

# Voices, Faces, and Feelings: Multi-modal Emotion-Cognition Captioning for Mental Health Understanding

Zhiyuan Zhou<sup>1</sup>, Yanrong Guo<sup>1\*</sup>, Shijie Hao<sup>1\*</sup>

<sup>1</sup>Hefei University of Technology  
zhouzhiyuan.hfut@gmail.com, yrguo@hfut.edu.cn, hfut.hsj@gmail.com

## Abstract

Emotional and cognitive factors are essential for understanding mental health disorders. However, existing methods often treat multi-modal data as classification tasks, limiting interpretability especially for emotion and cognition. Although large language models (LLMs) offer opportunities for mental health analysis, they mainly rely on textual semantics and overlook fine-grained emotional and cognitive cues in multi-modal inputs. While some studies incorporate emotional features via transfer learning, their connection to mental health conditions remains implicit. To address these issues, we propose ECMC, a novel task that aims at generating natural language descriptions of emotional and cognitive states from multi-modal data, and producing emotion-cognition profiles that improve both the accuracy and interpretability of mental health assessments. We adopt an encoder-decoder architecture, where modality-specific encoders extract features, which are fused by a dual-stream BridgeNet based on Q-former. Contrastive learning enhances the extraction of emotional and cognitive features. A LLaMA decoder then aligns these features with annotated captions to produce detailed descriptions. Extensive objective and subjective evaluations demonstrate that: 1) ECMC outperforms existing multi-modal LLMs and mental health models in generating emotion-cognition captions; 2) the generated emotion-cognition profiles significantly improve assistive diagnosis and interpretability in mental health analysis.

## Introduction

Mental health disorders, such as depression, significantly affect physical and psychological well-being. According to the World Health Organization (WHO), depression affects over 300 million people worldwide, and untreated mental disorders will account for 13% of the total disease burden by 2030 (WHO 2017). One reason many patients struggle to receive evaluation is the lack of effective assistive mental health analysis methods. Current diagnostic methods rely heavily on self-reports and clinical interviews (Habtamu et al. 2023), which are subjective and inefficient. The shortage of psychiatrists cannot keep up with increasing patients (Johnson 2022), highlighting the urgent need for efficient and scalable assistive mental health analysis.

\*Corresponding authors.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

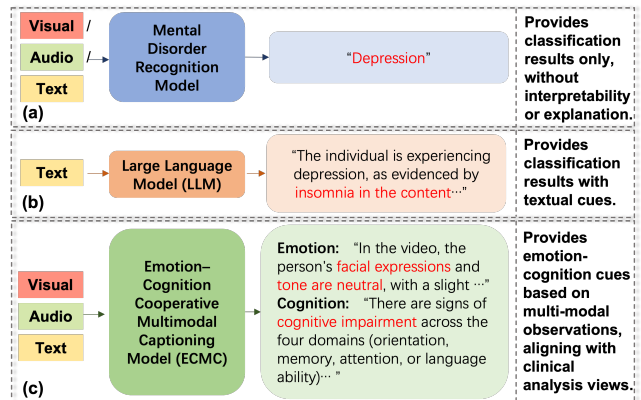


Figure 1: Different paradigms for mental disorder detection. (a) Classification-based. (b) LLM-based. (c) Ours.

However, existing assistive mental health analysis methods face challenges. Firstly, most methods (Liu et al. 2024; Ji et al. 2021; Zhang et al. 2022) treat the task as a classification problem by categorizing multi-modal data to a mental disorder category, which lacks interpretability (Figure 1(a)). They do not reveal which cues are associated with mental health conditions. As a result, these methods are often unconvincing for clinical diagnostic support. Secondly, large language models (LLMs) have strong natural language understanding abilities (Guo et al. 2025; Achiam et al. 2023). However, when applied to mental health analysis, they rely on textual information, such as detecting symptom-related words (Yang et al. 2024; Xu et al. 2024a) (Figure 1(b)). They cannot capture emotional and cognitive signals from other modalities, like facial expressions or vocal tone, which are important in real-world clinical observations. Despite the above two categories of methods contain works (Wu, Zhang, and Woodland 2023; Wu, Wu, and Yu 2022) that introduce affective features using transfer learning, the high-dimensional features offer limited insight into the correlation between emotion and mental disorders. They also ignore the role of cognitive factors in mental health analysis.

Emotion-cognition patterns play a vital role in understanding mental health disorders. Neuroscience research (Shackman, Fox, and Seminowicz 2015) suggests mental

disorders are not only reflected in observable symptoms, but also in the dynamic mechanisms of emotional and cognitive processing. For example, depression is often associated with a prolonged emotional stagnation and suppressed cognitive functioning, while anxiety disorders exhibit rapid emotional and cognitive fluctuations, often manifesting as panic attacks. Prior studies (Uban, Chulvi, and Rosso 2021; Li et al. 2022) have also identified significant differences in emotional and cognitive characteristics between individuals with and without mental health conditions based on linguistic features. However, existing methods rarely consider a multi-modal perspective to examine these patterns, and often fail to provide interpretable insights for clinical diagnosis.

To address these limitations, we introduce an emotion–cognition cooperative multi-modal captioning (ECMC) task, which aims to generate interpretable emotion–cognition profiles for mental health understanding (Figure 1(c)). Inspired by recent work on emotion captioning task (Xu et al. 2024b; Ye et al. 2024), ECMC requires models to extract emotion- and cognition-relevant features from diverse modalities, and generate coherent, high-quality natural language descriptions. Unlike traditional approaches that compress complex signals to categorical labels, ECMC leverages natural language to unify multi-modal information into rich, contextualized narratives that capture both emotional states and cognitive impairments. This not only enhances interpretability for clinicians but also supports fine-grained comparisons across individuals and time, making ECMC a promising direction for accurate and explainable mental health assessment.

To this end, we design a multi-modal encoder–decoder framework, which comprises modality-specific encoders, a dual-stream BridgeNet, and a LLaMA-based decoder. Given the challenge of training encoders from scratch with limited data, we initialize the modality-specific encoders using pre-trained models tailored to each modality to extract initial features. However, such pre-trained models typically focus on frame-level representations and fail to capture emotion- and cognition-relevant semantics. Inspired by BLIP-2 (Li et al. 2023), we design a dual-stream BridgeNet based on Q-former, which compresses multi-modal features and disentangles emotion and cognition representations with contrastive learning. Finally, we employ a LLaMA-based decoder, leveraging the powerful language understanding abilities of LLM, to generate emotion–cognition captions from the encoded features. The decoder guides the BridgeNet in the training process, to align emotion–cognition features with the target textual space.

The contributions of our work are summarized as follows:

- We propose a novel emotion–cognition cooperative multi-modal captioning (ECMC) task, which enables LLMs to generate emotion–cognition profiles for mental health understanding. To the best of our knowledge, this perspective has not been explored in previous work.
- We design a multi-modal encoder–decoder framework for the ECMC<sup>1</sup> task, which extracts emotion and cogni-

tion features from multi-modal data and aligns them with the LLM input space to produce informative captions.

- Extensive objective and subjective evaluations demonstrate that our method can generate high-quality emotion–cognition captions from multi-modal observations.

## Related Work

### Mental Disorders Detection

Most mental disorder detection methods focus on designing various deep learning models to classify samples into disorder categories (Liu et al. 2024; Ji et al. 2021; Zhang et al. 2022). However, the underlying reasons behind these predictions are often implicit, limiting their utility in clinical applications. The emergence of LLMs enhances mental disorders detection by fine-tuning on relevant data (Yang et al. 2024; Xu et al. 2024a). Nevertheless, these approaches typically rely solely on textual semantic information and cannot capture facial expressions or other observable reactions from patients. Although prior works (Uban, Chulvi, and Rosso 2021; Li et al. 2022) have incorporated affective information via transfer learning, the relationship between emotion-related features and prediction outcomes remains difficult to interpret. In contrast, our approach enables multi-modal observation of patients’ emotion–cognition patterns and generates descriptions of salient cues, thus providing more interpretable evidence to support effective clinical diagnosis.

### Multi-modal Large Language Model

Recent advances in multi-modal large language models (MLLMs) have achieved remarkable performance in cross-modal understanding and generation tasks. For example, BLIP-2 (Li et al. 2023) introduces a lightweight query-based transformer that effectively bridges frozen vision encoders and language models, enabling strong zero-shot performance. Qwen2.5-Omni (Xu et al. 2025) supports seamless multi-turn dialogue grounded in visual content, highlighting its strength in multi-modal scenarios. InternVL-2.5 (Chen et al. 2024) unifies visual and textual representations through dense alignment, enhancing complex cross-modal reasoning and fine-grained understanding. Sa2VA (Yuan et al. 2025) focuses on fine-grained semantic alignment, leading to improvements in visual question answering and detailed image captioning. Overall, MLLMs demonstrate impressive generalization across diverse modalities, laying a foundation for downstream multi-modal applications.

### Automatic Captioning

Automatic captioning methods typically aim to generate natural language descriptions from visual or multi-modal inputs by leveraging encoder–decoder architectures. For example, generating emotion–cause–aware video captions helps capture fine-grained multi-modal cues, which can then be used to summarize the causes of emotions in conversations (Wang et al. 2024). SECap (Xu et al. 2024b) uses HuBERT and Q-Former to extract emotion-focused speech features and LLaMA to generate natural language captions of speech emotions. By providing fine-grained and semantically rich

<sup>1</sup>Code and more results: <https://github.com/zhouzyhfut/ECMC>.

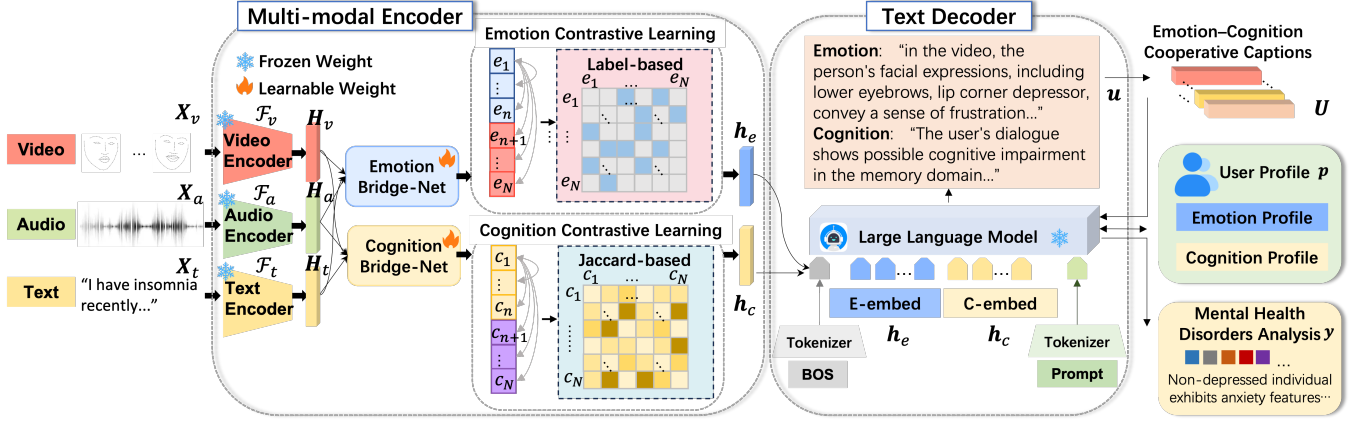


Figure 2: The framework of the proposed method.

textual descriptions, captioning tasks naturally offer detailed cues that can be highly valuable for mental health understanding, enabling interpretable insights into patients’ emotional and cognitive states.

## Method

We first provide an overview of our ECMC. Then, we describe in detail how the emotion bridge-net and cognition bridge-net extract emotional and cognitive features, respectively. The overall training process is finally described.

### Model Architecture

As illustrated in Figure 2, inspired by automatic captioning in emotion recognition (Wang et al. 2024; Xu et al. 2024b; Ye et al. 2024), we adopt an encoder–decoder architecture.

**Encoder** For a given utterance sample  $x_i = \{X_v, X_a, X_t\}$  where the subscripts  $v$ ,  $a$ , and  $t$  denote the video, audio, and text modalities respectively, we first extract modality-specific initial representations with modal-specific pre-trained models:

$$H_v = \mathcal{F}_v(X_v), \quad H_a = \mathcal{F}_a(X_a), \quad H_t = \mathcal{F}_t(X_t). \quad (1)$$

Here, we employ VideoMAE (Tong et al. 2022), HuBERT (van Niekerk et al. 2022), and BERT (Devlin et al. 2019) as  $\mathcal{F}_v$ ,  $\mathcal{F}_a$ , and  $\mathcal{F}_t$ , respectively.  $H_m \in \mathbb{R}^{T_m \times d_m}$ ,  $m \in \{v, a, t\}$  represents initial representation of each modality, where  $T_m$  is the number of time steps and  $d_m$  is the dimension of representation.

Subsequently, we design a dual-stream BridgeNet comprising the emotion bridge-net and the cognition bridge-net, which fuses multi-modal information and further compresses the initial representations, since the sequential representations are often redundant. This process yields two embeddings, E-embedding and C-embedding, denoted as:

$$\{h_e, h_c\} = \mathcal{F}_{bridge}(H_v, H_a, H_t) \quad (2)$$

Here,  $h_e \in \mathbb{R}^{1 \times d_e}$  captures features relevant to various emotional categories, while  $h_c \in \mathbb{R}^{1 \times d_c}$  encodes features related to potential cognitive impairments.

**Decoder** We then employ LLaMA (Touvron et al. 2023) as a text decoder to generate fine-grained emotion–cognition descriptions. Leveraging the powerful language understanding and generation capabilities of LLMs, we guide the BridgeNet toward a semantically rich textual feature space and produce high-quality captions. Specifically, we prepend a  $\langle \text{BOS} \rangle$  token to the input, and concatenate it with the E-embedding  $h_e$  and C-embedding  $h_c$ , followed by the tokenized prompt. The input to LLM is thus formulated as:

$$u = \mathcal{F}_{llm}(\text{concat}(\langle \text{BOS} \rangle, h_e, h_c, \text{Prompt})) \quad (3)$$

which constrains the output space of LLM to generate accurate captions  $u$  that reflect both emotional and cognitive aspects of the utterance.

After obtaining the emotion–cognition descriptions  $U = \{u_1, \dots, u_n\}$  for multiple utterances, we instruct LLM to produce a user profile  $p$ . This profile summarizes salient patterns such as negative emotions and cognitive impairments observed in the utterances:

$$p = \mathcal{F}_{llm}(U, \text{Prompt}) \quad (4)$$

The user profile  $p$  provides valuable evidence for mental disorder diagnosis, and further assists a general-purpose LLM in enhancing its mental disorder detection capabilities:

$$y = \mathcal{F}_{llm}(p, \text{Prompt}, X_t) \quad (5)$$

where  $y$  is the prediction of mental disorder.

### Emotional Representation Extraction

To fuse multi-modal features and further compress the initial representations for extracting an emotional representation to feed into the decoder, we design the Emotion BridgeNet, as illustrated in Figure 3.

**Q-former** For each modality  $m$ , we introduce a set of learnable query tokens  $Q_m \in \mathbb{R}^{l_q \times d_q}$ , where  $l_q$  is the number of query tokens and  $d_q$  is the hidden dimension. We first apply self-attention over the query tokens to model intra-query dependencies:

$$Q'_m = \text{softmax} \left( \frac{Q_m W_q (Q_m W_k)^T}{\sqrt{d_k}} \right) Q_m W_v \quad (6)$$

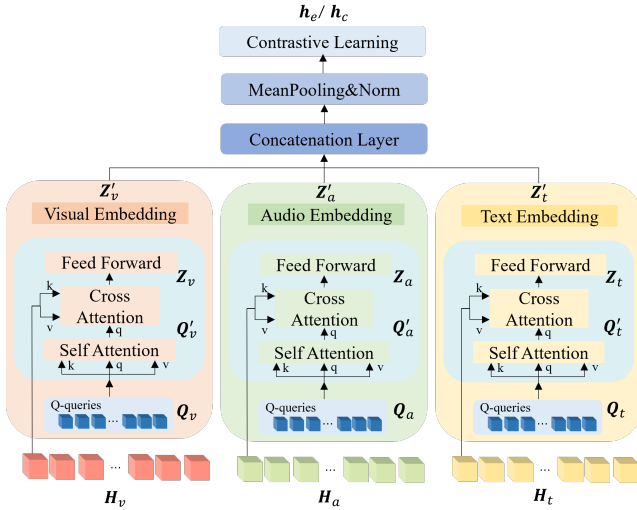


Figure 3: The architecture of BridgeNet.

where  $\mathbf{W}_q \in \mathbb{R}^{d_q \times d_k}$ ,  $\mathbf{W}_k \in \mathbb{R}^{d_q \times d_k}$ ,  $\mathbf{W}_v \in \mathbb{R}^{d_q \times d_v}$  are the learnable weight matrices for queries, keys, and values. Next, the updated queries  $\mathbf{Q}'_m \in \mathbb{R}^{l_q \times d_v}$  used as queries for the cross-attention:

$$\mathbf{Z}_m = \text{softmax} \left( \frac{\mathbf{Q}'_m \mathbf{W}_q^{(c)} (\mathbf{H}'_m \mathbf{W}_k^{(c)})^\top}{\sqrt{d_k}} \right) \mathbf{H}'_m \mathbf{W}_v^{(c)} \quad (7)$$

where  $\mathbf{H}'_m = \mathbf{H}_m \mathbf{W}_m$ ,  $\mathbf{W}_q^{(c)} \in \mathbb{R}^{d_v \times d_k}$ ,  $\mathbf{W}_k^{(c)} \in \mathbb{R}^{d_q \times d_k}$ ,  $\mathbf{W}_v^{(c)} \in \mathbb{R}^{d_q \times d_v}$ ,  $\mathbf{W}_m \in \mathbb{R}^{d_m \times d_q}$  are the learnable weight matrices. Subsequently, the attended query representation  $\mathbf{Z}_m \in \mathbb{R}^{l_q \times d_v}$  is passed through a feed-forward layer:

$$\mathbf{Z}'_m = \text{ReLU}(\mathbf{Z}_m \mathbf{W}_1 + \mathbf{b}_1) \mathbf{W}_2 + \mathbf{b}_2 \quad (8)$$

where  $\mathbf{W}_1 \in \mathbb{R}^{d_v \times d_e}$ ,  $\mathbf{W}_2 \in \mathbb{R}^{d_e \times \frac{d_e}{3}}$ , and  $\mathbf{b}_1, \mathbf{b}_2$  are biases.

We concatenate the representations  $\mathbf{Z}'_m$  and project them into the feature space:

$$\mathbf{h}_e = \text{Norm}(\text{Pooling}(\text{Concat}(\mathbf{Z}'_v, \mathbf{Z}'_a, \mathbf{Z}'_t) \mathbf{W}_3 + \mathbf{b}_3)), \quad (9)$$

where  $\mathbf{h}_e \in \mathbb{R}^{l_d \times d_e}$ ,  $\mathbf{W}_3 \in \mathbb{R}^{d_e \times d_e}$  is the projection matrix and  $\mathbf{b}_3$  is the bias. Pooling indicates meanpooling operation, and Norm is L2 normalization.

**Emotion Contrastive Learning** For mental disorder analysis, psychological researches (Joormann and Stanton 2016; Vanderlind et al. 2020; Zakiei et al. 2014) have highlighted negative emotions as a significant characteristic. Accordingly, we divide the feature representations  $\mathbf{h}_{e,i}$  into negative, neutral, and positive categories according to the valence dimension (Kensinger 2004) in the affective space. To better capture the representations of different emotional categories, we introduce a label-matching based contrastive learning approach. We first compute the similarity matrix:

$\mathbf{S}_{ij} = \frac{\mathbf{h}_{e,i}^\top \mathbf{h}_{e,j}}{\tau}$ . Then we build a binary mask  $\mathbf{M}_{ij} = 1$  if  $y_{e,i} = y_{e,j}$  and  $i \neq j$ , and 0 otherwise, where  $y_e \in$

$\{-1, 0, 1\}$  denotes emotion class label, to identify intra-class positives. The loss combines intra-class compactness and inter-class separability:

$$\mathcal{L}_{\text{emo}} = -\frac{1}{N} \sum_{i=1}^N \frac{1}{|\{j : \mathbf{M}_{ij} = 1\}|} \sum_{j: \mathbf{M}_{ij}=1} \log p_{ij} + \frac{1}{N} \sum_{i=1}^N \log \left( 1 + \sum_{j: \mathbf{M}_{ij}=0} \exp(\mathbf{S}_{ij}) \right), \quad (10)$$

where  $\log p_{ij} = \mathbf{S}_{ij} - \log \sum_{k \neq i} \exp(\mathbf{S}_{ik})$ , and  $N$  represents the batch size.

### Cognitive Representation Extraction

Similar to the emotional representation extraction process, we employ a Q-former to compress and fuse the multi-modal initial features  $\mathbf{H}_m$ , resulting in the cognition representation  $\mathbf{h}_c$ . Guided by clinical cognitive scale MMSE (Arevalo-Rodriguez et al. 2015), we focus on four specific cognitive impairments: orientation deficit, attention deficit, memory deficit, and language disorder, and aim to discover representations corresponding to these distinct cognitive categories. Since a sample  $\mathbf{h}_{c,i}$  may be associated with multiple cognitive impairments simultaneously, we design a multi-label contrastive learning objective based on the Jaccard similarity to better capture the shared and unique aspects of these impairments. For cognition features  $\mathbf{h}_{c,i}$ , each sample may have multiple labels  $\mathbf{y}_{c,i} \in \{0, 1\}^4$ . We replace the binary mask with a soft similarity weight based on Jaccard index:

$$\mathbf{W}_{ij} = \begin{cases} 1, & \mathbf{y}_{c,i} = \mathbf{y}_{c,j} = \emptyset \\ 0, & \mathbf{y}_{c,i} = \emptyset \text{ or } \mathbf{y}_{c,j} = \emptyset \\ \frac{|\mathbf{y}_{c,i} \cap \mathbf{y}_{c,j}|}{|\mathbf{y}_{c,i} \cup \mathbf{y}_{c,j}|}, & \text{otherwise.} \end{cases} \quad (11)$$

The loss softly pulls samples with higher label overlap closer and pushes others apart:

$$\mathcal{L}_{\text{cog}} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\sum_j \exp(\mathbf{S}_{ij}) \cdot \mathbf{W}_{ij}}{\sum_j \exp(\mathbf{S}_{ij})} + \frac{1}{N} \sum_{i=1}^N \log \left( 1 + \sum_j \exp(\mathbf{S}_{ij}) (1 - \mathbf{W}_{ij}) \right). \quad (12)$$

### Training Process

To enhance multi-modal emotion-cognition captions with LLMs, we design a two-stage training process. In the first stage, we jointly train the emotional representation extraction and cognitive representation extraction processes. This collaborative training effectively compresses and fuses multi-modal information, enabling the extraction of compact emotional and cognitive representations:

$$\mathcal{L}_1 = \mathcal{L}_{\text{emo}} + \mathcal{L}_{\text{cog}}, \quad (13)$$

where we keep the parameters of the modal-specific encoder frozen, and the Q-Former is initialized with BERT (Devlin et al. 2019) pre-trained parameters.

Method	Modality	Domain	BLEU-1	BLEU-2	BLEU-4	METEOR	ROUGE-L	CIDEr	F_BERT
InternVL-2.5-8B	VT	General	8.35	3.30	0.96	10.06	13.98	0.96	6.48
DeepSeek-VL-2-small	VT		2.98	1.33	0.42	5.99	10.00	0.42	2.02
Sa2VA-8B	VT		14.36	7.05	2.14	<u>15.35</u>	<u>20.34</u>	2.14	<u>12.28</u>
Qwen2.5-Omni-7B	AVT		12.74	4.86	1.42	<u>13.31</u>	<u>16.99</u>	1.42	8.93
PsycoLLM	T	Mental	<u>4.86</u>	<u>1.76</u>	<u>0.50</u>	<u>5.68</u>	<u>7.68</u>	<u>0.50</u>	<u>0.35</u>
MindChat	T		5.44	1.75	0.50	6.73	9.71	0.50	0.87
EmoLLM	T		8.50	3.50	1.19	10.08	14.25	1.19	6.25
CPsyCoun	T		<u>17.44</u>	<u>7.24</u>	<u>2.33</u>	15.07	18.90	<u>2.33</u>	9.30
<b>Ours</b>	AVT		<b><u>34.76</u></b>	<b><u>19.40</u></b>	<b><u>8.28</u></b>	<b><u>29.47</u></b>	<b><u>24.91</u></b>	<b><u>8.28</u></b>	<b><u>27.13</u></b>

Table 1: Comparison of emotion captioning performance with other methods. A,V,T denote audio, video, and text, respectively. Bold indicates the best, while underline indicates the second-best.

Method	Modality	Domain	BLEU-1	BLEU-2	BLEU-4	METEOR	ROUGE-L	CIDEr	F_BERT
InternVL-2.5-8B	VT	General	8.17	3.05	0.84	9.35	16.52	0.84	18.43
DeepSeek-VL-2-small	VT		8.49	4.40	1.72	10.34	15.45	1.72	6.75
Sa2VA-8B	VT		15.82	6.24	1.61	13.15	22.59	1.61	18.78
Qwen2.5-Omni-7B	AVT		12.82	5.30	1.48	12.04	19.56	1.48	<u>20.03</u>
PsycoLLM	T	Mental	6.22	3.19	1.30	10.59	10.48	1.30	<u>6.60</u>
MindChat	T		6.81	3.02	1.11	8.33	11.74	1.11	5.41
EmoLLM	T		<u>17.00</u>	<u>6.66</u>	<u>1.98</u>	13.29	19.08	<u>1.98</u>	14.99
CPsyCoun	T		9.32	3.89	1.06	<u>18.88</u>	<u>23.07</u>	1.06	4.47
<b>Ours</b>	AVT		<b><u>35.92</u></b>	<b><u>25.05</u></b>	<b><u>15.32</u></b>	<b><u>35.86</u></b>	<b><u>39.82</u></b>	<b><u>15.32</u></b>	<b><u>41.04</u></b>

Table 2: Comparison of cognition captioning performance with other methods

In the second stage, we fine-tune the emotion Bridge-Net and cognition Bridge-Net to effectively integrate the extracted features into the LLM. Specifically, we keep both the modal-specific encoders and LLM parameters frozen, and employ cross-entropy loss to make the LLM generate annotated captions  $\hat{u}$ , thereby guiding the Bridge-Net to produce features aligned with the LLM’s input space:

$$\mathcal{L}_2 = CELoss(\hat{u}, u) \quad (14)$$

### Dataset and Evaluation Metric

**Dataset** We evaluate our method on the MMDA dataset (Jiang, Zhang, and Sun 2022), a large-scale multimodal benchmark for mental disorder analysis containing 1,025 clinically validated samples. Each sample includes visual, acoustic, and textual data from clinical interviews, with depression and anxiety labels based on standardized assessments (HAMD and HAMA). To obtain utterance-level emotion–cognition captions, we combine automatic annotation with manual refinement. The interviews are first segmented into question–answer turns. Emotion-LLaMA (Cheng et al. 2024) then generates emotional descriptions integrating facial, vocal, and semantic cues. Using these outputs and the textual content, DeepSeek-671B (Guo et al. 2025) produces corresponding cognitive captions and categorical labels for both emotion and cognition, yielding 30,592 annotated utterance–caption pairs. To ensure data quality, non-informative segments (e.g., greetings, personal information inquiries, side conversations, or utterances shorter than 2 seconds) were removed to retain core doctor–patient interactions. Annotators further discarded inconsistent labels to

minimize noise and bias. We computed inter-rater agreement, yielding substantial consistency (Fleiss’s  $\kappa = 0.64$  for emotion and 0.73 for cognition). After filtering and verification, the dataset comprises 13,536 segments for training, 1,402 for validation, and 3,790 for testing. The dataset’s category distribution is 0.91% orientation deficit, 8.54% attention deficit, 10.38% memory deficit, 15.89% language disorder, and 31.85% negative emotion.

**Objective metrics** For the ECMC task, we adopt the natural language evaluation metrics as used in automatic captioning, including BLEU-1, BLEU-2, BLEU-4, METEOR, ROUGE-L, CIDEr, and F\_BERT. Among these, BLEU, METEOR, ROUGE-L, and CIDEr are primarily word-level matching metrics, while F\_BERT is a semantic similarity-based metric. For evaluating the generated emotion-cognition profiles on mental disorder detection, we use accuracy (ACC) and F1 score (F1).

**Subjective Metrics** To evaluate both the quality of our dataset annotations and the effectiveness of captions generated by different methods, we conduct a human evaluation using a five-point Likert scale (Joshi et al. 2015) across the following designed dimensions: (1) Emotional Accuracy (EAcc), measuring whether the generated emotional captions accurately reflect the content of the video; (2) Cognitive Accuracy (CAcc), measuring whether the captions of cognitive impairments correctly correspond to the dialogue content; (3) Detail Richness (Detail), evaluating the richness and granularity of details in the captions; (4) Label Consistency (Consis), assessing whether the captions align with the corresponding classification labels; and (5) Overall Quality

Method	ACC	F1	Method	ACC	F1	Method	ACC	F1	Im_ACC	Im_F1
Llama-3-8B	44.95	46.43	Qwen3-8B	50.46	60.29	GLM4-9B	47.71	58.39	0.00	0.00
InternVL-2.5-8B	48.62	46.15	InternVL-2.5-8B	69.72	69.81	InternVL-2.5-8B	53.21	59.84	9.48↑	3.56↑
DeepSeek-VL-2-small	41.28	47.54	DeepSeek-VL-2-small	51.38	59.54	DeepSeek-VL-2-small	51.38	60.15	0.31↑	0.71↑
Sa2VA-8B	48.62	49.09	Sa2VA-8B	67.89	69.03	Sa2VA-8B	57.80	62.90	10.40↑	5.30↑
Qwen2.5-Omni-7B	49.54	54.24	Qwen2.5-Omni-7B	65.14	67.24	Qwen2.5-Omni-7B	56.88	58.46	9.48↑	4.94↑
PsycoLLM	47.71	44.66	PsycoLLM	54.13	62.12	PsycoLLM	53.21	61.07	3.98↑	0.91↑
MindChat	29.36	36.36	MindChat	58.72	62.81	MindChat	57.80	62.90	0.92↑	1.01↓
EmoLLM	45.87	54.96	EmoLLM	61.47	63.16	EmoLLM	55.05	63.16	6.42↑	5.39↑
CPsyCounX	48.62	48.15	CPsyCounX	67.89	68.47	CPsyCounX	55.05	60.80	9.48↑	4.10↑
<b>Ours</b>	<b>50.46</b>	<b>56.69</b>	<b>Ours</b>	<b>70.64</b>	<b>70.27</b>	<b>Ours</b>	<b>59.63</b>	<b>65.62</b>	<b>12.54↑</b>	<b>9.16↑</b>

Table 3: Depression Detection by Baseline LLMs Using Emotion–Cognition Profiles from Different Methods. Row 1: results of baseline models. 'Im\_ACC' and 'Im\_F1' show average gains over Llama-3-8B, Qwen3-8B, and GLM4-9B in ACC and F1.

Method	ACC	F1	Method	ACC	F1	Method	ACC	F1	Im_ACC	Im_F1
Llama-3-8B	38.54	52.80	Qwen3-8B	41.67	57.58	GLM4-9B	43.75	57.14	0.00	0.00
InternVL-2.5-8B	56.25	56.25	InternVL-2.5-8B	55.21	63.87	InternVL-2.5-8B	42.71	58.02	10.07↑	3.54↑
DeepSeek-VL-2-small	48.96	55.05	DeepSeek-VL-2-small	45.83	59.38	DeepSeek-VL-2-small	41.67	57.58	4.17↑	1.50↑
Sa2VA-8B	55.21	52.75	Sa2VA-8B	54.17	62.71	Sa2VA-8B	44.79	58.46	10.07↑	2.13↑
Qwen2.5-Omni-7B	50.00	50.00	Qwen2.5-Omni-7B	65.62	65.96	Qwen2.5-Omni-7B	43.75	58.46	11.80↑	2.30↑
PsycoLLM	52.08	57.41	PsycoLLM	48.96	59.50	PsycoLLM	43.75	58.46	6.94↑	2.62↑
MindChat	47.92	46.81	MindChat	46.88	59.84	MindChat	41.67	57.58	4.17↑	1.10↓
EmoLLM	54.17	55.10	EmoLLM	61.46	66.06	EmoLLM	43.75	58.46	11.81↑	4.03↑
CPsyCounX	46.88	54.05	CPsyCounX	56.25	63.16	CPsyCounX	42.71	58.02	7.29↑	2.57↑
<b>Ours</b>	<b>57.29</b>	<b>58.59</b>	<b>Ours</b>	<b>66.67</b>	<b>69.16</b>	<b>Ours</b>	<b>44.95</b>	<b>58.91</b>	<b>14.98↑</b>	<b>6.38↑</b>

Table 4: Anxiety Detection by Baseline LLMs Using Emotion–Cognition Profiles from Different Methods

(Overall), providing an overall assessment of the quality.

## Experiments and Analysis

### Experiment Setup

For the first stage, we train our model on two NVIDIA RTX A6000 GPUs with a batch size of 64 for up to 500 epochs. The maximum sequence lengths of 1024 for audio and 512 for video. We use the AdamW optimizer with a learning rate of  $1.3 \times 10^{-5}$ , and a weight decay of  $1.0 \times 10^{-6}$ . The model has approximately 701M total parameters, with about 598M parameters being trainable. For the second stage, we train our model with a micro batch size of 8 per GPU and accumulation over 8 steps. We also apply DeepSpeed ZeRO stage-2 optimization and set mixed-precision training (fp16). In this stage, the model has approximately 7.6B total parameters, with about 605M trainable parameters.

### Performance Analysis

We evaluate our method through caption comparison with baselines, mental disorder detection performance, ablations on modalities and contrastive modules, statistical analysis across disorders, and human evaluation.

**Comparison of Emotion–Cognition Captions** We select general multi-modal large language models (LLMs) (InternVL-2.5 (Chen et al. 2024), DeepSeek-VL-2 (Lu et al. 2024), Sa2VA (Yuan et al. 2025), Qwen2.5-Omni (Xu et al.

2025)), as well as large language models (LLMs) specifically fine-tuned for the mental health domain (PsycoLLM (Hu et al. 2024), MindChat<sup>2</sup>, EmoLLM<sup>3</sup>, CPsyCoun (Zhang et al. 2024)), for comparison. Tables 1 and 2 present the ECMC performance of different models. We can see that our model consistently outperforms other competitors in both emotion and cognition captioning. General-domain MLLMs like Sa2VA and Qwen2.5-Omni perform relatively well due to better instruction-following and broader data coverage, but often produce short or inconsistent outputs. Mental health LLMs, though domain-specific, are limited to text and struggle with describing non-verbal cues, leading to lower scores. In contrast, our method effectively integrates multi-modal data, including facial expressions, vocal tone, and text semantics, to generate more comprehensive and higher-quality emotion-cognition captions.

**Effect on Mental Disorder Detection** We evaluate the impact of emotion–cognition profiles generated by different methods on depression and anxiety detection using existing baseline LLMs, including LLaMA-3-8B (Dubey et al. 2024), Qwen3-8B (Yang et al. 2025), and GLM4-9B (Zeng et al. 2024). For each subject, we first summarize the captions generated by different methods (Table 1 and 2) into emotion–cognition profiles using DeepSeek-671B. These profiles, together with the original dialogue content, are then

<sup>2</sup><https://github.com/X-D-Lab/MindChat>

<sup>3</sup><https://github.com/SmartFlowAI/EmoLLM>

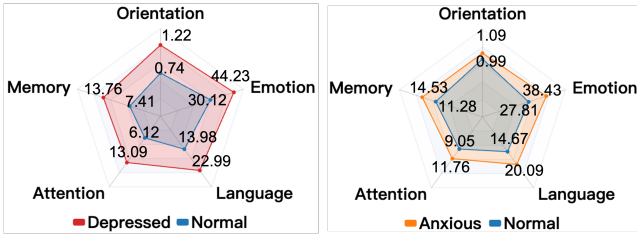


Figure 4: Emotional and cognitive differences between normal individuals and patients with different mental disorders

Audio	Video	Text	EmoCL	CogCL	F.BERT
✓			✓	✓	18.10
	✓		✓	✓	16.30
		✓	✓	✓	16.75
✓	✓		✓	✓	22.05
✓		✓	✓	✓	29.30
	✓	✓	✓	✓	27.21
✓	✓	✓		✓	23.89
✓	✓	✓		✓	27.24
✓	✓	✓	✓		26.94
✓	✓	✓	✓	✓	34.09

Table 5: Ablation Study on Different Modules. EmoCL and CogCL represent emotion contrastive learning and cognition contrastive learning, respectively.

provided as input to the baseline LLMs. As shown in Table 3 and 4, the profiles generated by our method yield higher accuracy and F1 scores across both disorder types, significantly improving the performance of different baselines, with average ACC and F1 improvements over 12.54% and 9.16% on depression detection, and 14.98% and 6.38% on anxiety detection. Notably, we also observe that ineffective captions can even degrade detection performance, as long but uninformative texts may increase the difficulty for LLMs to extract relevant indicators. The results validate that our method produces higher-quality emotion-cognition profiles, which effectively enhance the accuracy of mental disorder detection using existing general models.

**Ablation Study** We conduct ablation studies to verify the contribution of different modalities and contrastive learning modules to the final performance. We use the semantic similarity metric F.BERT to evaluate the overall effectiveness of the emotion-cognition captions. As shown in Table 5, among uni-modal inputs, audio achieves the best performance, indicating its discriminative power. Multi-modal combinations further improve results, with audio+text showing strong complementarity. Removing either EmoCL or CogCL causes the model to degenerate into a single-branch structure, resulting in performance drops, while removing both leads to a significant decline. This demonstrates the importance of the dual-branch structure for effectively decoupling emotion and cognition features. The full model with all modalities and contrastive objectives achieves the best score, confirming the effectiveness of each component.

Method/Dataset	EAcc	CACC	Detail	Consis	Overall
Dataset	4.2	3.9	4.6	4.3	4.3
Sa2VA	3.76	4.40	<b>4.36</b>	-	4.15
Qwen2.5-Omni	3.73	4.45	4.32	-	4.12
EmoLLM	3.6	3.91	3.96	-	3.79
CPsyCoun	3.2	3.6	3.68	-	3.46
<b>Ours</b>	<b>3.98</b>	<b>4.51</b>	<b>4.32</b>	-	<b>4.26</b>

Table 6: Human Evaluation on Dataset and Methods

**Emotional and Cognitive Statistics across Different Mental Disorders** To explore the differences in emotion-cognition patterns across mental disorders and healthy individuals, we calculate the proportion of negative emotions, orientation deficit, attention deficit, memory deficit, and language disorder in utterances of each individual, and then compute the average within each group (depression, anxiety, and healthy controls). As shown in Figure 4, both depressed and anxious individuals exhibit significantly higher proportions of negative emotions compared to healthy participants. In terms of cognitive impairments, depression is associated with notably higher frequencies than anxiety. Among the cognitive categories, orientation deficits appear less frequently, while attention deficits, memory deficits, and language disorders are more prevalent. The above observations validate the necessity of the ECMC task.

**Human Evaluation** We invited eight psychology practitioners with expertise in mental health to evaluate our dataset and model outputs. As shown in Table 6, the dataset receives an overall score of 4.3, indicating high-quality emotion-cognition captions. EAcc, Detail, and Consis all score above 4.0, suggesting accurate emotional content, rich detail, and reliable labels for contrastive learning. CACC scores slightly lower, reflecting the difficulty of generating cognitively accurate descriptions from multi-modal and textual inputs, and highlighting the need for better multi-modal cognitive modeling. We compare our method with two top-performing general MLLMs and two leading mental health LLMs. Our model achieves the highest scores in EAcc, CACC, and Overall, and ranks second in Detail. Notably, MLLMs receive higher subjective ratings than mental health LLMs, empirically validating the effectiveness of combining visual and acoustic cues, which mental health LLMs lack.

## Conclusion

In this paper, we introduce the ECMC task to generate interpretable emotion-cognition profiles from multi-modal data for mental health understanding. To support this task, we design a novel encoder-decoder framework that extracts and aligns emotion-cognition features with a LLaMA-based decoder. Experimental results show that our method produces high-quality, explainable captions, offering a promising direction for interpretable and fine-grained mental health assessment.

## Acknowledgments

This work was supported by the National Key Research and Development Program (Grant No. 2023YFC2506800), the National Natural Science Foundation of China under Grant No. 62072152, 62172137.

## References

- Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Arevalo-Rodriguez, I.; Smailagic, N.; i Figuls, M. R.; Ciapponi, A.; Sanchez-Perez, E.; Giannakou, A.; Pedraza, O. L.; Cosp, X. B.; and Cullum, S. 2015. Mini-Mental State Examination (MMSE) for the detection of Alzheimer’s disease and other dementias in people with mild cognitive impairment and other dementias in people with mild cognitive impairment (MCI). *Cochrane Database of Systematic Reviews*, (3).
- Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24185–24198.
- Cheng, Z.; Cheng, Z.-Q.; He, J.-Y.; Wang, K.; Lin, Y.; Lian, Z.; Peng, X.; and Hauptmann, A. 2024. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems*, 37: 110805–110853.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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)*, 4171–4186.
- Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv e-prints*, arXiv-2407.
- Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Habtamu, K.; Birhane, R.; Demissie, M.; and Fekadu, A. 2023. Interventions to improve the detection of depression in primary healthcare: systematic review. *Systematic Reviews*, 12(1): 25.
- Hu, J.; Dong, T.; Gang, L.; Ma, H.; Zou, P.; Sun, X.; Guo, D.; Yang, X.; and Wang, M. 2024. Psycollm: Enhancing llm for psychological understanding and evaluation. *IEEE Transactions on Computational Social Systems*.
- Ji, S.; Zhang, T.; Ansari, L.; Fu, J.; Tiwari, P.; and Cambria, E. 2021. Mentalbert: Publicly available pretrained language models for mental healthcare. *arXiv preprint arXiv:2110.15621*.
- Jiang, Y.; Zhang, Z.; and Sun, X. 2022. MMDA: A Multimodal Dataset for Depression and Anxiety Detection. In *Proceedings of the International Conference on Pattern Recognition*, 691–702. Springer.
- Johnson, R. M. 2022. Addressing shortages in the psychiatry workforce. *Psychiatric Times*.
- Jormann, J.; and Stanton, C. H. 2016. Examining emotion regulation in depression: A review and future directions. *Behaviour Research and Therapy*, 86: 35–49.
- Joshi, A.; Kale, S.; Chandel, S.; and Pal, D. K. 2015. Likert scale: Explored and explained. *British Journal of Applied Science & Technology*, 7(4): 396.
- Kensinger, E. A. 2004. Remembering emotional experiences: The contribution of valence and arousal. *Reviews in the Neurosciences*, 15(4): 241–252.
- Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on Machine Learning, 19730–19742*. PMLR.
- Li, M.; Zhang, J.; Song, J.; Li, Z.; and Lu, S. 2022. A clinical-oriented non-severe depression diagnosis method based on cognitive behavior of emotional conflict. *IEEE Transactions on Computational Social Systems*, 10(1): 131–141.
- Liu, H.; Li, C.; Zhang, X.; Zhang, F.; Wang, W.; Ma, F.; Chen, H.; Yu, H.; and Zhang, X. 2024. Depression detection via capsule networks with contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 22231–22239.
- Lu, H.; Liu, W.; Zhang, B.; Wang, B.; Dong, K.; Liu, B.; Sun, J.; Ren, T.; Li, Z.; Yang, H.; et al. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.
- Shackman, A. J.; Fox, A. S.; and Seminowicz, D. A. 2015. The cognitive-emotional brain: Opportunities and challenges for understanding neuropsychiatric disorders. *Behavioral and Brain Sciences*, 38: e86.
- Tong, Z.; Song, Y.; Wang, J.; and Wang, L. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in Neural Information Processing Systems*, 35: 10078–10093.
- Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Uban, A.-S.; Chulvi, B.; and Rosso, P. 2021. An emotion and cognitive based analysis of mental health disorders from social media data. *Future Generation Computer Systems*, 124: 480–494.
- van Niekerk, B.; Carbonneau, M.-A.; Zaïdi, J.; Baas, M.; Seuté, H.; and Kamper, H. 2022. A Comparison of Discrete and Soft Speech Units for Improved Voice Conversion. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*.

- Vanderlind, W. M.; Millgram, Y.; Baskin-Sommers, A. R.; Clark, M. S.; and Joormann, J. 2020. Understanding positive emotion deficits in depression: From emotion preferences to emotion regulation. *Clinical Psychology Review*, 76: 101826.
- Wang, F.; Ma, H.; Shen, X.; Yu, J.; and Xia, R. 2024. Observe before Generate: Emotion-Cause aware Video Caption for Multimodal Emotion Cause Generation in Conversations. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 5820–5828.
- WHO. 2017. Depression and other common mental disorders: global health estimates. Technical report, World Health Organization.
- Wu, W.; Wu, M.; and Yu, K. 2022. Climate and weather: Inspecting depression detection via emotion recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 6262–6266. IEEE.
- Wu, W.; Zhang, C.; and Woodland, P. C. 2023. Self-Supervised Representations in Speech-Based Depression Detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1–5.
- Xu, J.; Guo, Z.; He, J.; Hu, H.; He, T.; Bai, S.; Chen, K.; Wang, J.; Fan, Y.; Dang, K.; et al. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.
- Xu, X.; Yao, B.; Dong, Y.; Gabriel, S.; Yu, H.; Hendler, J.; Ghassemi, M.; Dey, A. K.; and Wang, D. 2024a. MentalLLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*.
- Xu, Y.; Chen, H.; Yu, J.; Huang, Q.; Wu, Z.; Zhang, S.-X.; Li, G.; Luo, Y.; and Gu, R. 2024b. Secap: Speech emotion captioning with large language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 19323–19331.
- Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Yang, K.; Zhang, T.; Kuang, Z.; Xie, Q.; Huang, J.; and Ananiadou, S. 2024. MentalLaMA: interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM Web Conference 2024*, 4489–4500.
- Ye, C.; Chen, W.; Li, J.; Zhang, L.; and Mao, Z. 2024. Dual-path Collaborative Generation Network for Emotional Video Captioning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 496–505.
- Yuan, H.; Li, X.; Zhang, T.; Huang, Z.; Xu, S.; Ji, S.; Tong, Y.; Qi, L.; Feng, J.; and Yang, M.-H. 2025. Sa2VA: Marrying SAM2 with LLaVA for Dense Grounded Understanding of Images and Videos. *arXiv preprint arXiv:2501.04001*.
- Zakiei, A.; Karami, J.; Alikhani, M.; Ghasemi, S.; and Gilan, N. R. 2014. The relationship between alexithymia, irrational beliefs, positive and negative emotions with mental disorders. *Depression and Anxiety*, 17: 18.
- Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Zhang, D.; Rojas, D.; Feng, G.; Zhao, H.; Lai, H.; Yu, H.; Wang, H.; Sun, J.; Zhang, J.; Cheng, J.; Gui, J.; Tang, J.; Zhang, J.; Sun, J.; Li, J.; Zhao, L.; Wu, L.; Zhong, L.; Liu, M.; Huang, M.; Zhang, P.; Zheng, Q.; Lu, R.; Duan, S.; Zhang, S.; Cao, S.; Yang, S.; Tam, W. L.; Zhao, W.; Liu, X.; Xia, X.; Zhang, X.; Gu, X.; Lv, X.; Liu, X.; Liu, X.; Yang, X.; Song, X.; Zhang, X.; An, Y.; Xu, Y.; Niu, Y.; Yang, Y.; Li, Y.; Bai, Y.; Dong, Y.; Qi, Z.; Wang, Z.; Yang, Z.; Du, Z.; Hou, Z.; and Wang, Z. 2024. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. *arXiv:2406.12793*.
- Zhang, C.; Li, R.; Tan, M.; Yang, M.; Zhu, J.; Yang, D.; Zhao, J.; Ye, G.; Li, C.; and Hu, X. 2024. Cpsycoun: A report-based multi-turn dialogue reconstruction and evaluation framework for chinese psychological counseling. *arXiv preprint arXiv:2405.16433*.
- Zhang, Z.; Chen, S.; Wu, M.; and Zhu, K. Q. 2022. Psychiatric scale guided risky post screening for early detection of depression. *arXiv preprint arXiv:2205.09497*.