

Technical Appendix

A. Loss Convergence Analysis



Figure 1: Training and validation loss of stage 1.



Figure 2: Training and validation loss of stage 2.

We provide the training and validation loss curves of both Stage 1 and Stage 2 to analyze the optimization dynamics and convergence behavior of our model.

As shown in Figure 1, during stage 1, the training loss steadily decreases throughout the training process, indicating that the model effectively learns emotion and cognition features from the initial representations extracted by the pre-trained backbone. However, the validation loss, after an initial decline, begins to increase gradually after approximately 20,000 steps, suggesting potential overfitting as training progresses. To address this, we select the model checkpoint corresponding to the lowest validation loss for stage 2 training process, as it is expected to yield the most generalizable feature representations.

Figure 2 illustrates the loss curves of stage 2, where the model is trained to align the multi-modal emotion–cognition features obtained from stage 1 with the textual feature space of the large language model (LLM), thereby enabling the generation of emotion–cognition captions. Both the training

and validation losses decrease steadily during the early training steps. Notably, the validation loss reaches its minimum around step 4000 and slightly increases thereafter, indicating the onset of overfitting. Compared to stage 1, stage 2 converges faster and more stably, suggesting that the representations learned in stage 1 provide a strong foundation for effective alignment with the LLM feature space.

These results highlight the effectiveness of our two-stage training paradigm: stage 1 enables the model to learn multi-modal emotion–cognition features, while stage 2 fine-tunes the model efficiently toward the captioning objective.

B. Feature Visualization with t-SNE

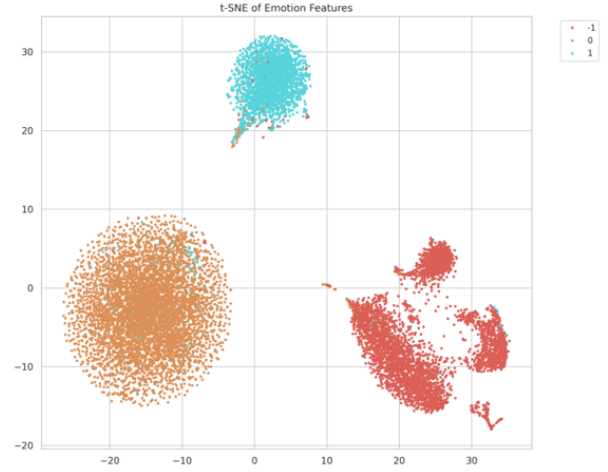


Figure 3: t-SNE visualization of the learned emotion feature representations.

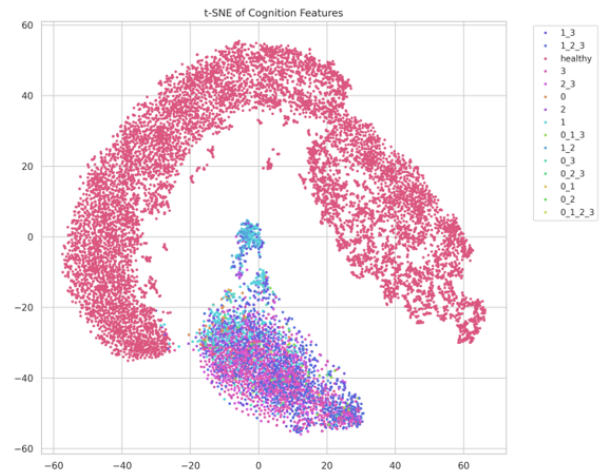


Figure 4: t-SNE visualization of the learned cognition feature representations.

To qualitatively assess the learned representations of emotion and cognition, we apply t-distributed Stochastic Neighbor Embedding (t-SNE) to project the high-dimensional fea-

tures into a two-dimensional space. We present the features of the model selected in Appendix A, as shown in Figure 3 and 4.

Figure 3 presents the t-SNE visualization of the learned emotion feature representations from the emotion bridge-net, where each point is color-coded based on the emotion label: -1 (negative), 0 (neutral), and 1 (positive). The visualization reveals three clearly separated clusters, indicating that the model is able to effectively distinguish between different emotional states.

Figure 4 shows the t-SNE visualization of the cognition feature representations, where each sample is labeled according to the combination of cognitive deficits it exhibits. The numbers 0, 1, 2, and 3 respectively denote the presence of orientation deficit, attention deficit, memory deficit, and language disorder. Due to the multi-label nature of this task, some samples may exhibit multiple cognitive impairments. The resulting visualization shows a more continuous and overlapping distribution compared to the emotion features, reflecting the higher complexity and co-occurrence patterns among cognitive deficits. Notably, we observe a meaningful grouping of samples with similar deficit profiles. In particular, cognitively healthy samples form a distinct cluster that is clearly separated from those with cognitive impairments. This separation highlights the model’s ability to capture structured and discriminative cognitive representations, effectively distinguishing between healthy individuals and those exhibiting one or more deficits.

These visualizations confirm that our model is capable of capturing meaningful and interpretable latent spaces for both emotional and cognitive aspects, supporting its effectiveness in multi-modal mental health analysis.

C. Case Analysis

Figures 5 and 6 present comparative generation results from five models: Sa2VA, Qwen2.5-Omni, EmoLLM, CPsyCoun, and our proposed method. Among them, Sa2VA and Qwen2.5-Omni are general-purpose Multi-modal LLMs (MLLMs), while EmoLLM and CPsyCoun are domain-specific LLMs fine-tuned for mental health analysis. For each model’s output, we annotated the details related to negative emotion, orientation deficit, attention deficit, memory deficit, and language disorder, using background colors that match the corresponding ground truth categories. For each category judgment, correct predictions are highlighted in green text, while incorrect ones are marked in red.

Specifically, Figure 5 shows a case where the patient explicitly expresses difficulties in memory and attention. Sa2VA provides detailed descriptions across all five categories but misclassifies attention deficit and memory deficit, leading to inaccurate supporting details. Qwen2.5-Omni correctly predicts all categories, but its descriptions are sparse, lacking nuanced observations such as subtle facial expressions. EmoLLM makes category-level predictions based on textual semantics and incorrectly infers the presence of an orientation deficit. CPsyCoun accurately identifies negative emotion, attention deficit, and memory deficit, but its generated captions lack grounding evidence. In contrast, our method offers fine-grained multi-modal analysis

such as facial micro-expressions and tone of voice, which supports predictions and correctly identifies cognitive impairments.

Figure 6 presents a more subtle case that involves negative emotion, orientation deficit, and language disorder. Compared to Figure 5, both orientation and language impairments are more implicit and thus harder to detect. Sa2VA again produces judgments and corresponding details for all categories, but it fails to identify orientation deficit and language disorder, and its descriptions rely solely on text without leveraging visual cues. Qwen2.5-Omni incorporates captions of prosody and vocal tone and accurately detects emotional state, but it makes incorrect predictions regarding cognitive impairments. EmoLLM misattributes the emotional state to weather conditions and concludes that the subject is cognitively healthy. CPsyCoun correctly detects the emotional state based on specific lexical cues but also concludes a healthy cognitive status. Our method again utilizes multi-modal cues such as facial expression and tone, providing accurate emotional classification and identifying cognitive impairments in orientation and language. It highlights the subject’s temporal disorientation and sentence-level abnormalities, such as syntactic irregularities and abnormal pauses.

In summary, MLLMs demonstrate a strength in generating detailed multi-modal captions, enabling them to support judgments on both emotion and cognition. However, when errors occur, especially in cognition, they may suffer from hallucinated explanations. Mental health-specific LLMs, on the other hand, rely heavily on semantic understanding and domain-specific knowledge, enabling them to better detect cognitive symptoms but often lacking multi-modal detail or emotional sensitivity. Our proposed method achieves more accurate detection across both emotion and cognition domains, offering grounded multi-modal evidence. However, it sometimes provides less structured captions compared to general-purpose MLLMs and tends to focus only on the impaired domains without providing a full cognitive profile. To address this, we plan to explore more comprehensive and structured representations of cognitive symptoms in future work, further bridging the gap between mental health reasoning and multi-modal understanding.

Content	Method	Emotion-Cognition Captions
<p>Utterance: Doctor: Do you feel that your memory has deteriorated in the past two months and you can't concentrate? Patient: Yes. I feel like I can't remember things clearly recently.</p> <p>Ground Truth: Negative emotion: Yes Orientation deficit: No Memory deficit: Yes Attention deficit: Yes Language disorder: No</p>	<p>Sa2VA</p> <p>Qwen2.5-Omni</p> <p>EmoLLM</p> <p>CPsyCoun</p> <p>Ours</p>	<p>Emotion: The person in the video appears to be concerned and possibly anxious. Their facial expressions may include furrowed brows, a slightly open mouth, and a tense posture. Orientation: The person seems to be aware of their current situation and can communicate their concerns effectively, indicating that their orientation is intact. Memory: The person is able to recall their recent experiences and express their concerns about their memory decline, suggesting that their memory function is preserved. Attention: The person is able to focus on the conversation and express their thoughts clearly, indicating that their attentional abilities are not significantly impaired. Language Ability: The person is able to communicate their concerns effectively, using appropriate grammar and vocabulary, suggesting that their language abilities are intact.</p> <p>The patient's facial expression appears somewhat distressed or concerned, and her vocal tone suggests frustration or irritation. These indicators reflect an emotion of distress or concern, possibly related to her difficulties with memory and focus. There is cognitive impairment present. The type of cognitive impairment observed is attention and memory impairment. The patient's expression and her response to the questions indicate difficulty in concentrating and recalling information, which are key aspects of attention and memory. No abnormalities in language ability are observed. The patient is able to answer the questions clearly, indicating that there are no noticeable issues with her language skills.</p> <p>Emotion: The user expresses a concern about their memory and attention. This reflects a negative emotion, possibly frustration or disappointment. Orientation: The user mentions that they feel disoriented, which suggests that they may have difficulty with spatial awareness and navigation. Memory: The user states that their memory has deteriorated, indicating that they may struggle to recall recent events or information. Attention: The user complains about their inability to concentrate, implying that they may have a short attention span. Language ability: There are no signs of aphasia or other language disorders.</p> <p>The user seems to be expressing concern about their recent memory and attention issues. They are seeking reassurance from the doctor and may be experiencing anxiety or worry about these changes. The user is experiencing some level of cognitive impairment, particularly in the areas of memory and attention. They are aware of these issues and are seeking reassurance from the doctor. It is important for the doctor to address these concerns and provide the user with appropriate support and guidance. I hope this analysis helps you understand the emotional state and cognitive impairment of the user in the conversation.</p> <p>Emotion: In the video, the person's facial expressions, including lower eyebrows, slight frown, and lip corner puller, suggest a sense of sadness or disappointment the person's tone is also somewhat flat, which further supports this interpretation their responses. Cognition: Several signs of potential cognitive decline are evident in this brief exchange. The patient reports memory and attention difficulties. No clear orientation or language ability difficulty are present at this stage.</p>

Figure 5: Comparative Case Analysis of Different Methods (Case 1). This utterance manifests negative emotion, memory and attention deficit.

Content	Method	Emotion-Cognition Captions
<p>Utterance: Doctor: Is it severe? Patient: Uh... when it's... the sky... it's not bright, it hurts. Very much. But when... the... um, blue... it's better. I think. Not yesterday. Or was it?</p> <p>Ground Truth: Negative emotion: Yes Orientation deficit: Yes Memory deficit: No Attention deficit: No Language disorder: Yes</p>	<p>Sa2VA</p> <p>Qwen2.5-Omni</p> <p>EmoLLM</p> <p>CPsyCoun</p> <p>Ours</p>	<p>Emotion: The person in the video appears to be experiencing discomfort or pain, as indicated by their description of the weather and their pain. They mention that the weather is bad and cloudy, causing them a lot of pain, while it is a bit better when the weather is sunny. This suggests that the person is experiencing physical discomfort or pain related to the weather conditions. Orientation: The person seems to be aware of their surroundings and the current weather conditions, indicating that their orientation is intact. Memory: The person is able to recall their past experiences with the weather and its impact on their pain, suggesting that their memory is functioning well. Attention: The person is focused on the conversation and is able to provide a detailed description of their pain and its relation to the weather, indicating that their attention is not impaired. Language Ability: The person is able to communicate their thoughts and feelings effectively, using appropriate vocabulary and grammar, which suggests that their language ability is intact.</p> <p>The patient appears to be in discomfort, as indicated by the tone of their voice and the attention they seem to be giving to the doctor's questions. They seem to be expressing a mix of pain and confusion, which could suggest distress or frustration. Persistently mentioning the weather conditions to the doctor indicates that the patient may be experiencing Attentional impairment. Additionally, the fact that the patient is focusing on how the weather affects their pain may suggest Memory impairment as they may not remember the condition of their health. There seems to be no noticeable issues with orientation or language ability.</p> <p>Emotion: The user describes the weather, which can be interpreted as a metaphor for their emotional state. When the weather is 'cloudy day', it reflects the user's emotional state of being 'serious'. When the weather is 'sunny day, it suggests that the user's emotional state is relatively better. Therefore, the intended meaning behind the user's words is to express their emotional state, which is serious due to the weather. Orientation: The user refers to the weather as a metaphor for their emotional state, indicating that they are able to relate the weather to their internal emotional experiences. Memory: The user recalls the weather conditions that led to their pain, suggesting that they have a good memory of the events. Attention: The user pays close attention to the weather conditions, as they directly affect their pain. Language Ability: The user is able to articulate their experiences and needs, indicating that they have good language ability.</p> <p>Emotion: The patient is expressing physical pain and discomfort. The use of the word 'hurt' indicates that the patient is experiencing physical pain. The patient also mentions that the weather is affecting their pain, suggesting that they may be experiencing some level of emotional distress due to the pain. Orientation: The patient is able to identify the current weather conditions and how it affects their pain, indicating that they have good orientation. Memory: The patient is able to recall past experiences with the weather and how it affects their pain, indicating that they have good memory. Attention: The patient is able to focus on the current situation and provide a detailed description of their pain, indicating that they have good attention. Language ability: The patient is able to communicate their pain and discomfort using clear and concise language, indicating that they have good language ability.</p> <p>Emotion: The person in the video is sitting at a desk, looking directly into the camera with a negative expression, which suggests a lack of emotional cues his words. The combination of these factors suggests that the person is likely experiencing a mild discomfort or pain, but their facial expression and tone do not convey any strong emotions such as fear, anger, or distress. Cognition: There are signs of cognitive impairment in the domains of orientation and language. The patient demonstrates difficulty articulating clear temporal or environmental references, suggesting disorientation. Additionally, the fragmented syntax, vague word choices and frequent pauses indicate a mild language disorder. Despite these issues, the patient maintains a logical connection between weather and physical symptoms, showing preserved memory and sustained attention throughout the conversation.</p>

Figure 6: Comparative Case Analysis of Different Methods (Case 2). This utterance manifests negative emotion, orientation deficit and language disorder.