Z} = \text{TransformerEncoder}(\mathbf{H}\mathbf{W}_{\text{proj}} + \mathbf{b}_{\text{proj}})$, where $\text{TransformerEncoder}(\cdot)$ consists of L_{enc} Transformer encoder layers, $\mathbf{H} \in \mathbb{R}^{N \times d_{\text{bert}}}$ is the sequence of utterance embeddings, $\mathbf{W}_{\text{proj}} \in \mathbb{R}^{d_{\text{bert}} \times d_{\text{enc}}}$ is a learnable projection matrix that maps BERT embeddings to the Transformer hidden dimension d_{enc}, and $\mathbf{b}_{\text{proj}} \in \mathbb{R}^{d_{\text{enc}}}$ is the corresponding bias term.that capture both local utterance semantics and global dialogue context. We then apply global average pooling over the sequence dimension to obtain a multi-task shared representation $\mathbf{z} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{Z}_n$, where $\mathbf{Z}_n \in \mathbb{R}^{d_{\text{enc}}}$ denotes the n-th row of $\mathbf{Z}$ (i.e., the representation of the n-th utterance).

3.2.4 Multitask Correlation Modeling

To explicitly model the structured relationships among clinical subscores, we construct a **task correlation graph** (shown in Fig.3(c)) where each node represents a subscore and edges encode symptom correlations based on clinical prior knowledge. We employ a Graph Attention Network (GAT) (Velickovic et al., 2017) to learn task embeddings that capture these relationships.

The task correlation graph is constructed based on clinically validated symptom clusters (3 clusters for PHQ-8, 5 for HAMD-13; see Appendix C for details). The graph includes two types of edges with learnable weights parameterized from clinical priors:

- **Intra-cluster edges:** Connect subscores within each symptom cluster, with weight ω_{intra} (initialized to 1.0).
- **Inter-cluster edges:** Connect subscores between clusters based on clinical correlations (e.g., Core Depression ↔ Cognitive Function),

with weight ω_{inter} (initialized to 0.6).

These weights are learned during training, allowing the model to adapt the graph structure to the data while respecting clinical priors. Given this graph, we maintain learnable task embeddings $\mathbf{E} \in \mathbb{R}^{K \times d_e}$, where K is the number of subscores and d_e is the task-embedding dimension. We propagate $\mathbf{E}$ over the clinically informed task correlation graph using a series of GAT layers, $\mathbf{T}_G = \text{GAT}(\mathbf{E}, \mathcal{G}; \omega_{\text{intra}}, \omega_{\text{inter}})$, where $\mathcal{G}$ is the task correlation graph and $\omega_{\text{intra}}, \omega_{\text{inter}}$ control intra- and inter-group edge weights. The resulting $\mathbf{T}_G \in \mathbb{R}^{K \times d_g}$ encodes both task-specific information and cross-task dependencies.

3.2.5 Multi-Task Prediction

For each subscore $k \in \{1, \ldots, K\}$, we first obtain a task-specific feature by concatenating the pooled text representation and the projected GAT-enhanced task embedding as $\mathbf{Z}_k = [\mathbf{z}; \text{Proj}_t(\mathbf{t}_k)]$, where $\mathbf{z} \in \mathbb{R}^{d_z}$ is the shared task embedding, $\mathbf{t}_k = \mathbf{T}_G[k] \in \mathbb{R}^{d_g}$ is the embedding for subscore k obtained from the task correlation graph, and $\text{Proj}_t(\cdot)$ is a learnable linear projection applied only to $\mathbf{t}_k$. For continuous score prediction, we normalize predictions to $[0, 1]$ during training and denormalize to the valid range $[0, C_k]$ in inference, where C_k is the maximum score for subscore k. To this end, the subscore is then predicted as $\hat{y}_k = C_k \cdot \sigma(\mathbf{w}_k^\top \mathbf{Z}_k + b_k)$, where $\mathbf{w}_k$ and b_k are the parameters of the task-specific head.

3.2.6 Subscore Prediction Loss

The primary supervision signal is the loss on predictions from each subscore. For sample i and subscore k, let $y_k^{(i)}$ and $\hat{y}_k^{(i)}$ denote the ground-truth and predicted scores, respectively. We use the mean squared error between $\hat{y}_k^{(i)}$ and $y_k^{(i)}$ as the prediction loss, and at the end of each epoch e we compute the average loss for each subscore $k \in \{1, \ldots, K\}$ over the training set as $\bar{\mathcal{L}}_k^{(e)} = \frac{1}{M} \sum_{i=1}^{M} \big(\hat{y}_k^{(i)} - y_k^{(i)}\big)^2$, where M is the number of training samples.

3.2.7 Prior Loss

Given the task correlation graph, we incorporate a constraint loss that encourages predictions of highly correlated subscores within each cluster to be similar across all training examples. Given a training set with M examples, the prior loss is defined as

$$\mathcal{L}_{\text{prior}} = \sum_{i=1}^{M} \sum_{c \in \mathcal{C}} \gamma_c \sum_{(k,k') \in \mathcal{P}_c} w_{kk'}^{(c)} \big(\hat{y}_{ik} - \hat{y}_{ik'}\big)^2, \tag{1}$$

where $\mathcal{C}$ is the set of symptom clusters, $\mathcal{P}_c$ is the set of subscore pairs (k, k') within cluster c, $\hat{y}_{ik}$ and $\hat{y}_{ik'}$ are the predicted scores for subscores k and k' of example i, $w_{kk'}^{(c)} \in [0, 1]$ is the correlation weight between subscores k and k' within cluster c (derived from clinical literature and empirical analysis), and $\gamma_c > 0$ controls the relative importance of cluster.

3.2.8 Task-Level Self-Paced Learning

In practice, different subscores exhibit substantially different learning difficulty: tasks such as *Depressed* are relatively easy to optimize because patients usually respond to these questions explicitly, whereas *Appetite* and *Concentrating* are harder since related information is rarely discussed directly, and *Suicide* is even more challenging because cues about suicidal ideation are often expressed implicitly or euphemistically, leading to noisier supervision. To address this, following self-paced learning (Kumar et al., 2010), we propose a **task-level** curriculum strategy that dynamically adjusts the contribution of each task based on its learning difficulty.

Task Weights. Given the average task-level loss $\bar{\mathcal{L}}_k^{(e)}$, we assign a self-paced weight $w_k^{(e)}$ to each task. Tasks with loss below the threshold receive positive weights proportional to their easiness, while tasks with loss exceeding the threshold are temporarily excluded from training.

$$w_k^{(e)} = \begin{cases} 1 - \dfrac{\bar{\mathcal{L}}_k^{(e)}}{\lambda_{\text{spl}}^{(e)}}, & \text{if } \bar{\mathcal{L}}_k^{(e)} < \lambda_{\text{spl}}^{(e)}, \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

where $\lambda_{\text{spl}}^{(e)}$ is the pace parameter at epoch e.

Pace Schedule. The pace parameter λ_{spl} controls the curriculum progression and increases linearly over epochs: $\lambda_{\text{spl}}^{(e)} = \lambda_{\text{min}} + \frac{e}{E} \cdot (\lambda_{\text{max}} - \lambda_{\text{min}})$, where E is the total number of epochs. We set $\lambda_{\text{min}} = 0.5$ and $\lambda_{\text{max}} = 2.0$. Early in training, only easy tasks (with small loss) are learned, and as training progresses, harder tasks are gradually incorporated. The task weights can be further smoothed using a logarithmic pace function.

3.2.9 Overall Objective.

The overall training objective combines: (1) subscore prediction losses weighted by task-level self-paced weights $w_k^{(e)}$, and (2) the medical domain prior constraint loss $\mathcal{L}_{\text{prior}}$. The task correlation graph with learnable edge weights (ω_{intra} and ω_{inter}) is learned jointly via the GAT, while the cluster constraint loss penalizes deviations among subscores within the same clinically defined clusters. The overall objective function is defined as:

$$\mathcal{L} = \frac{1}{M} \sum_{i=1}^{M} \frac{\sum_{k=1}^{K} w_k^{(e)} \left(\hat{y}_k^{(i)} - y_k^{(i)}\right)^2}{\sum_{k=1}^{K} w_k^{(e)}} + \beta \, \mathcal{L}_{\text{prior}}, \tag{3}$$

where $\beta > 0$ is a hyperparameter controlling the weight of the constraint loss. The normalization by $\sum_{k=1}^{K} w_k^{(e)}$ ensures stable gradient magnitudes as the number of active subscores changes during curriculum learning.

4 Experiments

4.1 Datasets

We evaluate MTSP on three clinical interview datasets, i.e., public **E-DAIC** (Ringeval et al., 2019), **PDCH** (Cao et al., 2025) and private **CIDH (Clinical interview dialogue based on the Hamilton Depression Rating Scale)**, covering both PHQ-8 and HAMD-13 scales. Dataset statistics and experimental settings are provided in Appendix A and E.

4.2 Experimental Results

Table 1 reports results on the E-DAIC dataset, including both development and test performance. All results are averaged over N=5 random seeds; we report mean ± standard deviation.

Comparison with the state of the art. On E-DAIC (PHQ-8), MTSP (Full) achieves a test MAE of 3.48 ± 0.16, outperforming all prior methods. Compared to the best prior work PTTSD-seq2seq (Schmidt et al., 2025) (MAE 3.85), MTSP achieves a 9.6% relative improvement. For RMSE, MTSP obtains 4.57 ± 0.22 versus 4.52 for Schmidt (2025); the difference is marginal, indicating comparable performance. MTSP also outperforms the Qwen3-14B zero-shot baseline (MAE 5.39, 35.4% improvement), demonstrating that our compact model with multi-task subscore supervision is more effective than relying solely on large language model prompting. **Ablation analysis.** We conduct ablation experiments across all three datasets (Table 2) to assess each component's contribution. Comparing individual contributions, the medical domain prior constraint (Prior) provides the largest improvement: on E-DAIC, adding Prior alone reduces MAE from 3.93 to 3.58 (8.9% improvement), while adding T-SPL alone reduces MAE to 3.76 (4.3% improvement). Notably, combining T-SPL and Prior Loss yields additional gains (MAE 3.48 vs. 3.58 / 3.76), demonstrating that these two modules are complementary because they address different aspects of the learning problem: Prior Loss constrains the prediction space based on medical knowledge, while T-SPL optimizes the learning dynamics by scheduling tasks according to difficulty. Moreover, we observe that Prior alone achieves MAE 3.65 (1.6% improvement) and MAE 3.66 (3.9% improvement) on the CIDH and PDCH dataset, respectively. The larger improvements on the PDCH suggest that medical domain priors are especially valuable when training data is limited (PDCH has only 99 samples).

	Dev		Test	
Model	**MAE↓**	**RMSE↓**	**MAE↓**	**RMSE↓**
LLM Baseline:				
Qwen3-14B (Zero-Shot)	5.43	5.49	5.39	5.47
Other Work on E-DAIC:				
(Ray et al., 2019)	–	4.37	4.02	4.73
(Sadeghi et al., 2023)	3.65	5.27	4.26	5.37
(Sadeghi et al., 2024) (Pr3 + Whisper)	3.17	4.51	4.22	5.07
(Schmidt et al., 2025) (s2o-MentalBERT)	3.55	4.58	4.18	5.23
(Schmidt et al., 2025) (s2s-all-MiniLM-L6-v2)	3.47	4.57	3.85	**4.52**
Our Methods (N=5):				
MTSP (Full)	**2.94**±0.12	**4.18**±0.15	**3.48**±0.16	4.57±0.22

Table 1: Main results on the E-DAIC (PHQ-8) dataset. We report mean absolute error (MAE) and root mean squared error (RMSE) on the development and test sets. Our results are averaged over 5 runs with different random seeds. MTSP achieves state-of-the-art MAE with comparable RMSE to (Schmidt et al., 2025).

4.3 Analysis

We conduct detailed analyses to investigate (1) task difficulty variation and the effectiveness of self-paced learning, and (2) the ability to capture fine-grained symptoms.

4.3.1 Task Difficulty Analysis

To understand why task-level self-paced learning is effective, we analyze per-task performance on

Model Variant	E-DAIC (PHQ-8)		CIDH (HAMD-13)		PDCH (HAMD-13)	
	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓
LLM Baseline:						
Qwen3-14B (Zero-Shot)	5.39±0.04	5.47±0.05	3.86±0.06	4.76±0.08	4.00±0.05	5.01±0.07
Our Ablation (MTSP Variants, N=5):						
MTSP (Total Score)	4.02±0.18	4.76±0.20	3.72±0.15	4.65±0.17	4.58±0.22	5.47±0.25
MTSP (Subscore)	3.93±0.17	4.91±0.21	3.71±0.16	4.72±0.19	3.81±0.19	4.79±0.21
+ Task-Level SPL	3.76±0.15	4.82±0.19	3.66±0.14	4.64±0.17	3.74±0.18	4.73±0.20
+ Prior	3.58±0.14	4.55±0.18	3.65±0.14	4.57±0.16	3.66±0.17	4.68±0.19
MTSP (Full)	**3.48**±0.16	**4.57**±0.22	**3.63**±0.14	**4.55**±0.18	**3.54**±0.15	**4.56**±0.20

Table 2: Ablation study of MTSP variants across all three datasets (test set). We report mean ± std (N=5) of total score MAE and RMSE on E-DAIC (PHQ-8), CIDH (HAMD-13), and PDCH (HAMD-13). LLM baselines are included for all datasets.

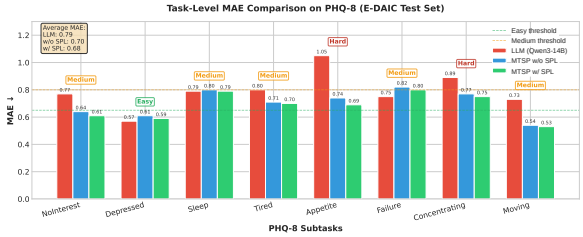


Figure 4: **The Task-level MAE comparison on PHQ-8 (E-DAIC test set).** Each subscore shows three bars: LLM (Qwen3-14B zero-shot, red), MTSP w/o SPL (blue), and MTSP w/ SPL (green). Tasks are annotated with difficulty levels based on LLM performance: *Easy* (MAE $<$ 0.65), *Medium* (0.65–0.80), and *Hard* ($>$ 0.80). T-SPL consistently improves performance across most subscores, especially on hard tasks like *Appetite* and *Concentrating*.

PHQ-8. Figure 4 compares the task-level MAE of Qwen3-14B (zero-shot), MTSP without T-SPL, and MTSP with T-SPL on the E-DAIC test set.

The results reveal a significant difficulty disparity across tasks. Depressed is classified as "Easy" (MAE 0.57) because depressed mood is explicitly discussed in interviews. In contrast, Appetite (MAE 1.05) and Concentrating (MAE 0.89) are "Hard" tasks because their symptoms are often expressed implicitly and in ambiguous ways rather than being mentioned directly. By investigating the learning procedure of T-SPL, we find that it learns tasks in the order of increasing task difficulty. In early epochs (pace $\lambda = 0.2$), only easy tasks receive non-zero weights: *Depressed* ($w = 0.85$) and *NoInterest* ($w = 0.78$) are prioritized, while hard tasks like *Appetite* ($w = 0$) and *Concentrating* ($w = 0$) are excluded. As training progresses and λ increases linearly, medium-difficulty tasks (*Tired*, *Failure*, *Sleep*) are incorporated around epoch 5–10, and hard tasks enter the curriculum after epoch 10. By epoch 15, all tasks have comparable weights ($w \approx 0.7$–0.9), allowing the model to refine predictions across all subscores. This progressive curriculum prevents gradient interference from difficult tasks during early training, reducing average MAE from 0.70 to 0.68. The improvement is especially obvious on hard tasks, i.e., *Appetite*: 0.74 → 0.69, *Concentrating*: 0.77 → 0.75.

4.3.2 Subscore analysis

To evaluate whether MTSP enables fine-grained subscore prediction by capturing individual symptom patterns, especially for patients with identical total scores, we conduct the following analyses on the E-DAIC and CIDH test sets.

Subscore prediction performance. Table 3 reports the task-level comparison between Qwen3-14B (zero-shot), MTSP without task-level SPL, and the full MTSP model on the E-DAIC test set. The full MTSP achieves an average subscore MAE of 0.68, versus 0.79 for the LLM baseline—a 13.9% relative reduction overall. MTSP outperforms the LLM on 6 of the 8 PHQ-8 subscores and ties on *Sleep* (both 0.79). The largest gains are observed on tasks whose symptoms tend to be expressed implicitly in dialogue: *Appetite* (0.69 vs. 1.05, 34.3% improvement), *Moving* (0.53 vs. 0.73, 27.4% improvement), and *NoInterest* (0.61 vs. 0.77, 20.8% improvement). These are precisely the subscores for which cues are rarely stated directly in interview transcripts, and therefore benefit most from subscore-aware multi-task supervision. On two subscores—*Depressed* (0.59 vs. 0.57) and *Failure* (0.80 vs. 0.75)—MTSP performs comparably to or slightly worse than the zero-shot LLM. This is consistent with clinical intuition: "depressed mood" and "feelings of failure" are among the most semantically overt tasks in PHQ-8, typically verbalized

Subscore	LLM (Qwen3-14B) MAE↓	MTSP w/o SPL MAE↓	MTSP (Full) MAE↓	Diff.
NoInterest	0.77	0.64	**0.61**	Med
Depressed	**0.57**	0.61	0.59	Easy
Sleep	**0.79**	0.80	**0.79**	Med
Tired	0.80	0.71	**0.70**	Med
Appetite	1.05	0.74	**0.69**	Hard
Failure	**0.75**	0.82	0.80	Med
Concentrating	0.89	0.77	**0.75**	Hard
Moving	0.73	0.54	**0.53**	Med
Average	0.79	0.70	**0.68**	–

Table 3: Task-level MAE comparison on PHQ-8 (E-DAIC test set). We report the LLM baseline (Qwen3-14B zero-shot), MTSP without task-level self-paced learning (w/o SPL), and the full MTSP model. Difficulty levels are derived from LLM MAE: *Easy* (< 0.65), *Medium* (0.65–0.80), *Hard* (> 0.80). Best result per subscore in **bold**.

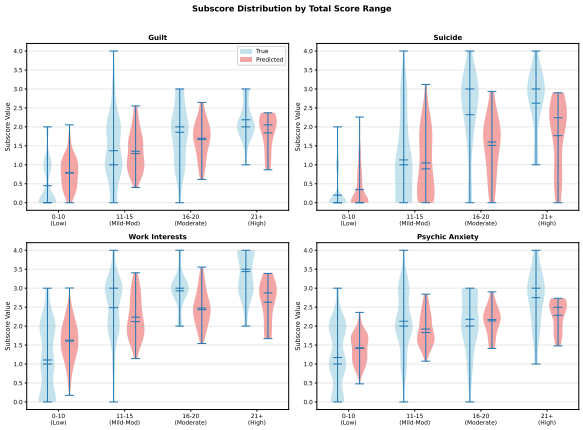


Figure 5: **Subscore distribution comparison by total score range.** Violin plots showing true (blue) vs. predicted (red) subscore distributions for four representative subscores across different total score ranges. The similar distribution shapes indicate that MTSP preserves symptom heterogeneity within each severity group.

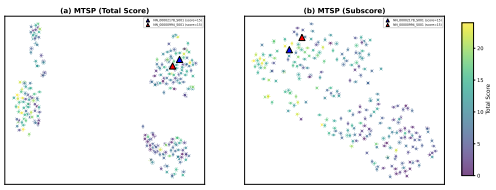
explicitly by patients, so a strong pre-trained LLM can already score them near the annotation noise floor. Comparing the two MTSP variants, adding task-level SPL reduces the average MAE from 0.70 to 0.68, with especially visible gains on the hard tasks *Appetite* (0.74 $\rightarrow$ 0.69) and *Concentrating* (0.77 $\rightarrow$ 0.75), confirming that curriculum scheduling refines difficult subscores without degrading easy ones.

Subscore distribution analysis. To validate that MTSP captures fine-grained subscores, Figure 5 shows the subscore distributions for each total score range using violin plots. For each total score range, we compare the true subscore distributions (blue) with predicted distributions (red). The similarity in distribution shapes indicates that MTSP captures the underlying symptom patterns rather than merely predicting average values.

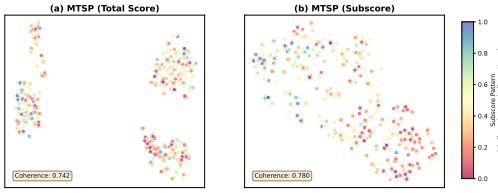
To further illustrate how MTSP learns clinically meaningful representations, we visualize the learned features using t-SNE (van der Maaten and Hinton, 2008). We extract the final-layer features (before the output heads) for all test samples and project them to 2D space. Fig. 6 shows the data distributions obtained by MTSP (Total Score) and MTSP (Subscore). Patients A (▲) and B (▲) mentioned in the introduction section have the same total score. In Fig. 6a, for MTSP (Total Score), A and B are mapped close together because they share the same total score. In contrast, MTSP (Subscore) correctly separates them, recognizing their clinically distinct symptom profiles. This validates that MTSP (Subscore) learns representations that separate patients with different symptom patterns even when total scores are the same—a capability absent in total score supervision. Figure 6b provides complementary evidence by coloring points according to subscore pattern similarity. We project each patient's 13-dim subscore vector to a 1D score via PCA, where similar colors indicate similar symptom profiles. In MTSP (Total Score), the points within the same clusters exhibit varying colors. In contrast, MTSP (Subscore) shows that spatially close points share similar colors, indicating that patients with similar symptom profiles are mapped to nearby regions in the feature space. This confirms that multi-task subscore supervision enables the model to learn symptom-aware and discriminative representations that preserve patient heterogeneity, enabling fine-grained clinical assessment that captures individual patient differences.

5 Conclusion

We have proposed MTSP, a multi-task subscore prediction framework for automated depression assessment that shifts the paradigm from coarse total score regression to fine-grained subscore modeling. MTSP comprises three key components: (1) a multi-task prediction framework that jointly predicts all subscores, providing richer supervision signals than single total score prediction; (2) a task correlation graph with prior constraints that explicitly models clinically validated symptom relationships; and (3) task-level self-paced learning that adaptively handles varying subscore difficulty



(a) Colored by total score. Two patients with a total score of 15 are highlighted: A (▲) and B (▲) have different symptom patterns.



(b) Colored by subscore pattern similarity: similar colors indicate similar symptom profiles.

Figure 6: **Data distribution visualization via t-SNE.**

during training. Experiments on E-DAIC (PHQ-8), CIDH (HAMD-13), and PDCH (HAMD-13) datasets demonstrate that MTSP achieves state-of-the-art MAE, outperforming both total score regression baselines and LLM scoring. These results suggest that fine-grained subscore modeling via multi-task learning is a promising direction for building clinically meaningful automated depression assessment systems.

Limitations

Our evaluation is limited to three datasets (E-DAIC, CIDH, PDCH) from specific clinical contexts. While CIDH provides a relatively large sample (1,689 interviews), PDCH contains only 99 samples, and E-DAIC focuses on virtual interviewer interactions rather than real doctor-patient consultations. Future work should validate on larger, more diverse populations across different healthcare systems.

We assess only 13 text-based HAMD subscores, excluding 4 items (Depressed Mood, Retardation, Agitation, Depersonalization) that require direct behavioral observation. Extending MTSP to multimodal inputs would enable complete HAMD-17 assessment but requires careful handling of video/audio data.

The model has not undergone prospective clinical validation. Real-world deployment would require collaboration with psychiatrists to evaluate failure modes, examine potential biases across patient subgroups, and establish appropriate clinical workflows. Additionally, HAMD annotations themselves have known inter-rater variability, and our model may inherit these annotation inconsistencies.

Finally, we emphasize that MTSP is designed as a clinical decision support tool, not a replacement for professional psychiatric evaluation. Automated depression assessment raises important questions about privacy, accountability, and the appropriate role of AI in mental healthcare.

Acknowledgements

We thank the anonymous reviewers and the area chair for their constructive feedback. This work was supported by the National Natural Science Foundation of China (62076059, 82501861), the Science and Technology Joint Project of Liaoning Province (2023JH2/101700367), the Natural Science Foundation of Jiangsu Province (BK20240272), the Special Funds for Health Science and Technology Development of Nanjing Municipal Health Commission (YKK24188), and the General Project of the Science and Technology Development Foundation of Nanjing Medical University (NMUB20230205).

References

Navneet Agarwal, Kirill Milintsevich, Lucie Metivier, Maud Rotharmel, Gaël Dias, and Sonia Dollfus. 2024. Analyzing symptom-based depression level estimation through the prism of psychiatric expertise. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 974–983, Torino, Italia. ELRA and ICCL.

Robert Michael Bagby, Andrew G. Ryder, Deborah R. Schuller, and Margarita B. Marshall. 2004. The hamilton depression rating scale: has the gold standard become a lead weight? *The American journal of psychiatry*, 161 12:2163–77.

Guanqun Bi, Zhuang Chen, Zhou Liu, Hongkai Wang, Xiyao Xiao, Yuqiang Xie, Wen Zhang, Yongkang Huang, Yuxuan Chen, Libiao Peng, Yi Feng, and Minlie Huang. 2025. Magi: Multi-agent guided interview for psychiatric assessment. In *Annual Meeting of the Association for Computational Linguistics*.

Amanda M. Buch and Conor Liston. 2020. Dissecting diagnostic heterogeneity in depression by integrating neuroimaging and genetics. *Neuropsychopharmacology*, 46:156–175.

Isobel M. Cameron, John R. Crawford, Kenneth Lawton, and Ian Cameron Reid. 2008. Psychometric

comparison of phq-9 and hads for measuring depression severity in primary care. *The British journal of general practice : the journal of the Royal College of General Practitioners*, 58 546:32–6.

Pengfei Cao, Yuanzhe Zhang, Chenxiang Zhang, Wei Chen, Yan Liu, Shuang Xu, Miao Xu, Wenqing Jin, Jinjie Xu, Dan Wang, Wei Wang, Xue Wang, Wen Wang, Yan ping Ren, Jun Zhao, Rena Li, and Kang Liu. 2025. A multimodal depression consultation dataset of speech and text with hamd-17 assessments. *Scientific Data*, 12.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. *Proceedings of the International AAAI Conference on Web and Social Media*.

Glen A. Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in twitter. In *CLPsych@ACL*.

Chengfeng Dou, Chong Liu, Fan Yang, Fei Li, Jiyuan Jia, Mingyang Chen, Qiang Ju, Shuai Wang, Shunya Dang, Tianpeng Li, Xiangrong Zeng, Yijie Zhou, Chenzheng Zhu, Dawei Pan, Fei Deng, Guangwei Ai, Guosheng Dong, Hongda Zhang, Jinyang Tai, and 14 others. 2025. Baichuan-m2: Scaling medical capability with large verifier system. *ArXiv*, abs/2509.02208.

Robert D. Gibbons, David C. Clark, and David J. Kupfer. 1993. Exactly what does the hamilton depression rating scale measure? *Journal of psychiatric research*, 27 3:259–73.

George Gkotsis, Anika Oellrich, Sumithra Velupillai, Maria Liakata, Tim J. P. Hubbard, Richard J. B. Dobson, and Rina Dutta. 2017. Characterisation of mental health conditions in social media using informed deep learning. *Scientific Reports*, 7.

Max Hamilton. 1960. A rating scale for depression. *Journal of Neurology, Neurosurgery & Psychiatry*, 23:56 – 62.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and E. Cambria. 2021. Mentalbert: Publicly available pretrained language models for mental healthcare. *ArXiv*, abs/2110.15621.

Kurt Kroenke, Robert L. Spitzer, and Janet B W Williams. 2001. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16 9:606–13.

Kurt Kroenke, Tara W. Strine, Robert L. Spitzer, Janet B W 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 1-3:163–73.

M. Pawan Kumar, Ben Packer, and Daphne Koller. 2010. Self-paced learning for latent variable models. In *Neural Information Processing Systems*.

Jianning Li, Amin Dada, Behrus Puladi, Jens Kleesiek, and Jan Egger. 2023. Chatgpt in healthcare: A taxonomy and systematic review. *Computer methods and programs in biomedicine*, 245:108013.

Aishik Mandal, Dana Atzil-Slonim, Thamar Solorio, and Iryna Gurevych. 2025. Enhancing depression detection via question-wise modality fusion. *ArXiv*, abs/2503.20496.

Kirill Milintsevich, Kairit Sirts, and Gael Dias. 2023. Towards automatic text-based estimation of depression through symptom prediction. *Brain Informatics*, 10.

Syed Arbaaz Qureshi, G. Dias, Mohammed Hasanuzzaman, and Sriparna Saha. 2020. Improving depression level estimation by concurrently learning emotion intensity. *IEEE Computational Intelligence Magazine*, 15:47–59.

Anupama Ray, Siddharth Krishna Kumar, R. Venkata Siva Reddy, Prerana Mukherjee, and Ritu Garg. 2019. Multi-level attention network using text, audio and video for depression prediction. *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*.

Fabien Ringeval, Björn Schuller, Michel F. 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, M. Soleymani, and Maja Pantic. 2019. Avec 2019 workshop and challenge: State-of-mind, detecting depression with ai, and cross-cultural affect recognition. *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*.

Misha Sadeghi, Bernhard Egger, Reza Agahi, Robert Richer, Klara Capito, Lydia Helene Rupp, Lena Schindler-Gmelch, Matthias Berking, and Bjoern M. Eskofier. 2023. Exploring the capabilities of a language model-only approach for depression detection in text data. *2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–5.

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

Fabian Schmidt, Seyedehmoniba Ravan, and Vladimir Vlassov. 2025. Probabilistic textual time series depression detection. *ArXiv*, abs/2511.04476.

Floris Van Steijn, Gizem Sogancioglu, and Heysem Kaya. 2022. Text-based interpretable depression severity modeling via symptom predictions. *Proceedings of the 2022 International Conference on Multimodal Interaction*.

Laurens van der Maaten and Geoffrey E. Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2017. Graph attention networks. *ArXiv*, abs/1710.10903.

World Health Organization. 2025. Depressive disorder (depression). https://www.who.int/news-room/fact-sheets/detail/depression. Accessed: 2025-12-15.

Danai Xezonaki, Georgios Paraskevopoulos, Alexandros Potamianos, and Shrikanth Narayanan. 2020. Affective Conditioning on Hierarchical Attention Networks Applied to Depression Detection from Transcribed Clinical Interviews. In *Interspeech 2020*, pages 4556–4560.

Feihong Yang, Xuwen Wang, and Jiao Li. 2021. Exploration and research on applying BERT to chinese clinical natural language processing. GitHub repository. Original title in Chinese: "BERT". Accessed: 2026-01-04.

Linhai Zhang, Ziyang Gao, Deyu Zhou, and Yulan He. 2025. Explainable depression detection in clinical interviews with personalized retrieval-augmented generation. In *Annual Meeting of the Association for Computational Linguistics*.

Xianbing Zhao, Yi Lyu, Di Wang, and Buzhou Tang. 2025. Predicting depression in screening interviews from interactive multi-theme collaboration. In *Annual Meeting of the Association for Computational Linguistics*.

A Dataset Details

E-DAIC (Extended Distress Analysis Interview Corpus). E-DAIC (Ringeval et al., 2019) is a publicly available dataset from the AVEC 2019 challenge, containing clinical interviews with PHQ-8 annotations. The dataset includes 275 audio recordings and transcripts of structured psychiatric interviews conducted by a virtual interviewer (Ellie). Each interview is annotated with 8 PHQ-8 subscore scores (each 0–3) and a total score (0–24). We use the standard train/validation/test split (163/56/56 samples) for evaluation.

CIDH (Clinical interview dialogue based on the Hamilton Depression Rating Scale). This dataset contains 1,689 clinical interviews with HAMD-13 annotations, collected in collaboration with Nanjing Brain Hospital (Nanjing Medical University Affiliated Brain Hospital). The dataset includes 271 male and 621 female participants, with ages ranging from 11 to 54 years. Each interview is a structured psychiatric evaluation transcript conducted by trained clinicians. Due to the excessive length of raw clinical transcripts (often exceeding 10,000 tokens), we use Baichuan-M2-32B (Dou et al., 2025) to generate structured summaries for each HAMD-13 subscore. Each summarized transcript contains concise descriptions relevant to all 13 HAMD subscores (see Appendix G for the summarization prompt). All data were collected with informed consent and anonymized. We use an approximate 70/15/15 split, resulting in 1,182/253/254 samples for train/validation/test. All models are trained on the CIDH training set, and hyperparameters are selected on the validation set.

PDCH (Public Depression Consultation in Hospital). PDCH (Cao et al., 2025) is a publicly available multimodal dataset containing clinical psychiatric interviews with HAMD-17 annotations. The dataset includes audio, video, and text transcripts of real doctor-patient consultations. We extract the 13 text-assessable tasks (excluding 4 items requiring direct behavioral observation: Depressed Mood, Retardation, Agitation, and Depersonalization/Derealization) for evaluation. We use a 70/15/15 split, resulting in 69/15/15 samples for train/validation/test.

Statistic	E-DAIC	CIDH	PDCH
Scale	PHQ-8	HAMD-13	HAMD-13
Total samples	275	1,689	99
Age range	-	11–54	18–65
Gender (M/F)	-	271/621*	-
Subscore	8	13	13
Score range	0–24	0–36	0–36
Avg. total score	-	10.2	11.4

Table 4: Summary of dataset statistics. E-DAIC uses PHQ-8 (8 subscales), while CIDH and PDCH use HAMD-13 (13 text-assessable subscales). *Gender counts refer to 892 unique patients; some patients have multiple interview sessions.

B Subscore Definitions

PHQ-8 Subscores (each scored 0–3): (1) NoInterest, (2) Depressed, (3) Sleep, (4) Tired, (5) Appetite, (6) Failure, (7) Concentration, (8) Psychomotor.

HAMD-13 Subscores: (1) Guilt (0–4), (2) Suicide (0–4), (3) Insomnia-initial (0–2), (4) Insomnia-middle (0–2), (5) Insomnia-late (0–2), (6) Work and Interests (0–4), (7) Psychic Anxiety (0–4), (8) GI Symptoms (0–2), (9) Somatic Symptoms (0–2),

(10) Genital Symptoms (0–2), (11) Hypochondriasis (0–4), (12) Weight Loss (0–2), (13) Insight (0–2).

C Symptom Cluster Definitions

The task correlation graph is constructed based on clinically validated symptom clusters (Kroenke et al., 2001, 2009; Cameron et al., 2008; Bagby et al., 2004; Gibbons et al., 1993).

PHQ-8 Clusters (3 clusters):

- *Core Depression*: NoInterest, Depressed, Failure
- *Cognitive Function*: Concentration, Psychomotor
- *Somatic Symptoms*: Sleep, Tired, Appetite

HAMD-13 Clusters (5 clusters):

- *Cognitive*: Guilt, Suicide
- *Sleep*: Insomnia-Initial, Insomnia-Middle, Insomnia-Late
- *Retardation*: Work & Interests, Genital Symptoms
- *Anxiety/Somatization*: Psychic Anxiety, GI Symptoms, Somatic Symptoms, Hypochondriasis, Insight
- *Weight*: Weight Loss

The graph structure includes two types of edges:

- **Intra-cluster edges:** Connect subscores within each symptom cluster, reflecting higher correlations among related symptoms. Initialized with weight $\omega_{\text{intra}} = 1.0$.
- **Inter-cluster edges:** Connect subscores between clusters based on clinical correlations (e.g., Core Depression $\leftrightarrow$ Cognitive Function, Sleep $\leftrightarrow$ Concentration). Initialized with weight $\omega_{\text{inter}} = 0.6$.

D Evaluation Metrics

Let $y_{i,k}$ and $\hat{y}_{i,k}$ denote the ground-truth and predicted scores for sample $i \in \{1, \ldots, M\}$ and task $k \in \{1, \ldots, K\}$.

- **Mean Absolute Error (MAE):** Our primary metric, defined as

$$\text{MAE} = \frac{1}{MK} \sum_{i=1}^{M} \sum_{k=1}^{K} \left| \hat{y}_{i,k} - y_{i,k} \right|.$$

- **Root Mean Squared Error (RMSE):** Defined as

$$\text{RMSE} = \sqrt{\frac{1}{MK} \sum_{i=1}^{M} \sum_{k=1}^{K} (\hat{y}_{i,k} - y_{i,k})^2},$$

which penalizes larger errors more heavily than MAE.

- **Task-level MAE:** MAE computed separately for each task k,

$$\text{MAE}_k = \frac{1}{M} \sum_{i=1}^{M} \left| \hat{y}_{i,k} - y_{i,k} \right|,$$

which allows us to analyze which symptoms are easier or harder to predict.

- **Total Score MAE and RMSE:** Error metrics on the summed score per interview,

$$\text{MAE}_{\text{total}} = \frac{1}{M} \sum_{i=1}^{M} \left| \sum_{k=1}^{K} \hat{y}_{i,k} - \sum_{k=1}^{K} y_{i,k} \right|,$$

$$\text{RMSE}_{\text{total}} = \sqrt{\frac{1}{M} \sum_{i=1}^{M} \left(\sum_{k=1}^{K} \hat{y}_{i,k} - \sum_{k=1}^{K} y_{i,k} \right)^2},$$

reflecting error at the overall severity level.

- **Task Accuracy:** Exact-match accuracy, i.e., the proportion of task predictions that match the ground-truth label, averaged over all tasks and samples:

$$\text{Acc} = \frac{1}{MK} \sum_{i=1}^{M} \sum_{k=1}^{K} \mathbb{I}\left[\hat{y}_{i,k} = y_{i,k} \right].$$

E Detailed Experimental Setup

E.1 Training Setup

All models are trained on the respective training splits (E-DAIC for PHQ-8, CIDH and PDCH for HAMD-13), and we select hyperparameters based on validation performance. We use Adam optimizer with learning rate 2×10^{-4} and cosine annealing learning rate schedule with minimum learning rate 1×10^{-4}. The mini-batch size is 50. We train for 80 epochs with early stopping (patience=15) based on validation total score MAE. Experiments are repeated with multiple random seeds, and we report mean and standard deviation.

E.2 Model Configuration

Encoder. For the English E-DAIC dataset, we use MentalBERT (Ji et al., 2021), a BERT model pre-trained on mental health-related text. For the Chinese HAMD-13 datasets (CIDH and PDCH), we use MedBERT-Chinese (Yang et al., 2021), a BERT model pre-trained on Chinese medical corpora. The BERT embedding dimension is d_{bert}=768. All experiments are conducted on a single NVIDIA A100 80GB GPU.

Transformer encoder. MTSP uses L_{enc}=2 Transformer encoder layers with hidden dimension d=200, 4 attention heads, feedforward dimension $4d$=800, and dropout rate 0.3.

Task correlation graph GAT. The task correlation graph module is configured as follows:

- Task embedding dimension: d_e=64
- GAT hidden dimension: d_g=128
- Number of GAT layers: L_{gat}=2
- Number of attention heads: H=4
- GAT dropout rate: 0.1
- LeakyReLU negative slope: α=0.2
- Edge weight initialization: ω_{intra}=1.0, ω_{inter}=0.6
- Fusion type: concatenation

Loss weights. For PHQ-8, we set medical domain prior constraint weights γ_1=1.0 (Core Depression), γ_2=0.5 (Cognitive), γ_3=0.8 (Somatic) and overall constraint weight β=0.01. For HAMD-13, we use five clusters with weights γ_1=1.0 (Cognitive), γ_2=0.8 (Sleep), γ_3=0.6 (Retardation), γ_4=0.5 (Anxiety/Somatization), γ_5=0.3 (Weight) and β=0.01.

Self-paced learning. The pace parameter λ_{spl} increases linearly over epochs, with task weights computed using a linear pace function. We set λ_{min}=0.5 and λ_{max}=2.0.

F Hyperparameter Sensitivity

F.1 Implementation Details

We implement MTSP in PyTorch with Hugging Face Transformers and run all experiments on NVIDIA GPUs. For E-DAIC (English), we use MentalBERT (Ji et al., 2021) as the base encoder; for HAMD-13 datasets (Chinese), we use MedBERT-Chinese (Yang et al., 2021). The Transformer encoder uses $L = 2$ layers with hidden dimension $d_{\text{tfm}} = 200$, 4 attention heads, and dropout rate 0.3. The task correlation graph module uses learnable task embeddings with dimension $d_e = 64$, GAT hidden dimension $d_g = 128$, $L_{\text{gat}} = 2$ GAT layers with $H = 4$ attention heads, and dropout rate 0.1. We optimize with Adam (learning rate 2×10^{-4}) using mini-batch size 50, and train for 80 epochs with early stopping (patience=15) based on validation performance. For PHQ-8, we apply prior constraint with weights $\gamma_1 = 1.0$, $\gamma_2 = 0.5$, $\gamma_3 = 0.8$ and constraint weight $\beta = 0.01$. Task-level self-paced learning uses linear lambda growth, where $\lambda_{\text{min}} = 0.5$ and $\lambda_{\text{max}} = 2.0$.

We conducted sensitivity analysis on key hyperparameters:

Hyperparameter	Range	Optimal
Transformer Encoder:		
Transformer layers (L_{enc})	1, 2, 3	2
Hidden dimension (d)	100, 200, 300	200
Attention heads	2, 4, 8	4
Task Correlation Graph GAT:		
Task embed dim (d_e)	32, 64, 128	64
GAT hidden dim (d_g)	64, 128, 200	128
GAT layers (L_{gat})	1, 2, 3	2
GAT heads (H)	2, 4, 8	4
ω_{inter}	0.4, 0.6, 0.8	0.6
Fusion type	concat, add, gate	concat
Training:		
Learning rate	1e-4, 2e-4, 5e-4	2e-4
Batch size	16, 32, 50	50

Table 5: Hyperparameter sensitivity analysis for MTSP

G Prompts

G.1 Summarization Prompt (HAMD-13)

The following prompt generates structured summaries from raw clinical transcripts using Baichuan-M2-32B:

```
You are a professional psychiatrist.
Analyze the clinical interview and extract
information for HAMD-13 assessment.

subscores:
1. Guilt (self-blame feelings)
2. Suicide (ideation or attempts)
3. Insomnia-Initial (falling asleep)
4. Insomnia-Middle (waking at night)
5. Insomnia-Late (early awakening)
6. Work/Interests (reduced interest)
7. Psychic Anxiety (anxiety symptoms)
8. GI Symptoms (appetite, digestion)
9. Somatic Symptoms (fatigue)
10. Genital Symptoms (libido loss)
```

```
11. Hypochondriasis (health concerns)
12. Weight Loss (weight changes)
13. Insight (illness awareness)

[Transcript]: {transcript}

Output a summary for each subscore.
```

G.2 HAMD-13 Scoring Prompt

The prompt for Qwen3-14B zero-shot HAMD-13 scoring:

```
You are a psychiatrist using HAMD-13.
Score each subscore based on the summary.

Guidelines:
- Guilt (0-4): 0=absent, 4=delusions
- Suicide (0-4): 0=absent, 4=attempts
- Insomnia-Initial (0-2): 0=none, 2=nightly
- Insomnia-Middle (0-2): 0=none, 2=nightly
- Insomnia-Late (0-2): 0=none, 2=nightly
- Work/Interests (0-4): 0=normal, 4=stopped
- Psychic Anxiety (0-4): 0=absent, 4=severe
- GI Symptoms (0-2): 0=absent, 2=severe
- Somatic Symptoms (0-2): 0=absent, 2=severe
- Genital Symptoms (0-2): 0=absent, 2=severe
- Hypochondriasis (0-4): 0=absent, 4=delusions
- Weight Loss (0-2): 0=none, 2=obvious
- Insight (0-2): 0=aware, 2=denies

[Summary]: {summary}

Output JSON: {"Guilt": X, "Suicide": X,
  "Insomnia-Initial": X, "Insomnia-Middle": X,
  "Insomnia-Late": X, "Work/Interests": X,
  "Psychic Anxiety": X, "GI Symptoms": X,
  "Somatic Symptoms": X, "Genital Symptoms": X,
  "Hypochondriasis": X, "Weight Loss": X,
  "Insight": X}
```

G.3 PHQ-8 Scoring Prompt

The prompt for Qwen3-14B zero-shot PHQ-8 scoring:

```
# Task:
Based on the given interview conversation
excerpts with the patient, assign a PHQ-8
score (0-3) for each of the 8 factors below.

If a factor is not explicitly mentioned in
the interview, you must still estimate a
reasonable score by inferring from the
overall patient situation and context.

# PHQ-8 items and scoring rules:
Over the last 2 weeks, how often has the
patient been bothered by the following?

# Items:
1. NoInterest - Little interest or pleasure
   in doing things.
2. Depressed - Feeling down, depressed,
   or hopeless.
3. Sleep - Trouble falling or staying asleep,
   or sleeping too much.
4. Tired - Feeling tired or having
   little energy.
5. Appetite - Poor appetite or overeating.
6. Failure - Feeling bad about yourself,
   or that you are a failure.
7. Concentrating - Trouble concentrating
   on things.
8. Moving - Moving or speaking so slowly
   that others noticed, or being fidgety.

# Each PHQ-8 item is rated on a 4-point scale:
0 = Not at all
1 = Several days
2 = More than half the days
3 = Nearly every day

[Transcript]: {transcript}

# Response Format:
Output all 8 factors in this format:
- NoInterest: 0
- Depressed: 0
- Sleep: 0
- Tired: 0
- Appetite: 0
- Failure: 0
- Concentrating: 0
- Moving: 0
```