Measuring the Unspoken: A Disentanglement Model and Benchmark for Psychological Analysis in the Wild

Yigui Feng
College of Computer Science, National University of Defense Technology
Deya Road 109, Changsha, 410073, Hunan, China

Qinglin Wang
College of Computer Science, National University of Defense Technology
Deya Road 109, Changsha, 410073, Hunan, China

Haotian Mo
College of Computer Science, National University of Defense Technology
Deya Road 109, Changsha, 410073, Hunan, China

Yang Liu
Shien-Ming Wu School of Intelligent Engineering, South China University of Technology
xingye Road 777, Guangzhou, 511442, Guangdong, China

Ke Liu
College of Computer Science, National University of Defense Technology
Deya Road 109, Changsha, 410073, Hunan, China

Gencheng Liu
College of Computer Science, National University of Defense Technology
Deya Road 109, Changsha, 410073, Hunan, China

Xinhai Chen
College of Computer Science, National University of Defense Technology
Deya Road 109, Changsha, 410073, Hunan, China

Siqi Shen
School of Informatics Xiamen University
Furong Road, Xiang'an District, Xiamen, 361005, Fujian, China

Songzhu Mei
College of Computer Science, National University of Defense Technology
Deya Road 109, Changsha, 410073, Hunan, China

Jie Liu
College of Computer Science, National University of Defense Technology
Deya Road 109, Changsha, 410073, Hunan, China

Abstract

Generative psychological analysis of in-the-wild conversations faces two fundamental challenges: (1) existing Vision-Language Models (VLMs) fail to resolve Articulatory-Affective Ambiguity, where visual patterns of speech mimic emotional expressions; and (2) progress is stifled by a lack of verifiable evaluation metrics capable of assessing vi-

sual grounding and reasoning depth. We propose a complete ecosystem to address these twin challenges. First, we introduce Multilevel Insight Network for Disentanglement(MIND), a novel hierarchical visual encoder that introduces a Status Judgment module to algorithmically suppress ambiguous lip features based on their temporal feature variance, achieving explicit visual disentanglement. Second, we construct ConvoInsight-DB, a new large-scale dataset with expert annotations for micro-expressions and deep psychological inference. Third, Third, we designed the Mental Reasoning Insight Rating Metric (PRISM), an automated dimensional framework that uses expert-guided LLM to measure the multidimensional performance of large mental vision models. On our PRISM benchmark, MIND significantly outperforms all baselines, achieving a +86.95% gain in micro-expression detection over prior SOTA. Ablation studies confirm that our Status Judgment disentanglement module is the most critical component for this performance leap. Our code has been opened.

1. Introduction

The aspiration to create artificial intelligence that genuinely understands human beings is a long-standing goal in computer science, holding the key to transformative applications from mental healthcare to trustworthy AI [8]. The true frontier for this endeavor lies not in controlled laboratory settings, but in the complex, unconstrained dynamics of "in-the-wild" conversations—the very scenarios where human understanding is most challenged, as psychological research like Truth-Default Theory (TDT) [17]suggests. It is in these low-stakes interactions where the only reliable clues to a person's inner state are often the most subtle, involuntary, and purely visual "emotional leakages" [33].

Capturing these visual signals is paramount, yet this presents a dual challenge. First, there is a critical technical obstacle in the visual domain which we term Articulatory-Affective Ambiguity. The same facial muscles used for emotional expression are also used for speech, creating profound ambiguity in a video-only analysis [32]. Current Vision-Language Models (VLMs) [6, 12], with their holistic feature aggregation, are architecturally incapable of resolving this, frequently misinterpreting the visual patterns of articulation as affective signals, as shown in Figure 1. Second, and equally important, is a critical void in meaningful evaluation. While evaluation protocols are established for affective computing in controlled or semi-controlled settings, they are fundamentally ill-equipped for the fluid, context-dependent nature of in-the-wild dialogue. This leaves the community without a standardized and automated method to assess a generated psychological analysis for its visual grounding, logical depth, and descriptive richness in these challenging, realistic scenarios.

In this work, we tackle these twin challenges of modeling and evaluation head-on. We introduce Multilevel Insight Network for Disentanglement(MIND), a novel hierarchical vision-language architecture designed to explicitly disentangle the visual features of speech from those of emotion. Critically, to measure its true capabilities, we also propose what is, to our knowledge, the first automatic evaluation framework:Psychological Reasoning Insight Scoring Metric(PRISM) specifically designed for the multifaceted assessment of generative psychological analysis in in-the-wild conversational settings. By combining a sophisticated model with a bespoke, meaningful benchmark, we create a complete ecosystem for advancing research in this challenging domain.

To summarize, we make the following contributions: (1) We propose MIND, a novel hierarchical framework that explicitly distinguishes articulatory movements from emotional signals. (2) We design and present a comprehensive evaluation framework, PRISM, which is the first of its kind for automatic, vision-based psychological analysis, particularly for natural conversation scenarios. (3) We construct a new dataset, ConvoInsight-DB, which features macro-expressions, micro-expressions, character emotion analysis, and character psychological analysis.

2. Related Work

Our research is positioned at the confluence of VLMs and fine-grained Facial Affective Computing, addressing critical gaps in both representation learning and evaluation.

Vision-Language Models and Facial Affective Computing. The current paradigm in video understanding is dominated by large VLMs like Video-LLaMA [47] and Qwen2.5-VL [2], which couple a holistic visual encoder with a Large Language Model (LLM). Concurrently, the field of Facial Affective Computing has a rich history of developing specialized models for recognizing discrete emotions or detecting micro-expressions [1, 34]. However, both lines of research fall short of our goal. VLMs are architecturally blind to the fleeting, localized facial cues, while traditional affective computing models are framed as recognition tasks, outputting mere labels rather than the rich, inferential generative analysis required to explain the 'why' behind an expression. Our work bridges this divide, reformulating the task from recognition to generative inference, enabled by a fine-grained visual encoder.

Disentangled Representation for Facial Analysis. The principle of learning disentangled representations has proven effective for separating static facial factors like identity from expression [41]. This body of work inspires our approach but has yet to address the core challenge of in-the-wild conversations: disentangling two dynamic, co-occurring, and visually ambiguous processes—the visual patterns of articulation and the subtle signals of affect. Therefore, to address

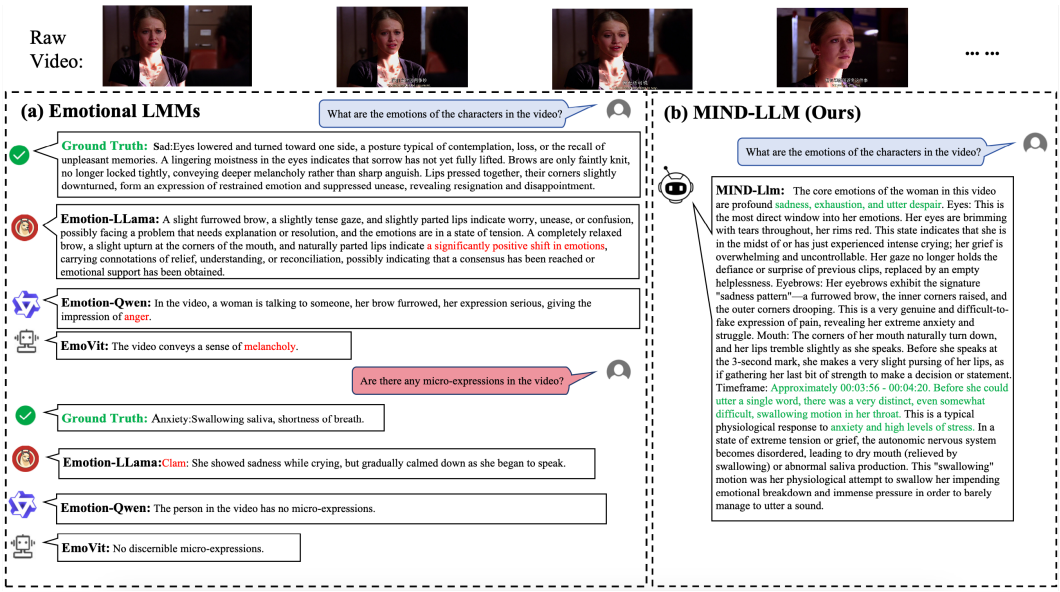

Figure 1. The motivation behind MIND (zoom in for detailed Q&A): (a) Current state-of-the-art large multimodal models (LMMs) for emotion recognition are inaccurate for video-based emotion recognition, unable to detect and analyze micro-expressions, and have limited ability to infer a person's psychological activities. (b) In contrast, MIND effectively addresses these limitations, enabling efficient psychological analysis and profiling of individuals. Incorrect outputs are marked in red; correct outputs are marked in green.

Benchmark	Granularity	Mental activity	Modality	Macro-expression analysis	Micro-expression detection	Micro-expressions analysis
CR [3]	Coarse	✗	Text	✗	✗	✗
Yelp [38]	Coarse	✗	Text	✗	✗	✗
SemEval [28]	Fine	✗	Text	✓	✗	✗
TOWE [10]	Fine	✗	Text	✗	✗	✗
ACOS [5]	Fine	✗	Text	✓	✗	✗
ASTE [27]	Fine	✗	Text	✓	✗	✗
DiaASQ [18]	Fine	✗	Text	✓	✗	✗
VER [14]	Fine	✗	Text, Video	✓	✗	✗
CMU-MOSEI [46]	Coarse	✗	Text, Audio, Video	✗	✗	✗
IEMOCAP [4]	Coarse	✗	Text, Audio, Video	✗	✗	✗
MELD [29]	Coarse	✗	Text, Audio, Video	✗	✗	✗
M3ED [48]	Coarse	✗	Text, Audio, Video	✗	✗	✗
CAS(ME)^2 [30]	Coarse	✗	Video	✗	✓	✗
PanoSent [24]	Fine	✗	Text, Image, Audio, Video	✓	✗	✗
PRISM	Fine	✓	Text, Audio, Video	✓	✓	✓

Table 1. Summary of existing popular benchmarks of sentiment analysis (representatively summarized, not fully covered).

this pain point, we designed the MIND visual network to decouple the visual features of pronunciation from the visual features of facial expression changes.

Micro-expression information characteristics are linked to human psychological activities. Previous research has not fully explored the role of micro-expressions in large-scale models' inference of human psychology. In most cases, micro-expression information features are considered as part of macro-expression features. Furthermore, micro-expressions are rarely combined with the specific psychological activities of the characters.Micro-expression information is a key source of information for analyzing a person's true inner thoughts [35]. To address this, this paper proposes a new dataset, ConvoInsight-DB, designed to bridge these gaps and improve the ability of large-scale models to infer human psychology.

Evaluation of Psychological Analysis in Conversational Settings. A significant bottleneck hindering progress is the lack of meaningful evaluation. While evaluation protocols exist for traditional affective computing—such as F1 scores for emotion classification on datasets like AffectNet [25]—these are designed for tasks with discrete, unambiguous labels, often in controlled or semi-controlled environments. Standard text generation metrics (e.g., BLEU [26], ROUGE [20]) are inadequate because they only measure superficial n-gram overlap and fail to capture semantic accuracy, factual basis, or psychological depth. They lack the capacity to evaluate the inferential and compositional nature of analysis required for in-the-wild conversations, where an expression's meaning is heavily context-dependent and intertwined with speech. More recent "LLM-as-a-judge" approaches [37, 42], while powerful, cannot verify visual grounding and are not tailored to the specific psychological constructs we aim to measure. Table 1 summarizes the

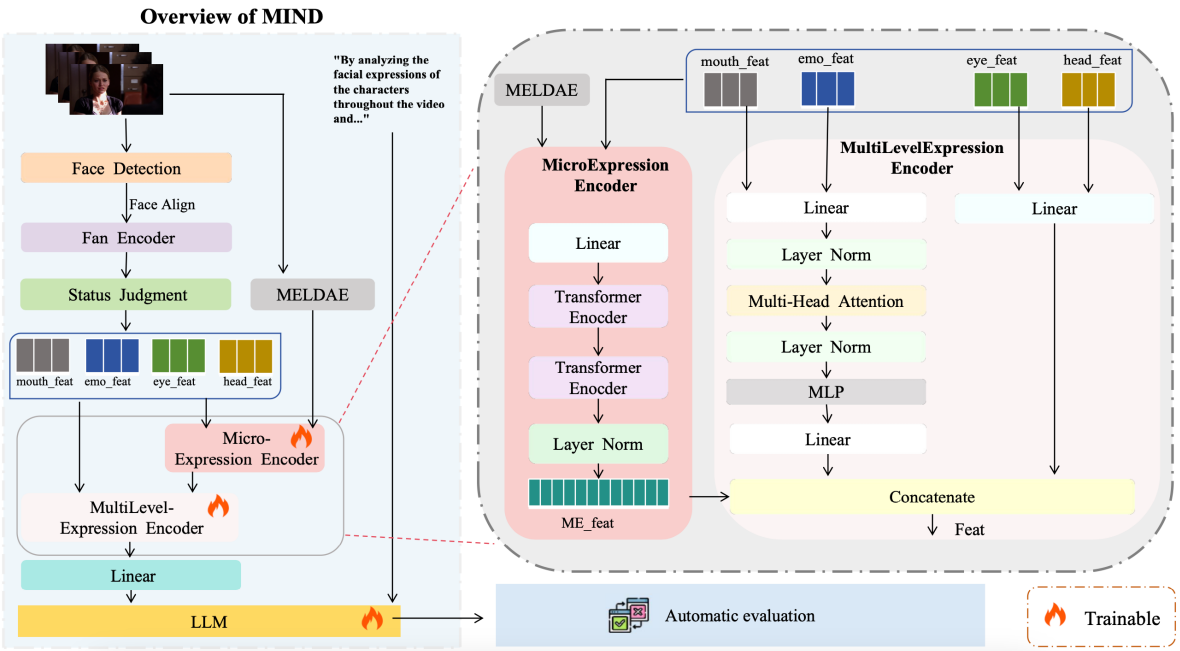

Figure 2. An overview of the MIND training process and model architecture. The FanEncoder module decouples expression features from the video, while the MicroExpressionEncoder extracts micro-expression features. The MultiLevelExpressionEncoder integrates micro-expression emotion features with macro-expression features. The detailed structure of the MicroExpressionEncoder and MultiLevelExpressionEncoder are shown in the figure to the right.

key differences between this work and existing benchmarks. Our work directly fills this gap, introducing PRISM, for generative analysis of human psychology in unconstrained conversational contexts.

3. Method

We propose `MIND`, whose overall architecture is shown in Figure 2. It consists of three main stages: (1) Spatiotemporal Feature Decomposition and Disentanglement; (2) The `MIND` Core Encoder for parallel, multi-level encoding; and (3) Vision-Language Alignment and Generative Reasoning.

3.1. Spatiotemporal Feature Decomposition and Disentanglement

This stage, corresponding to the left flow in Figure 2, aims to: (1) extract a set of structured, multi-dimensional facial dynamics features from the raw video; and (2) perform our key disentanglement step to resolve Articulatory-Affective Ambiguity.

Expert Feature Extraction Given a raw video clip V, we first utilize a set of pre-trained, frozen expert models to decompose it into parallel, semantically-disentangled feature streams. This flow includes three sub-steps:

- **Face Detection [15]:** Localizes facial bounding boxes in each frame.
- **Face Align [40]:** Normalizes the detected facial regions to mitigate scale and rotation effects.
- **Dynamics Encoding:** Feeds the aligned facial sequence into the `Fan Encoder` [39], which explicitly decomposes the complex facial dynamics into four key feature streams:
  - **Head Pose** (E_{head})**:** `head_feat`
  - **Eye Dynamics** (E_{eye})**:** `eye_feat`
  - **Core Emotion** (E_{emo})**:** `emo_feat`
  - **Lip Movement** (E_{lip})**:** `mouth_feat`

Context-Aware Disentanglement This is the key innovation in our methodology for resolving Articulatory-Affective Ambiguity. After feature extraction, we introduce the `Status Judgment` module, which operates exclusively on the most ambiguous E_{lip} stream.

Principle: Our core hypothesis is that the articulation state visually manifests as continuous and high-amplitude variations in the `mouth_feat` feature over time, whereas silent or simple emotional expressions (e.g., holding a smile) produce features that are comparatively static or change only briefly.

Implementation: The `Status Judgment` module performs a Temporal Contrastive Test on the E_{lip} feature sequence $\mathbf{f}_{1...T}$ (where $\mathbf{f}_t \in \mathbb{R}^{D_{lip}}$). Specifically, we assess the magnitude and continuity of its variation by computing the

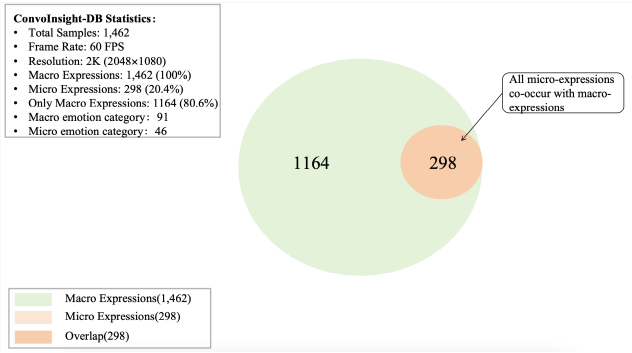

Figure 3. Main statistics of ConvoInsight-DB dataset.

temporal variance (V_{var}) and the sum of adjacent differences (V_{sad}), as defined in Equation (1):

$$c = \mathbb{I}\left(\alpha \cdot V_{var} + \beta \cdot V_{sad} > \tau\right) \tag{1}$$

where V_{var} and V_{sad} are defined as:

$$V_{var} = \frac{1}{T}\sum_{t=1}^{T} ||\mathbf{f}_t - \boldsymbol{\mu}||_2^2 \quad \text{and} \quad V_{sad} = \sum_{t=2}^{T} ||\mathbf{f}_t - \mathbf{f}_{t-1}||_2 \tag{2}$$

Here, $\boldsymbol{\mu}$ is the mean feature vector over all frames, and $\mathbb{I}(\cdot)$ is the indicator function. $c = 1$ indicates an "articulation state," and $c = 0$ indicates a "non-articulation state." α, β, and τ are hyperparameters.

Purification: Based on this judgment, we perform a gated suppression on the `mouth_feat` stream to obtain the purified lip feature $\hat{E}_{lip}$:

$$\hat{E}_{lip} = (1 - c) \cdot E_{lip} \tag{3}$$

This step ensures that the subsequent `MIND` core encoder is not misled by purely articulatory motions, achieving noise disentanglement at the source of the visual modality.

3.2. The MIND Core Encoder

This is the core innovation of our method, corresponding to the right-hand detail in Figure 2. The `MIND` framework receives the (purified) feature streams from Step 1 and processes them in a parallel, multi-level manner to produce a single, information-rich visual representation. It consists of two parallel sub-modules: the `MicroExpressionEncoder` and the `MultiLevelExpressionEncoder`.

Transient Path: MicroExpressionEncoder This module (the pink box in Figure 2) is specialized in capturing fleeting yet critical micro-expression signals, guided by explicit temporal localization.

- **Inputs:** The module receives two distinct sets of inputs:
  1. **Feature Streams:** $E_{\text{head}}, E_{\text{eye}}, E_{\text{emo}}, \hat{E}_{\text{lip}}$.
  2. **Temporal Localization:** Start and end timestamps $(t_{\text{start}}, t_{\text{end}})$ identified by a parallel `MELDAE` [11] module.
- **Processing:** The module extracts the segment $S_{ME} = [E_{\text{head}}, E_{\text{eye}}, E_{\text{emo}}, \hat{E}_{\text{lip}}]_{t_{\text{start}}:t_{\text{end}}}$.
- **Output:** Generates a compact feature $F_{ME} \in \mathbb{R}^{D_{me}}$.

Fusion Path: MultiLevelExpressionEncoder This module (the light pink box in Fig. 2) is the "fusion center" of the entire system, and its task is to enhance facial features. It processes features from different levels of abstraction in parallel and intelligently combines them.

- **Input:** The module receives five parallel inputs: F_{ME}, $\hat{E}_{lip}$,E_{emo}, E_{eye},E_{head}
- **Processing:** The module features four parallel processing paths corresponding to different feature hierarchies:
  - **Path 1 (Transient):** F_{ME} is treated as the highest-level feature and is passed directly to the final fusion step without further processing.
  - **Path 2 (Complex Dynamics):** $\hat{E}_{lip}$ and E_{emo}, the two most complex expression features, are first concatenated. They are then fed into a deep processing stream: `Linear` → `Layer Norm` → `Multi-Head Attention` → `Layer Norm` → `MLP` → `Linear`. This path allows the model to fully mine the complex relationships between the lip and core emotion features.
  - **Path 3 (Simple Dynamics - Eye):** E_{eye} is passed through a separate `Linear` layer for simple feature mapping.
  - **Path 4 (Simple Dynamics - Head):** E_{head} is similarly passed through a separate `Linear` layer.
- **Fusion:** As depicted, the output features from all four paths ($F'_{ME}, F'_{complex}, F'_{eye}, F'_{head}$) are finally concatenated at the `Concatenate` step to form a single, comprehensive fusion vector V_{MIND} (`Feat` in the figure):

$$V_{MIND} = \text{Concat}(F'_{ME}, F'_{complex}, F'_{eye}, F'_{head}) \tag{4}$$

3.3. Vision-Language Alignment and Generative Reasoning

This final stage corresponds to the last part of the flowchart, responsible for translating our distilled visual evidence V_{MIND} into natural language analysis that the LLM can understand and reason with.

- **Visual Alignment:** The fused feature vector V_{MIND} is first mapped into the LLM's word embedding space via a `Linear` layer (the projector W_p) to obtain $V_{proj} \in \mathbb{R}^{D_{llm}}$.
- **Instruction Injection & Generation:** V_{proj} is injected into the `LLM` to replace a special visual token `<expr>` in the prompt template.

This step can be formulated as follows:

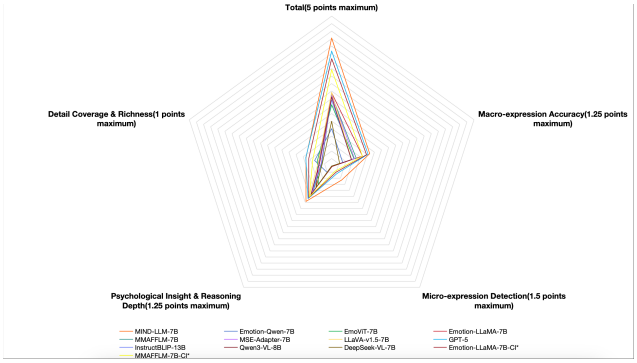

Figure 4. Balanced performance of large multimodal models (LMMs).MIND-LLM-8B demonstrates superior performance across all evaluation dimensions. Evaluation details are shown in Table 3.

Input Data: V_{MIND}
Instruction:You are an expert in psychology and micro-expression analysis. Please provide a detailed analysis of the person's facial expressions, emotional state, and inner psychology based on this expression feature.

Expected Output: Analysis Text

This step can be expressed as:

$$Analysis_{Text} \leftarrow f_{LLM}(P_{prompt} \oplus W_p(V_{MIND})) \quad (5)$$

where f_{LLM} is the Large Language Model, which we fine-tune using Parameter-Efficient Fine-Tuning (PEFT) with Low-Rank Adaptation (LoRA). The $\oplus$ operation denotes the injection of the projected visual vector V_{proj} to replace the `<expr>` token's embedding. Finally, the generated $Analysis_{Text}$ is fed into our `Automatic evaluation` framework for assessment.

3.4. Automatic Evaluation Framework

PRISM leverages the powerful external large-scale language model GPT-4o [16], enabling it to act as a "psychology expert." For each video sample, LLM assessors receive detailed scoring guidelines across four dimensions set by psychologists: ground-truth "micro-expression labels" (out of 1.25 points), ground-truth "macro-expression labels" (out of 1.5 points), psychological insight and reasoning depth (out of 1.25 points), and detail coverage and richness (out of 1 point). The final score for each generative analysis is calculated as the sum of these four dimensions (out of 5). This provides a robust, reproducible, and semantically rich benchmark for our task. The complete and detailed scoring criteria provided to LLM evaluators can be found in the Appendix C.

4. Experiments

4.1. ConvoInsight-DB

The ConvoInsight-DB dataset, used to analyze sentiment in natural dialogue scenarios, was primarily collected by crawling a large number of publicly available movies and TV shows.The focus on micro-expression clips from open-source movies is intended to lay the groundwork for our future micro-expression video generation model; the data we are currently using will be used to guide our micro-expression video generation model. We then conducted a rigorous screening process (combining manual review with automated filtering, such as keyword detection and Toxic-BERT detection [13]) to eliminate instances that were harmful, personal, or emotionally irrelevant. After obtaining the cleaned corpus, we collaborated with a team of psychology experts to label the video data. To ensure reliability, each video was independently annotated by at least three different psychology professionals, as well as an AI psychology expert (role-playing). After the annotations were completed, three different psychology professors discussed and jointly finalized the labels. Instances with unresolved ambiguities were removed. Figure 3 shows the details of ConvoInsight-DB.

4.2. Implementation Details

Following the pre-training process, MIND uses eight NVIDIA H800 80GB GPUs for distributed training. The training strategy employs the LoRA efficient fine-tuning method, with rank set to 16, alpha set to 32, and dropout set to 0.05. See Appendix B for specific implementation details.

4.3. ConvoInsight-DB dataset validity verification

To assess the effectiveness of the dataset, we conducted a comprehensive evaluation. We fine-tuned the Qwen2.5vl-7b [2] model (CIQwen-7b) using the ConvoInsight-DB dataset and compared CIQwen-7b with the GPT4o [16] direct role-playing baseline method using the PRISM. As shown in Table.2, CIQwen-7b consistently and significantly outperformed the baseline method in all dimensions: Macro-expression Accuracy (+23.5%), Micro-expression Detection (+83.3%), Psychological Insight & Reasoning Depth(+31.6%), and Detail Coverage & Richness (+48%). This clearly demonstrates the effectiveness of our dataset in enhancing the model's sentiment analysis capabilities.

4.4. Comparison with Baselines

Baselines and Metrics. We selected several large-scale models focused on video sentiment analysis and general large-scale models as baselines for comparative evaluation, including Emotion-Qwen-7B [14], EmoViT-7B [43], Emotion-LLaMA-7B [7], MMAFFLM-7B [22],

Method	Mac	Mic	PIRD	DCR
Role-play	0.51	0.06	0.57	0.25
CIQwen-7b	0.63	0.11	0.75	0.37
Improvement	**+23.5%**	**+83.3%**	**+31.6%**	**+48%**

Table 2. Comparison of Role-play vs. CIQwen-7b across emotion categories. Metric abbreviations: Mac (Macro-expression Accuracy), Mic (Micro-expression Detection), PIRD (Psychological Insight & Reasoning Depth), DCR (Detail Coverage & Richness).

MSE-Adapter-7B [45], LLaVA-v1.5-7B [21],GPT-5 [1], InstructBLIP-13B [19], Qwen3-VL-8B [31], and DeepSeek-VL-7B [23]. The LLM in our MIND uses the Qwen3-8B [44] model(MIND-LLM-8B). We evaluate all methods on PRISM.See Appendix A for model prompts.

Quantitative Results. Table 3 compares the performance of MIND-LLM-8B with other baseline models on the PRISM evaluation framework, yielding the following observations. First, due to the rich micro-expression features in the data, we found that both general-purpose large models and sentiment analysis large models performed very poorly, failing to recognize almost any micro-expressions. However, enhancing sentiment understanding and inference capabilities can address this issue to some extent, thereby improving overall performance. For example, compared to the best-performing baseline model (GPT-5), our MIND-LLM-8B achieved an 86.95% performance improvement in micro-expression detection. Interestingly, we found that GPT-5 performed best among all baseline models, which was unexpected, indicating that larger parameters do indeed improve the model's understanding capabilities. For fairness, we also fine-tuned the two best-performing sentiment analysis large models, Emotion-LLaMA-7B and MMAFFLM-7B, using our ConvoInsight-DB dataset (Emotion-LLaMA-7B-CI*, MMAFFLM-7B-CI*). It is evident that Emotion-LLaMA-7B-CI and MMAFFLM-7B-CI received good feedback in each dimension, almost comparable to GPT-5, but still slightly inferior to MIND-LLM-8B. Overall, MIND-LLM-8B achieved significant results in facial expression analysis and psychological interpretation.

Finally, we can examine the task evaluation results from different perspectives. For micro-expression detection, current mainstream emotion models all perform well. However, it can be seen that nearly all models perform significantly worse in this area. Recognizing and reasoning micro-expressions is the greatest challenge facing AI in understanding human psychology, providing a challenging benchmark for subsequent research.

[1] https://chatgpt.com/

Model	Metrics			
	Mac	Mic	PIRD	DCR
Emotion LMMs				
Emotion-Qwen-7B	0.233	0	0.156	0.358
EmoViT-7B	0.471	0	0.415	0.305
Emotion-LLaMA-7B	0.649	0.013	0.538	0.268
MMAFFLM-7B	0.517	0.006	0.636	0.207
MSE-Adapter-7B	0.482	0.002	0.692	0.219
Emotion-LLaMA-7B-CI*	0.735	0.152	0.781	0.483
MMAFFLM-7B-CI*	0.648	0.124	0.759	0.399
General LMMs				
GPT-5	0.767	0.184	0.805	0.545
LLaVA-v1.5-7B	0.422	0	0.774	0.279
InstructBLIP-13B	0.420	0	0.637	0.263
Qwen3-VL-8B	0.416	0.009	0.708	0.250
DeepSeek-VL-7B	0.167	0	0.533	0.195
MIND-LLM-8B	**0.802**	**0.344**	**0.875**	**0.541**

Table 3. Comparison of MIND-LLM-8B with other large emotion models and general large models on PRISM.Metric abbreviations: Mac (Macro-expression Accuracy), Mic (Micro-expression Detection), PIRD (Psychological Insight & Reasoning Depth), DCR (Detail Coverage & Richness).

Variant	**Mac**	**Mic**	**PIRD**	**DCR**
MIND (Full)	**0.802**	**0.344**	**0.875**	**0.541**
(a) w/o Status Judgment	0.739	0.286	0.727	0.425
(b) w/o Micro.Enc	0.718	0.302	0.760	0.441
(c) w/o Multi.Enc	0.653	0.228	0.713	0.391
(d) Basic Fusion (MLP)	0.730	0.236	0.656	0.309

Table 4. Ablation studies were conducted using the ConvoInsight-DB dataset on PRISM. Metric abbreviations: Mac (Macro-expression Accuracy), Mic (Micro-expression Detection), PIRD (Psychological Insight & Reasoning Depth), DCR (Detail Coverage & Richness).

4.5. Ablation Studies

To rigorously validate the effectiveness of our proposed `MIND` framework and quantify the contribution of each of its core components, we conduct a comprehensive series of simplification studies. We systematically degrade the performance of the full model by removing or replacing key modules and evaluate these variants using PRISM, our proposed automated evaluation framework. The results are summarized in Table 4.

Effectiveness of the Status Judgment Module. To verify our core disentanglement hypothesis, we create a variant (a) **w/o Status Judgment**. In this setting, we completely bypass the module described in Sec 3.1.2. The raw, unpurified `mouth_feat` (E_{lip}) is directly fed into the subsequent

LLM Backbone	Size	Final Score
ViT + Qwen3-8B	8B	1.496
MIND + Llama-3-8B	8B	2.419
MIND + Qwen3-8B	**8B**	**2.562**

Table 5. Comparison of different LLM backbones integrated with our frozen MIND visual encoder.

encoders. As shown in Table 4, the performance of this variant is severely degraded, with Micro-expression Detection performance decreasing by 16.86%. The impact is most devastating on the "Micro-expression Detection" and "Psychological Insight" dimensions. This empirically validates our central claim: without explicitly identifying and suppressing articulatory noise, the model is fatally misled by speech-related motions. This leads to factually incorrect visual grounding (mistaking speech for emotion) and, consequently, logically flawed psychological inferences.

Effectiveness of the MicroExpressionEncoder Module. To quantify the contribution of our specialized transient path, we design variant (b) **w/o Micro.Enc** Here, we remove the entire `MicroExpressionEncoder` and its corresponding path. The `MultiLevelExpressionEncoder` thus loses its most abstract input, F_{ME}, and must infer the entire state solely from the other four feature streams. Table 4 shows a significant performance drop, primarily concentrated in the "Micro-expression Detection" score, which plummets from 0.344 to 0.302. This confirms that a dedicated, attentive module is crucial for isolating the faint, transient signals of micro-expressions, which are otherwise lost or "averaged out" by the main fusion encoder.

Importance of the Micro-Expression Encoder. To verify the expression enhancement effect of the MultiLevelExpressionEncoder block, we created a variant (c) **w/o MultiL.Enc**. This variant directly inputs features into the subsequent Liner layer without the feature enhancement provided by the MultiLevelExpressionEncoder module. As shown in Table 4, this variant exhibits a significant performance decrease, with a 18.58% drop in Macro-expression Accuracy and a 31.40% drop in Micro-expression Detection. This effectively demonstrates the necessity of the MultiLevelExpressionEncoder module.

Contribution of the MultiLevel-Expression Encoder. To validate our hierarchical fusion architecture, we test variant (d) **Basic Fusion (MLP)**. We compares its `Hybrid Compressor` against a simpler `MLP projector`. Here, we replace our complex, parallel-path `MultiLevelExpressionEncoder` (the light pink box in Figure 2) with a simple "brute-force" fusion block: all five input features (F_{ME}, $\hat{E}_{lip}$, E_{emo}, E_{eye}, E_{head}) are first concatenated and then passed through a single, deep MLP to produce the final V_{MIND}. Table 4 show that this non-hierarchical approach underperforms our full model across all four dimensions. This validates our design of using specialized, parallel paths (e.g., the deep attention stream for E_{emo}/$\hat{E}_{lip}$ vs. simple `Linear` layers for E_{eye}/E_{head}) to handle features of different complexities, proving it is more effective than a monolithic fusion block.

Impact of the LLM Backbone. Finally, we explore the generalization ability of our framework by replacing the LLM backbone model. We keep the `MIND` visual encoder unchanged and retrain the projector and LoRA adapter for an alternative model `Llama-3-8B` [9] and compare it with the default `Qwen3-8B` [44]. The results are listed in Table 5. As shown, the more powerful LLM backbone model consistently achieves higher scores, especially on the "psychological insight" dimension, confirming that high-level reasoning capabilities are crucial for the final analysis quality [36]. However, we also observe that our `MIND` framework significantly outperforms the baseline model across all backbone models, indicating that our visual front-end is a robust, model-agnostic component that can provide critical, cleansed signals to any downstream LLM.

5. Discussion

Our experimental results, particularly the ablation studies, provide strong evidence for our central hypotheses. However, these results also illuminate several non-trivial challenges, limitations, and avenues for future inquiry.

On the Criticality of Disentanglement. The most significant finding from our ablations (Table 4) is the catastrophic performance collapse observed in the "w/o Status Judgment" variant. The final score dropped, which empirically validated our core argument: clearly addressing euphony is not merely a "nice-to-have" feature, but a fundamental prerequisite for accomplishing this task. It suggests that without a mechanism to filter speech-related visual noise at the source, the model is fatally misled, leading to factually incorrect visual grounding. This finding aligns with the broader trend in multimodal research, such as the `Emotion-Qwen` paper's use of specialized expert modules [14], which argues against one-size-fits-all processing and in favor of specialized components for distinct sub-tasks.

Reflections on our Evaluation Framework. A key contribution of this work is the proposal of a new, fine-grained automatic evaluation framework. Our experiments revealed its utility: baseline models, while sometimes generating

plausible-sounding psychological "fluff," scored near-zero on the "Micro-expression Detection" dimension. A standard metric like BLEU or even a generic "LLM-as-a-judge" would have been susceptible to these "plausible hallucinations." Our framework, by explicitly grounding the score in the detection of specific macro/micro expressions, provided a much-needed, verifiable assessment of the model's factual accuracy and reasoning depth. This demonstrates the necessity of domain-specific evaluation protocols for nuanced tasks.

Limitations of the Status Judgment Heuristic. While our `Status Judgment` module (Sec 3.1.2) proved highly effective, it remains a heuristic-based approach. The formula, $c = \mathbb{I}(\alpha \cdot V_{var} + \beta \cdot V_{sad} > \tau)$, is based on the assumption that speech always produces high-variance features. This assumption may falter in edge cases:

- **Subtle Talkers:** A person speaking very quietly or with minimal lip movement might be misclassified as "non-speaking" ($c = 0$), causing their articulatory movements to be incorrectly fed into the model as emotional cues.
- **Expressive Non-Speakers:** Conversely, a person exhibiting rapid, non-speech-related facial tics or chewing might be misclassified as "speaking" ($c = 1$), causing their genuine (but noisy) expression features to be erroneously suppressed.

A more robust, end-to-end learned disentanglement module is a promising direction for future work.

Limitations of the Visual-Only Modality. A more fundamental limitation is our model's deliberate reliance on a single modality (vision). By design, `MIND` is deaf. It cannot distinguish between a subject holding their breath in terror (a purely visual cue) and a subject merely pausing mid-sentence. While our work pushes the boundary of what can be inferred from vision alone, the integration of audio prosody (pitch, tone, energy) is the clear next step to resolve such deep ambiguities and achieve a truly holistic understanding.

Broader Impact and Ethical Considerations. As with any technology capable of inferring human psychological states, the potential for misuse must be soberly addressed. An automated system for psychological analysis, especially in non-clinical, "in-the-wild" settings, carries significant ethical risks, including surveillance, privacy violation, and use in biased contexts (e.g., hiring, interrogations). Furthermore, our model's interpretations are inherently tied to the cultural and demographic biases present in its training data; expressions of "sadness" or "confidence" are not universal. We position `MIND` not as a "lie detector" or a diagnostic tool, but as a research step towards building more empathetic AI assistants (e.g., for mental health chatbots or assistive technology). We strongly advocate for future work in this area to include rigorous auditing for fairness, bias, and transparency before any real-world deployment.

6. Conclusion

This paper tackled the critical challenge of Articulatory-Affective Ambiguity in generative psychological analysis of in-the-wild videos. We introduced MIND, a hierarchical visual encoder that explicitly disentangles speech-related motion from emotion, and proposed a novel, multi-dimensional automatic evaluation framework to assess visually-grounded psychological insight. Our experiments demonstrate that MIND significantly outperforms baselines, validating the necessity of explicit visual disentanglement. Together, our model, dataset (ConvoInsight-DB), and evaluation protocol provide a robust foundation for future research in this domain. Code and data will be released to foster reproducibility and further progress. Future work will explore integrating audio modalities for a more holistic understanding.

References

[1] G. Akilandasowmya and K. L. Joshitha. Emotion recognition from micro-expressions using logistic regression and stochastic gradient descent. In *2024 International Conference on Communication, Computing and Internet of Things (IC3IoT)*, pages 1–9. IEEE, 2024. 2

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 6

[3] John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the ACL*, pages 440–447, 2007. 3

[4] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42: 335–359, 2008. 3

[5] Hongjie Cai, Rui Xia, and Jianfei Yu. Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions. In *Proceedings of the ACL*, pages 340–350, 2021. 3

[6] Z. Chen, L. Xu, H. Zheng, and et al. Evolution and prospects of foundation models: From large language

models to large multimodal models. *Computers, Materials & Continua*, 80(2), 2024. 2

[7] Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. In *NeurIPS*, pages 110805–110853. Curran Associates, Inc., 2024. 6

[8] C. Du, K. Fu, B. Wen, and et al. Human-like object concept representations emerge naturally in multimodal large language models. *Nature Machine Intelligence*, 7:860–875, 2025. 2

[9] A. Dubey, A. Jauhri, A. Pandey, and et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 8

[10] Zhifang Fan, Zhen Wu, Xin-Yu Dai, Shujian Huang, and Jiajun Chen. Target-oriented opinion words extraction with target-fused neural sequence labeling. In *Proceedings of the ACL*, pages 2509–2518, 2019. 3

[11] Yigui Feng, Qinglin Wang, Yang Liu, Ke Liu, Haotian Mo, Enhao Huang, Gencheng Liu, Mingzhe Liu, and Jie Liu. Meldae: A framework for micro-expression spotting, detection, and automatic evaluation in in-the-wild conversational scenes. 2025. 5

[12] Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. *arXiv preprint arXiv:2404.07214*, 2024. 2

[13] Laura Hanu and Unitary team. Detoxify. Github. https://github.com/unitaryai/detoxify, 2020. 6

[14] Dawei Huang, Qing Li, Chuan Yan, Zebang Cheng, Yurong Huang, Xiang Li, Bin Li, Xiaohui Wang, Zheng Lian, and Xiaojiang Peng. Emotion-qwen: Training hybrid experts for unified emotion and general vision-language understanding, 2025. 3, 6, 8

[15] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: Adaptive curriculum learning loss for deep face recognition. In *CVPR*, pages 1–8, 2020. 4

[16] A. Hurst, A. Lerer, A. P. Goucher, and et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 6

[17] T. R. Levine. The relative frequency of true and false confessions and denials in an experimental paradigm designed to investigate deception detection. *Communication Studies*, 75:362–376, 2024. 2

[18] Bobo Li, Hao Fei, Fei Li, Yuhan Wu, Jinsong Zhang, Shengqiong Wu, Jingye Li, Yijiang Liu, Lizi Liao, Tat-Seng Chua, and Donghong Ji. Diaasq: A benchmark of conversational aspect-based sentiment quadruple analysis. In *Findings of the ACL*, pages 13449–13467, 2023. 3

[19] Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven C. H. Hoi. Lavis: A library for language-vision intelligence, 2022. 7

[20] C. Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 3

[21] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 7

[22] Zhiwei Liu, Lingfei Qian, Qianqian Xie, Jimin Huang, Kailai Yang, and Sophia Ananiadou. Mmaffben: A multilingual and multimodal affective analysis benchmark for evaluating llms and vlms. *arXiv preprint arXiv:2505.24423*, 2025. 6

[23] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. Deepseek-vl: Towards real-world vision-language understanding, 2024. 7

[24] Meng Luo, Hao Fei, Bobo Li, Shengqiong Wu, Qian Liu, Soujanya Poria, Erik Cambria, Mong-Li Lee, and Wynne Hsu. Panosent: A panoptic sextuple extraction benchmark for multimodal conversational aspect-based sentiment analysis. *arXiv preprint arXiv:2408.09481*, 2024. 3

[25] A. Mollahosseini, B. Hasani, and M. H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017. 3

[26] K. Papineni, S. Roukos, T. Ward, and et al. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 3

[27] Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. In *AAAI*, pages 8600–8607, 2020. 3

[28] Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the SemEval*, pages 27–35, 2014. 3

[29] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the ACL*, pages 527–536, 2019. 3

[30] F. Qu, S. J. Wang, W. J. Yan, and et al. Cas (me) 2: A database of spontaneous macro-expressions and micro-expressions. In *International Conference on*

Human-Computer Interaction, pages 48–59, Cham, 2016. Springer International Publishing. 3

[31] Qwen team, Alibaba Cloud. Qwen3-vl is the multimodal large language model series developed by qwen team, alibaba cloud. Web, 2025. Published on 2025.09.23. 7

[32] A. Seiger. The perception of mixed emotions from dynamic video-only and audio-only expressions. 2022. Journal name, volume, and pages not provided. 2

[33] L. Silla. Detection of simple and complex deceits through facial micro-expressions: a comparison between human beings' performances and machine learning techniques. 2022. Journal name, volume, and pages not provided. 2

[34] J. Singh and P. Malik. Unveiling hidden emotions: a review of microexpression recognition, classification, and datasets. *Multimedia Tools and Applications*, pages 1–56, 2025. 2

[35] J. Singh and P. Malik. Unveiling hidden emotions: a review of microexpression recognition, classification, and datasets. *Multimedia Tools and Applications*, pages 1–56, 2025. 3

[36] T. Susnjak and T. R. McIntosh. Chatgpt: The end of online exam integrity? *Education Sciences*, 14(6):656, 2024. 8

[37] A. Szymanski, N. Ziems, H. A. Eicher-Miller, and et al. Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 952–966, 2025. 3

[38] Duyu Tang, Bing Qin, and Ting Liu. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the EMNLP*, pages 1422–1432, 2015. 3

[39] D. Wang, Y. Deng, Z. Yin, H.-Y. Shum, and B. Wang. Progressive disentangled representation learning for fine-grained controllable talking head synthesis. In *CVPR*, pages 17979–17989, 2023. 4

[40] Duomin Wang, Yu Deng, Zixin Yin, Heung-Yeung Shum, and Baoyuan Wang. Progressive disentangled representation learning for fine-grained controllable talking head synthesis. In *CVPR*, 2023. 4

[41] Qiang Wang, Mengchao Wang, Fan Jiang, Yaqi Fan, Yonggang Qi, and Mu Xu. Fantasyportrait: Enhancing multi-character portrait animation with expression-augmented diffusion transformers. *arXiv preprint arXiv:2507.12956*, 2025. 2

[42] R. Wang, J. Guo, C. Gao, and et al. Can llms replace human evaluators? an empirical study of llm-as-a-judge in software engineering. *Proceedings of the ACM on Software Engineering*, 2(ISSTA):1955–1977, 2025. 3

[43] Hongxia Xie, Chu-Jun Peng, Yu-Wen Tseng, Hung-Jen Chen, Chan-Feng Hsu, Hong-Han Shuai, and Wen-Huang Cheng. Emovit: Revolutionizing emotion insights with visual instruction tuning. In *CVPR*, 2024. 6

[44] A. Yang, A. Li, B. Yang, and et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 7, 8

[45] Y. Yang, X. Dong, and Y. Qiang. Mse-adapter: A lightweight plugin endowing llms with the capability to perform multimodal sentiment analysis and emotion recognition. In *AAAI*, pages 25642–25650, 2025. 7

[46] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the ACL*, pages 2236–2246, 2018. 3

[47] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 543–553, Singapore, 2023. Association for Computational Linguistics. 2

[48] Jinming Zhao, Tenggan Zhang, Jingwen Hu, Yuchen Liu, Qin Jin, Xinchao Wang, and Haizhou Li. M3ed: Multi-modal multi-scene multi-label emotional dialogue database. In *Proceedings of the ACL*, pages 5699–5710, 2022. 3